Randomized trials for policy: a review of the external validity of treatment effects *

Seán M. Muller
School of Economics
University of Cape Town

Draft for Annual Bank Conference on Development Economics (ABCDE)

May 14, 2014

*This paper has benefited from presentation of related work at the Economic Society of South Africa Conference, University of Stellenbosch, University of Cape Town, the Evidence and Causality in the Sciences (ECitS) conference and the CEMAPRE Workshop on the Economics and Econometrics of Education. I am grateful to Martin Wittenberg for comments on an earlier version. All errors and omissions remain my own.

Abstract

The paper provides a first survey of the literature on external validity, using as a starting point recent debates regarding the use of randomized evaluations to inform policy. Besides synthesising contributions to the programme evaluation literature we consider definitions of external validity from other sub-disciplines within economics, such as experimental economics and the time-series forecasting literature, as well as from disciplines such as philosophy and medicine. Following Cook and Campbell (1979) we argue that the fundamental challenge arises from interactive functional forms. This somewhat neglected point provides a framework in which to understand how and why extrapolation may fail. In particular it suggests that replication cannot resolve the external validity problem unless informed by some prior theoretical understanding of the causal relationship of interest. Finally, the problem of interaction can be used to show that the assumptions required for simple external validity are conceptually equivalent to those required for obtaining unbiased estimates of treatment effects using non-experimental methods, undermining the idea that internal validity needs to be rigorously assessed whereas external validity can be ascertained subjectively. Theory may play a role in aiding extrapolation, but the extent to which this will be possible in practice remains an open question.

In the last decade some researchers in economics have taken the view that randomized trials are the 'gold standard' for evaluating policy interventions and identifying causal effects, with this approach being particularly influential in development economics. This has led to controversy and a series of exchanges, including not only econometricians but philosophers, statisticians and policy analysts, regarding the uses and limitations of different econometric methods. Much of this debate concerns reasons why randomized evaluations may not, in practice, identify the causal effect of interest or, alternatively, may not identify a causal effect that is of relevance to policy. These concerns are broadly of three types: whether many questions of interest can be even notionally addressed via experimentation; reasons why identification of the causal effect in the experimental sample ('internal validity') may fail; and, limitations of the extent to which such an effect is informative outside of that sample population ('external validity').

While the literature on experimental and quasi-experimental methods deals extensively with threats to internal validity, and despite the popularisation of randomized evaluations due to their apparent usefulness for policy, the literature on external validity is remarkably undeveloped. Work on the subject has increased in recent years but there remains little guidance - and no consensus - on how estimated treatment effects can be used to estimate the likely effects of a policy in a different, or larger, population. The vast majority of empirical studies, including in top journals, contain no formal analysis of external validity. That is a particular problem in development economics where many researchers, admirably, seek to find solutions for pressing development challenges. The concern of this paper is to provide a survey - the first of its kind to our knowledge - of the literature on external validity, including contributions from other disciplines.

Section 1 details the broader debate about randomized trials in economics, provides formal notation and an outline of some key results, and lists specific criticisms of experimental methods. Section 2 reviews the existing literature on external validity, including some contributions from outside the programme evaluation literature. It draws out a number of common themes across these literatures, focusing in particular on the basic intuition that external validity depends on similarity of the population(s) of interest to the experimental sample. The final contribution, in section 3, develops a perspective on external validity based on the role of variables that interact with the cause of interest to determine individuals' final outcomes. This, we suggest, provides a framework within which to examine the question of population similarity in a way that allows for some formal statements - already developed by other researchers - of the requirements for external validity. These, in turn, have close resemblance to requirements for internal va-

lidity, which provides some basis for comparing and contrasting these two issues for empirical analysis. The paper concludes by arguing that it is not coherent to insist on formal methods for obtaining internal validity, while basing assessments of external validity on qualitative and subjective guesses about similarity between experimental samples and the population(s) of policy interest. Insisting on the same standards of rigour for external validity as for obtaining identification of causal effects would imply that much of the existing applied literature is inadequate for policy purposes. The obstacles to econometric analysis that underlie this conclusion are not limited to randomized evaluations and therefore consideration of external validity suggests more modesty, in general, in claiming policy relevance for results produced by experimental *and* non-experimental methods.

1 The credibility controversy: randomized evaluations for policymaking

The possibility of using econometric methods to identify causal relationships that are relevant to policy decisions has been the subject of controversy since the early and mid-20th century. The famous Keynes-Tinbergen debate (Keynes, 1939) partly revolved around the prospect of successfully inferring causal relationships using econometric methods, and causal terminology is regularly used in Haavelmo (1944)'s foundational contribution to econometrics. Heckman (2000, 2008) provides detailed and valuable surveys of that history. randomized experiments began to be used in systematic fashion in agricultural studies (by Neyman (1923)), psychology and education, though haphazard use had been made of similar methods in areas such as the study of telepathy.[1] Although some studies involving deliberate randomization were conducted in, or in areas closely relating to, economics the method never took hold and economists increasingly relied on non-experimental data sources, either cross-sectional datasets with many variables but limited time periods ('large N, small T'), or time series datasets with small numbers of variables over longer time periods ('small N, large T'). The former tended to be used by microeconometricians while the latter was favoured by macroeconometricians and this distinction largely continues to the present day. (Our concern in this study is the use of microeconometric methods to inform policy decisions and so, although an integration of these literatures is theoretically possible, we will focus on data sources characterised by limited time periods).

[1] Herberich, Levitt, and List (2009) provide an overview of randomized experiments in agricultural research and Hacking (1988) provides an entertaining account of experiments relating to telepathy.

For much of that era econometricians relied on two broad approaches to obtaining estimates of causal effects: structural modelling and non-structural attempts to include all possibly relevant covariates to prevent confounding/bias of estimated coefficients. The latter relied on obtaining statistically significant coefficients in regressions that were robust to inclusion of ('conditioning on') plausibly relevant covariates, where the case for inclusion of particular variables and robustness to unobservable factors was made qualitatively (albeit sometimes drawing on contributions to economic theory). The structural approach involves deriving full economic models of the phenomena of interest by making assumptions about the set of relevant variables, the structure of the relationship between them and the behaviour of economic agents.

The rapid adoption of approaches based on random or quasi-random variation stems in part from dissatisfaction with both these preceding methods. Structural methods appear to be constrained by the need to make simplifying assumptions that are compatible with analytically producing an estimable model, but that may appear implausible, or at the least are not independently verified. On the other hand, non-structural regression methods seem unlikely to produce estimates of causal effects given the many possible relations between the variables of interest and many other, observed and unobserved, factors. This seeming inability to identify causal effects under plausible restrictions led to a period in which many econometricians and applied economists abandoned reference to causal statements - a point emphasised in particular by Pearl (2009), but see also Heckman (2000, 2008).

In this context, the further development and wider understanding of econometric methods for analysis using experimental, or quasi-experimental (Angrist and Krueger, 2001), data presented the promise of reviving causal analysis without needing to resort to seemingly implausible structural models. randomization, or variation from it, potentially severs the connection between the causal variable of interest and confounding factors. Many expositions of experimental methods cite LaLonde (1986)'s paper showing the superiority of experimental estimates to ones based on various quasi-structural assumptions in the case of job market training programmes.[2] Banerjee (2007) described randomized trials as the "gold standard" in evidence and Angrist and Pischke (2010) state that the adoption of experimental methods has led to a "credibility revolution" in economics. Such methodological claims have, however, been the subject of a great deal of criticism.

[2]We refer to these as 'quasi-structural' since in most cases they are not based on full structural models but rather specific assumptions on underlying structural relationships that, theoretically, enable identification using observational data.

Within economics, Heckman and Smith (1995), Heckman and Vytlacil (2007a), Heckman and Urzua (2010), Keane (2005, 2010a,b), Deaton (2008, 2009, 2010), Ravallion (2008, 2009), Leamer (2010) and Bardhan (2013), among others, have argued that the case for experimental methods has been overstated and that consequently other methods - particularly structural approaches (Rust, 2010) - are being displaced by what amounts to a fad. See also the contributions in Banerjee and Kanbur (2005), which anticipate some of the later points of contention. The more extreme proponents of these experimental methods have sometimes been referred to as 'randomiztas' (Deaton (2008), Ravallion (2009)).

Some of the concerns raised by Deaton are based on detailed work in the philosophy of science by Cartwright (2007, 2010). There also exists an active literature in philosophy on the so-called 'evidence hierarchies' developed in medicine; the notion that some forms of evidence are inherently superior to others. In standard versions of such hierarchies randomized evaluations occupy the top position. This is primarily due to the belief that estimates from randomized evaluations are less likely to be biased (Hadorn, Baker, Hodges, and Hicks, 1996) or provide better estimates of 'effectiveness' (Evans, 2003). Nevertheless, a number of contributions have critically addressed the implicit assumption that the idea of a 'gold standard' - a form of evidence unconditionally superior to all others - is coherent. Specifically, Concato, Shah, and Horwitz (2000) ask: is this view of evidence conceptually sound and is it confirmed empirically? Most of these references come from the medical literature in which randomized trials have been a preferred method for causal inference long before their adoption in economics. The generic problem of integrating different forms of evidence has not yet been tackled in any systematic fashion in economics, though studies delineating what relationships/effects various methodological approaches are identifying (Angrist (2004), Heckman and Vytlacil (2005, 2007b), Heckman and Urzua (2010)) may provide one theoretical basis for doing so. Nevertheless, some advocates of these methods continue to argue strongly that, "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top" (Imbens, 2010: 10).

1.1 Randomized evaluations

The great advantage of randomized evaluations is that they offer the prospect of simple estimation of individual causal effects by removing the risk of bias from confounding factors that plagues analysis using observational data. Introducing some formal notation, Y_i is the outcome variable for individual i, which becomes $Y_i(1) = Y_{1i}$ denoting the outcome state associated with receiving treatment ($T_i = 1$) and $Y_i(0) = Y_{0i}$ denoting the outcome state associated with not receiving

treatment ($T_i = 0$). The effect of treatment for any individual is $\triangle_i = Y_{1i} - Y_{0i}$.[3] This formulation can be seen to be based on a framework - the more complete version of which is known as the Neyman-Rubin model after Neyman (1923) and Rubin (1974) - of counterfactuals, since in practice the same individual cannot simultaneously be observed in treated and non-treated states. Holland (1986) is a key early review of this framework.

Assume we are interested in the average effect of treatment ($E[Y_{1i} - Y_{0i}]$).[4] To empirically estimate this, one might consider simply subtracting the average outcomes for those receiving treatment and the untreated. One can rewrite this difference as:

$$\begin{aligned} E[Y_i|T_i = 1] - E[Y_i|T_i = 0] =& \{E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1]\} \\ &+ \{E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 0]\} \end{aligned} \quad (1)$$

The second term, representing the difference between potential outcomes of treatment recipients and non-recipients in the non-treated state represents 'selection bias', the extent to which treatment receipt is associated with other factors that affect the outcome of interest. An ideal experiment in which individuals are randomly allocated into treatment and control groups, with no effects of the experiment itself beyond this, ensures that on aggregate individuals' potential outcomes are the same regardless of treatment receipt so $E[Y_{0i}|T = 1] = E[Y_{0i}|T = 0]$. A randomized evaluation can therefore estimate an unbiased effect of treatment on those who were treated (the first term on the right-hand side of (1)). Given randomization this is equal to the average treatment effect, as desired.

Implicit in the above is the assumption that a given individual's treatment effect does not vary with the treatment of others.[5] Furthermore, as Heckman and Smith (1995) point out, where there is selection into the experimental sample from the broader population randomization enables estimation of the average treatment effect in the sample by balancing this second form of selection bias across the treatment and control groups rather than eliminating it. Lastly, as various authors have pointed out, the above result need not hold for other properties of the

[3] In subsequent analysis, following a notational convention in some of the literature, $\triangle$ is used to signify a treatment effect and is subscripted accordingly if that is anything other than $Y_{1i} - Y_{0i}$.

[4] Note that in some treatments - see for instance Imbens (2004) - the 'i' subscript is used to denote sample treatment effects as opposed to those for the population. This distinction is not important for the above discussion but in later analysis we, instead, distinguish between populations using appropriately defined dummy variables.

[5] Often referred to, following Rubin (1980), as the 'stable unit treatment value assumption' (SUTVA).

treatment effect distribution, such as the median, unless one makes further assumptions. For instance, if one assumes that the causal effect of treatment is the same for all individuals ($\triangle_i = \triangle_j, \forall i$ and j), then the median treatment effect can also be estimated in the above fashion. That assumption, however, appears excessively strong and allowing for the possibility that treatment effect varies across individuals raises a host of other - arguably more fundamental - concerns, which we discuss in somewhat more detail below.

Nevertheless, the average effect is often of interest. To connect the above to one popular estimation method, least squares regression, one can begin by writing the outcome as a function of potential outcomes and treatment receipt:

$$\begin{aligned} Y_i &= (1-T)Y_{0i} + TY_{1i} \\ &= Y_{0i} + T(Y_{1i} - Y_{0i}) \end{aligned}$$

Writing the potential outcomes as:

$$\begin{aligned} Y_{0i} &= \alpha + u_{0i} \\ Y_{1i} &= \alpha + \tau + u_{1i} \end{aligned}$$

where $u_{0i} = Y_{0i} - E[Y_{0i}]$, and similarly for u_{1i}, and τ is then the average treatment effect ($\bar{\triangle}$). We can then write the previous equation as:

$$Y = \alpha + \tau T + [T(u_1 - u_0) + u_0]$$

taking expectations:

$$E[Y|T] = \alpha + \tau T + E[T(u_1 - u_0)] + E[u_0]$$

we have $E[u_0] = 0$ by definition and randomization ensures that the second last term is zero, so:

$$E[Y|T] = \alpha + \tau T \tag{2}$$

Equation 2 is just a conditional regression function, meaning that we can obtain an unbiased estimate of the average treatment effect through a least squares regression of Y on T. If there was selection bias then $E[T(u_1 - u_0)] \neq 0$, the regressor would be correlated with the error and a least squares estimate of τ would be biased.

1.2 Estimating average treatment effects conditional on covariates

The above discussion provides the basic rationale for the popular use of regression-based estimates of average treatment effects using data from randomized trials.

One can extend the analysis to somewhat weaker assumptions regarding random assignment that explicitly account for covariates. These in turn are the basis for contributions on *non-parametric* estimates of treatment effects. As we will see, some of the critical issues in that literature extend naturally to the question of external validity, so we briefly discuss these as a basis for subsequent analysis of that issue. Imbens (2004) and Todd (2006) are valuable surveys of these and related issues, providing extensive additional detail including on estimation of statistics of treatment effect distributions besides the mean.

A more general analysis than that outlined in the previous subsection would include covariates. In the case mentioned above where there is some selection bias, the weaker condition $E[T(u_1 - u_0)|X] = 0$ may hold. Rather than assuming that randomization ensures simple independence of potential outcomes from treatment (Y_{0i}, $Y_{1i} \perp\!\!\!\perp T_i$) it may be more plausible to assume that independence exists *conditional* on some covariates (X):

Assumption 1.1. *Unconfoundedness*

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp T_i|X \tag{3}$$

Unconfoundedness ensures that we can write the average treatment effect in terms of expectations of observable variables (rather than unobservable potential outcomes) conditional on a vector of covariates.[6] The probability of receiving treatment given the covariates (X) is known as 'the propensity score', written: $e(x) = Pr(T = 1|X = x)$. Where treatment is dichotomous: $e(x) = E[T|X = x]$. For a number of purposes it is useful to know a result by Rosenbaum and Rubin (1983) that unconfoundedness as defined above conditional on X implies unconfoundedness conditional on the propensity score. This has the notable advantage of reducing the 'dimensionality' of the estimation problem by summarising a possibly large number of relevant covariates into a single variable (Imbens, 2004).

In order to then obtain the preceding, desirable results under this weaker assumption, one also requires sufficient overlap between the distributions of covariates in the treated and non-treated populations:

Assumption 1.2. *Overlapping support*

$$0 < Pr(T = 1|X) < 1 \tag{4}$$

[6]Heckman, Ichimura, and Todd (1997) show that a weaker assumption can be used if the interest is in the effect of treatment on the treated, though Imbens (2004) argues that it is hard to see how this weaker form can be justified without also justifying the stronger unconfoundedness assumption - see also the discussion in Todd (2006).

This condition states that no covariate value, or combination of covariate values where X is a vector, perfectly predicts treatment receipt.

Three points about the above approach are particularly important for our later discussion of external validity. First, to implement it in practice a researcher must be able to accurately estimate the conditional average treatment effect for *every* realisation of X and T (denoted x and t), which in turn requires that these be represented in both treatment and control populations (the 'overlapping support' assumption) and with large enough sample size to enable accurate estimation.[7] Second, the unconditional average treatment effect is estimated by averaging over the distribution of x but that is often unknown and therefore requires further assumptions to make the approach empirically feasible. Finally, it is possible that both the above assumptions could be satisfied subject to knowledge of, and data on, the relevant conditioning variables even without experimental variation. In that case, which as a result is also often referred to as 'selection on observables', observational data is enough to secure identification of the average treatment effect. The experimental literature proceeds from the assumption that unconfoundedness, conditional or not, is - at the very least - more likely to hold in experimental data, a position which has some support from the empirical literature (see the previously mentioned paper by LaLonde (1986)) but is also contested.

1.3 Randomized evaluations: specific criticisms and defences

In its conditional formulation the formal case for experimental methods appears somewhat more nuanced, with experimental assignment increasing the likelihood of an unconfoundedness condition being satisfied. That in turn depends on a number of implicit assumptions about successful design and implementation of experiments as well as the broader applicability of such methods. Unsurprisingly, these are the issues on which many criticisms have focused. Table 1 summarises limitations to randomized evaluations that have been identified by critics and, in some cases, acknowledged by proponents of these methods.

[7] As various authors (Imbens (2004), Heckman and Vytlacil (2007a), Todd (2006)) have noted, where there is inadequate overlap in the support, identification can be obtained conditional on limiting the sample to the relevant part of the support. The substantive rationale for this is that it allows identification of *some* effect, but with the caveat that the restriction is otherwise ad hoc.

Table 1 – Criticisms of randomized or quasi-random evaluations †

Criticisms of randomized or quasi-random evaluations	
Limited applicability of method (Deaton (2010), Rodrik (2008), Ravallion (2008))	RCTs cannot address 'big questions' Many variables of interest are not amenable to deliberate randomization The question of interest is determined by method, rather than vice versa Policies often involve a combination of different interventions
Factors likely to confound experiments (Heckman and Smith (1995), Duflo, Glennerster, and Kremer (2006a))	Selection into the experimental sample The use of randomized assignment affects *ex ante* entry into the sample ('randomization bias') Individuals act to compensate for not receiving treatment ('substitution bias') Individuals in the control group respond to knowledge that they are not receiving treatment ('John Henry effects', may overlap with the above) Individuals' outcomes are affected simply by virtue of being observed ('Hawthorne effects')
Absence of ideal experiment means the ATE is not estimated(Heckman and Vytlacil (2005, 2007a))	Only identifies a 'local average treatment effect' (LATE) which is affected by the proportions of 'compliers' and 'non-compliers' The effect identified is a function of the 'marginal treatment effect' (MTE) which is affected by behavioural factors and treatment level
Limited relevance to other domains (Cartwright (2010), Keane (2010b,a), Manski (2013a))	Implementation details matter and in practice often vary There is an inherent trade-off between the use of experiments and generalisability of results The causal effect may differ for interventions implemented at a larger scale ('scale-up problem') We do not know *why* intervention worked/did not work (experiments are a 'black box') Experiments do not, on their own, allow for welfare analysis The treatment effect may be non-linear and therefore differ when the magnitude, or initial level, of a continuous treatment variable is different
RCTs are not conducive to learning(Heckman and Smith (1995), Keane (2010b), Deaton (2010), Rodrik (2008))	Provide information only on specific interventions Are inadequate for learning the underlying mechanism(s) No clear procedure for accumulating knowledge across experimental studies

† References are not intended to be exhaustive but rather to indicate particularly influential authors or those associated with specific criticisms.

To represent some of these concerns formally it is useful to distinguish between treated state, participation in a programme ($P \in \{0,1\}$) and participation in a randomized programme ($R \in \{0,1\}$), where $R = 1 \Rightarrow P = 1$ but not vice versa.[8]

$$\begin{aligned}
\text{Scale-up problem:} \quad & E(Y_{1i} - Y_{0i}) = (1/N)\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) = f(N) \\
\text{randomization bias:} \quad & E(Y_{1i}|T=1,R=1) \neq E(Y_{1i}|T=1,R=0) \\
\text{Hawthorne effect:} \quad & E(Y_{1i}|P=1,T=1) \neq E(Y_{1i}|T=1) \\
\text{John Henry effect:} \quad & E(Y_{0i}|P=1,T=0) \neq E(Y_{0i}|T=0)
\end{aligned}$$

There have been a variety of responses to the criticisms in Table 1 and we briefly survey some of the more important ones here, drawing to a significant extent on Banerjee and Duflo (2009), Imbens (2010) and Angrist and Pischke (2010) who provide some of the most detailed and cited expositions and defences of the use of randomized evaluations in economics.

First, it has been argued that many of the apparent limits on questions that can be meaningfully addressed with RCTs are a function of a lack of imagination. Angrist and Krueger (2001) suggest that creating experiments, or finding natural variation, to answer questions of interest, is "gritty work...[which requires] detailed institutional knowledge and the careful investigation and quantification of the forces at work in a particular setting" (Angrist and Krueger, 2001: 83). In a somewhat similar vein, Banerjee and Duflo (2008: 9) state that "experiments are...a powerful tool...in the hands of those with sufficient creativity". Second, the claim that experimental methods are particularly vulnerable to a trade-off between internal and external validity has been disputed. Banerjee and Duflo (2009) argue with reference to matching methods for observational data - which we discuss further below - that the same trade-off exists in such studies and without the advantage of a well-identified effect in a known population (as in experimental studies). Taking a stronger position, Imbens (2013) has argued, in disagreeing with Manski (2013a), that "studies with very limited external validity...should be [taken seriously in policy discussions]" (Imbens, 2013: 405). A partly complementary position has been to emphasise the existence of a continuum of evaluation methods (Roe and Just, 2009).

A popular position among RCT practitioners is that many concerns can be *empirically* assuaged by conducting more experimental and quasi-experimental eval-

[8]Here we partly follow the analysis by Heckman and Smith (1995).

uations in different contexts. Angrist and Pischke (2010), for instance, argue that "A constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge" (Angrist and Pischke, 2010: 23). The idea being that if results are relatively consistent across analyses then, for instance, this would suggest that the various concerns implying confounding or limited prospects for extrapolation are not of sufficient magnitude to be empirically important. This counterargument is particularly relevant for issues relating to external validity and we give it more detailed consideration in section 3.

A final point, made by critics and advocates, is that the use of randomized evaluations and formulation and estimation of structural models need not be mutually exclusive. Card, DellaVigna, and Malmendier (2011) classify experiments - evaluations ('field experiments') and lab-based experiments - into four categories based on the extent to which they are informed by theory: descriptive (estimating the programme effect); single model (interpreting results through a single model); competing model (examining results through multiple competing models); and, parameter estimation (specifying a particular model and using randomization to estimate a parameter/parameters of interest). They argue that there is no particular reason why experiments need be 'descriptive' and therefore subject to criticisms (Heckman and Smith (1995), Deaton (2010)) that they do little to improve substantive understanding. Those authors do, however, show that in practice a large proportion of the increase in experiment-based articles in top-ranked economics journals *is* due to descriptive studies. Ludwig, Kling, and Mullainathan (2011) make a related argument, that more attention should be directed to instances where economists feel confident in their *prior* knowledge of the structure of causal relationships so that randomized evaluations can be used to estimate parameters of interest.[9]

Many of the above criticisms of randomized trials can, in fact, be delineated by the two broad categories of internal and external validity. The former affect researchers' ability to identify the causal effect in the experimental sample and the latter the prospects of using estimated treatment effects to infer likely policy effects in other populations. While internal validity is the main concern of the experimental programme evaluation literature, in economics and elsewhere, the

[9]It is worth noting that while usefully expanding on the ways in which experiments can be employed, neither of these two analyses acknowledges the historical limitations of structural methods, "the empirical track record [of which] is, at best, mixed" (Heckman, 2000: 49). In short, while the claims made for descriptive randomized evaluations may be excessive, relating these more closely to theory simply reintroduces the concerns with structural work that partly motivated the rise in popularity of such methods.

issue of external validity is largely neglected. And yet by definition the usefulness of any estimate for policy necessarily depends on its relevance outside of the experiment. This concern is the focus of the present paper and the next section reviews the cross-disciplinary literature on the external validity of estimated treatment effects from randomized evaluations.

2 External validity of treatment effects: A review of the literature

The applied and theoretical econometric literatures that deal explicitly with external validity of treatment effects are still in the early stages of development. Here we provide an overview of the concept of external validity and contributions from different literatures. As noted above, there are currently two broad approaches to the evaluation problem in econometrics, albeit with increasing overlap between them. In what follows, the focus will be on critically engaging with the literature that builds on the Neyman (1923)-Rubin (1974) framework of counterfactuals and advocates the use of experimental or quasi-experimental methods in economics; Angrist and Pischke (2009) provide an accessible overview of this framework as applied to econometric questions, while Morgan and Winship (2007) use it for a broader discussion of causal inference in social science particularly in relation to the causal graph methods advocated by Pearl (2009). The alternative to this approach would be the framework of structural econometrics, but a correspondingly detailed assessment of that literature would go well beyond the scope of the present work. We will, however, note relevant insights from that literature in the analysis that follows.

Perhaps the earliest and best-known discussions of external validity in social science are in the work of Campbell and Stanley (1966) and Cook and Campbell (1979) on experimental and quasi-experimental analysis and design. Although not formally defined, the basic conception of external validity those authors utilise is that the treatment effect estimated in one population is the same as the effect that would occur under an identical intervention in another population. An alternative, though not mutually exclusive, conception of external validity concerns the extent to which the effect of one policy or intervention can be used to infer the effect of a related policy or intervention, whether in the same population or a different one. In reviewing the extant literature we will note contributions that have made preliminary efforts to address the question of predicting the effects of new policies. However, the problem of extrapolating the effect of the same programme from one context to another is of widespread interest and informative enough to merit

exclusive consideration, so that will be the focus of the analysis.

Operating within this conception of external validity, we now provide the first of a number of formal definitions of this concept. Adding to our previous notation, let D be a dummy equal to one for the population of policy interest and zero for the experimental sample. In what follows the focus is confined to the average treatment effect, which has been the focus of most contributions to the experimental literature, though the issues raised also apply to other properties of the treatment effect distribution. Given this we have:

Definition Simple external validity

$$E[Y_i(1) - Y_i(0)|D_i = 1] = E[Y_i(1) - Y_i(0)|D_i = 0] \tag{5}$$

The requirement of identical treatment effects, albeit in the aggregate, across contexts in equation (5) is strong and arguably unnecessarily so for many cases of interest. In subsections below we consider alternate approaches to, and formulations of, this concept. Three formal alternatives are suggested by different econometric literatures: external validity as a question of forecast accuracy; external validity as stability in policy decisions across contexts; and, external validity *conditional* on a vector of covariates. This last definition emerges from recent theoretical and empirical contributions on this subject in the experimental programme evaluation literature.

2.1 The medical literature on external validity [10]

One way of framing the debates on randomized evaluations discussed in section 1 is as a problem of assigning precedence to certain forms of evidence relative to others. A related problem is integrating different kinds of evidence. Both issues have been recognised in the medical literature for some time. Evans (2003) notes that the so-called 'evidence hierarchy' in medicine, with randomized controls trials at the top, goes back to Canadian guidelines developed in 1979. It is from this literature that the, now controversial, term 'gold standard' emerged. Authors differ on the interpretation of the hierarchy, with some suggesting that it is indicative of a (non-trivial) weighting of different sources of evidence while others see it as guiding a lexicographic process in which evidence only from the method highest on the hierarchy is considered. Given this, and that medical analogies are popular in methodological debates on RCTs in economics, it is somewhat instructive to consider developments in the medical literature.

[10] I am grateful to JP Vandenbroucke for drawing some of the references and arguments in this literature to my attention.

Mirroring some of the methodological debates in economics, two contributions to the medical literature by McKee, Britton, Black, McPherson, Sanderson, and Bain (1999) and Benson and Hartz (2000) caused controversy for suggesting that estimates from observational studies were not markedly different from experimental evaluations. This, in turn, prompted an editorial asserting that "the best RCT still trumps the best observational study" (Barton, 2000), while recognising that there ought to be some flexibility in relation to different kinds of evidence. Within these contributions, however, the *reasons* for the similarity across the different methods could only be the subject of speculation: the observational studies may have been successful in controlling for confounding factors, the randomized trials may have been poorly conducted or the problems studied may not have had the sources of bias that randomization is traditionally used to avoid. This reflects a broader problem that has perhaps been addressed more systematically in the econometrics literature: understanding conceptually what parameter a given randomized trial is estimating and why, therefore, it may differ from a parameter estimated in an observational study.

Parallel to such studies, in recent decades medical scientists and practitioners have increasingly expressed concerns about the external validity of randomized experiments. One particular area of interest has been selection of participants into the experimental sample. Unlike many of the experiments considered in the economics literature, medical RCTs often have strong, explicit exclusion and inclusion criteria. Falagasa, Vouloumanoua, Sgourosa, Athanasioud, Peppasa, and Siemposa (2010), for instance, review thirty RCTs relating to infectious diseases and argue, based on the authors' expertise, that many of these experiments exclude a significant proportion of patients that are treated by clinicians. That is problematic because such studies typically say little about external validity and it is left to clinicians to make a qualitative judgement as to whether, and how, the published results may be relevant for a given patient whose characteristics are not well-represented in the experimental sample. In statistics and econometrics this issue of 'adequate representation' of characteristics is dealt with formally via assumptions on the 'support' of relevant variables - an issue addressed in the next section.

In addition to *explicit* criteria, a number of studies have examined other reasons why patients and clinicians are hard to recruit into experimental samples. Ross, Grant, Counsell, Gillespie, Russell, and Prescott (1999) provide a survey of those contributions, noting that reasons for non-participation relate to decision-making by both the clinician and the patient. The decisions of both clinician and patient are affected by, among other factors: attitudes to risk; the possible costs (time,

travel, etc) imposed by the trial; preferences over treatment; perceived probability of success of the proposed intervention; and, experiment characteristics such as information provided and even the personality of the research/recruiter. The authors advocate gathering more information on reasons for non-participation. As Heckman and Smith (1995) note, such concerns in the medical literature go back at least as far as Kramer and Shapiro (1984), who noted markedly lower participation rates for randomized as opposed to non-randomized trials.

Besides selection problems, there are a variety of other factors that have been identified as likely to affect external validity of medical trials. Rothwell (2005a,b, 2006) has provided a number of influential discussions of the broader challenge where external validity is defined as, "whether the results [from randomized trials or systematic reviews] can be reasonably applied to a definable group of patients in a particular clinical setting in routine practice" (Rothwell, 2005a: 82). He notes that published results, rules and guidelines for designing and conducting clinical trials, treatment and medicine approval processes all largely neglect external validity, which is remarkable since ultimately it is external validity - here by definition - that determines the usefulness of any given finding (at least for clinicians). Besides the selection problem, he notes the following additional issues: the setting of the trial (healthcare system, country and type of care centre); variation of the effect by patient characteristics, including some that are inadequately captured and reported; differences between trial protocols and clinical practice; reporting of outcomes on particular scales, non-reporting of some welfare-relevant outcomes (including adverse treatment effects) and reporting of results only from short-term follow-ups. In relation to the debate regarding the merits of RCTs, Rothwell is strongly in favour of these over observational studies because of the likelihood of bias (failed internal validity) with the latter approach. Instead, his view is that a failure to adequately address external validity issues is limiting the relevance and uptake of results from experimental trials.

Dekkers, von Elm, Algra, Romijn, and Vandenbroucke (2010) take a somewhat different approach. Those authors make a number of key claims and distinctions:

- Internal validity is necessary for external validity;

- External validity (the same result for different patients in the same treatment setting) should be distinguished from applicability (same result in a different treatment setting);

- "The only formal way to establish the external validity would be to repeat the study in the specific target population" (Dekkers et al., 2010: 91).

The authors note three main reasons why external validity may fail: the official eligibility criteria may not reflect the actual trial population; there may be differences between the 'target population' and experimental population that affect treatment effects; treatment effects for those in the study population are not a good guide for patients outside the eligibility criteria. They conclude that external validity, unlike internal validity, is *too complex to formalise* and requires a range of knowledge to be brought to bear on the question of whether the results of a given trial are informative for a specific population.

In summary, the medical literature is increasingly moving away from rigid evidence hierarchies in which randomized trials always take precedence. Many studies are raising challenging questions about external validity, driven by the question asked by those actually treating patients "to whom do these results apply?" (Rothwell, 2005a). Medicine, therefore, can no longer be used to justify a decision-making process that is fixated on internal validity and the effects derived from randomized trials without regard to the generalisability of these results. However, methodological contributions to that literature have also expressed scepticism about the prospect of using formal methods to establish external validity.

2.2 Philosophers on external validity

The discussion in section 1 noted the contribution by philosopher Nancy Cartwright to the debate in economics on the merits of RCTs. Nevertheless, Guala (2003) notes that, "Philosophers of science have paid relatively little attention to the internal/external validity distinction." (Guala, 2003: 1198). This can partly be explained by the fact that many formulations of causality in philosophy do not lend themselves to making clean distinctions between these two concepts.

Cartwright, for example, advocates a view of causality that, in economics, bears closest relation to the approaches of structural econometricians (Cartwright, 1979, 1989, 2007). Structural approaches are more concerned with correct specification and identification of *mechanisms* rather than *effects*, whereas the literature developed from the Neyman-Rubin framework orients itself toward 'the effects of causes rather than the causes of effects' Holland (1986). Cartwright (2011a,b) makes explicit the rejection of the internal-external validity distinction, arguing that "'external validity' is generally a dead end: it seldom obtains and...it depends so delicately on things being the same in just the right ways" (Cartwright, 2011b: 14). She also differentiates between the external validity of effect size

and external validity of effect direction, arguing that both "require a great deal of background knowledge before we are warranted in assuming that they hold" (Cartwright, 2011a). Broadly speaking, Cartwright is sceptical of there being any systematic method for obtaining external validity and is critical of research programmes that fail to acknowledge the limitations and uncertainties of existing methods.

Nevertheless, not all philosophers take quite so pessimistic a view. Guala (2003), with reference to experimental economics which we discuss next, argues for the importance and usefulness of *analogical reasoning*, whereby populations of interest are deemed to be 'similar enough' to the experimental sample. Another notable exception is Steel (2008)'s examination of extrapolation in biology and social science. Steel's analysis is perhaps closer to Cartwright's in emphasising the role of mechanisms in obtaining external validity. Specifically, Steel advocates what he calls 'mechanism-based extrapolation'. In particular, he endorses (Steel, 2008: 89) a procedure of *comparative process tracing*: learn the mechanism (e.g. by experimentation); compare aspects of the mechanism where we expect the two populations to be most likely to differ; if the populations are adequately similar then we may have some confidence about the prospect of successful extrapolation.

The above proposals are not formalised in any way that would render them directly useful in econometrics. In relation to Steel's proposals one might note - following Heckman (2000)'s review of 20th century econometrics - that there has not been a great deal of success in identifying economic mechanisms. Nevertheless, as in the case of medicine we will see that the themes of similarity and analogies have formal counterparts in the econometric literature. Much of Guala's analysis of the validity issue has referred specifically to the case of experimental economics and it is to that literature that we now turn.

2.3 External validity in experimental economics

While the concern of this paper is 'experimental programme evaluation' and its role in informing policy, a related area of economics in which the issue of external validity has been explored in more detail is experimental economics. The majority of studies in that sub-discipline to date have been concerned with testing various hypotheses concerning agent behaviour, either of the choice theoretic or game theoretic variety. The motivation may be the testing of a specific prediction of a formal model of behaviour, but could also involve searching for empirical regularities premised on a simple hypothesis (Roth, 1988). The majority of these experiments have been conducted in what one might call laboratory settings, where

recruited participants play games, or complete choice problems, that are intended to test hypotheses or theories about behaviour and "the economic environment is very fully under the control of the experimenter" (Roth, 1988: 974). One famous example is the paper by Kahneman and Tversky (1979) in which experimental results revealed behaviour that violated various axioms or predictions of expected utility theory.

The main criticism of such results, typically from economic theorists, has been that the laboratory environment and the experiments designed for it may not be an adequate representation of the actual context in which individuals make economic decisions (Loewenstein (1999), Sugden (2005), Schram (2005), Levitt and List (2007)). One aspect of this emphasised by some authors (Binmore (1999)) is that behaviour in economic contexts contains important dynamic elements, including learning, history dependence and repetition. 'One-shot' experiments may, therefore, not be identifying behaviour that is meaningful on its own. Another is that subjects may not be adequately incentivised to apply themselves to the task, a criticism that has particularly been made of hypothetical choice tasks. Furthermore, participants have traditionally been recruited from among university students and even when drawn from the broader population are rarely representative.

Given our preceding definition of external validity it should come as no surprise that many of the above criticisms have been framed, or interpreted, as statements about the limited external validity of laboratory experiments. Loewenstein (1999: 25), arguing from the perspective of behavioural economics suggests that this is "the dimension on which [experimental economists'] experiments are particularly vulnerable" and raises some of the above reasons to substantiate this view. By contrast, Guala and Mittone (2005) argue that the failure of external validity as a generic requirement is 'inevitable'. Instead, they argue that experiments should be seen as contributing to a 'library of phenomena' from which experts will draw in order to determine on a case-by-case basis what is likely to hold in a new environment. A somewhat different position is taken by Samuelson (2005) who emphasises the role that theory can and, in his view, should play in determining how and to what contexts experimental results can be extended.

One response to the previous criticisms - and therefore indirectly concerns about external validity - has been to advocate greater use of 'field experiments' (Harrison and List (2004), Levitt and List (2009), List (2011)), the argument being that the contexts in which these take place are less artificial and the populations more representative. Depending on the research question and scale of the experiment, some such studies begin to overlap with the experimental programme

evaluation literature. Another, related, response is to advocate replication. As Samuelson (2005: 85) puts it, an "obvious observation is that more experiments are always helpful". The argument here is that conducting experiments across multiple, varying contexts will either reveal robustness of the result or provide variation that may assist in better understanding how and why the effect differs. Something like this position underpins the systematic review and meta analysis literatures, in which the results from different studies of (approximately) the same phenomenon are aggregated to provide an overarching finding.

The nature of the external validity challenge is different for experimental economics because while researchers appear to have control over a broader range of relevant factors, manipulation and control of these can potentially lead to to the creation of contexts that are too artificial and therefore the relevance of results obtained becomes questionable. Perhaps the most relevant point for our purposes is that no systematic or formal resolution to the external validity challenge has yet been presented in the experimental economics literature.

2.4 The programme evaluation and treatment effect literature

Although there are a number of alternative formulations within economics that are effectively equivalent to the notion of external validity, the issue - as formulated in the broader statistical literature - has arisen primarily in relation to experimental work. Remarkably, despite Campbell and Stanley (1966) and Cook and Campbell (1979)'s work, which itself was reviewed in one of the earliest and most cited overviews of experimental methods in programme evaluation by Meyer (1995), the external validity challenge has not been dealt with in the experimental evaluation literature in any detail. As Rodrik (2008: 20) notes, "considerable effort is devoted to convincing [readers] of the internal validity of the study. By contrast, the typical study based on a randomized field experiment says very little about external validity." More specifically, the lack of *formal* and rigorous analysis of external validity contrasts markedly with the vast theoretical and empirical literatures on experimental or quasi-experimental methods for obtaining internal validity. This disjunct continues to be the basis for disagreements between contributors to the field; see for instance the recent exchange between Imbens (2013) and Manski (2013b).

From the perspective of practitioners, and guides for practitioners, Banerjee and Duflo (2009) and Duflo, Glennerster, and Kremer (2006b) address the issue

of external validity informally.[11] As above, the authors discuss issues such as compliance, imperfect randomization and the like, which are recognised as affecting external validity *because* they affect internal validity. In addition, the authors note concerns regarding general equilibrium and scale-up effects (though not the possible non-linearity of effects in response to different levels of treatment intensity). Banerjee and Duflo (2009) deal with the basic external validity issue under the heading of 'environmental dependence', which can be separated into two issues: "impact of differences in the environment where the program is evaluated on the effectiveness of the program"; and, "implementer effects" (Banerjee and Duflo, 2009: 159-160).

Some empirical evidence on the latter has recently been provided by Allcott and Mullainathan (2012) and Bold, Kimenyi, Mwabu, Ngángá, and Sandefur (2013). Allcott and Mullainathan (2012) examine how the effect of an energy conservation intervention by a large energy company (OPower) - emailing users reports of consumption along with encouragement to conserve electricity - varied with the providers across 14 different locations. The first finding is that "there is statistically and economically signicant heterogeneity in treatment effects across sites, and this heterogeneity is not explained by individually-varying observable characteristics"(Allcott and Mullainathan, 2012: 22). Exploring this further, the authors find that the sites selected for participation in the programme were a non-random selection from OPower's full set of sites based on observable characteristics. In addition, the characteristics increasing the probability of participation were (negatively) correlated with the estimated average treatment effect. They conclude, however, that significant heterogeneity from unobservables remains and that therefore it is not possible to predict the effect of scaling-up the intervention with any confidence.

Bold et al. (2013) provide results on an intervention in Kenya that involved the hiring of additional contract teachers. An experiment embedded in a larger government programme randomized 192 schools into three different groups: those receiving a contract teacher via the government programme; those receiving the teacher via an NGO; and, the control group. They find that while the NGO-managed intervention had a positive effect on test scores, the same basic intervention when implemented by government had no significant effect. Using the geographical distribution of schools from a national sampling frame, Bold et al.

[11] Angrist and Pischke (2009) provide a guide to obtaining internally valid estimates and complications that arise in doing so and Morgan and Winship (2007) similarly focus on questions of identification using the framework of causal graphs, but with no substantive discussion of the generalisability of results.

(2013) also examine the heterogeneity of outcomes across location. They find no significant variation across space and therefore conclude that "we find no reason to question the external validity of earlier studies on the basis of their geographic scope"(Bold et al., 2013: 5). By contrast, both papers attribute differences in outcomes to implementing parties and obviously that constitutes evidence of a failure of external validity broadly defined.

In this review our interest lies, more narrowly, with the external validity question *abstracting from issues that compromise internal validity or similarity of the intervention across populations*. In this regard, Banerjee and Duflo (2009) correctly note that the basic problem arises from the fact that heterogeneity in the treatment effect across individuals means that it may well vary by covariates, which in turn may vary across contexts. How to address this? Those authors argue, in essence, for two approaches. First, researchers could use their expertise, theory or *ex ante* knowledge of populations to determine whether the population of policy interest is similar enough for the original experimental result(s) to carry-over to this new context. Conceptually this bears a close resemblance to the 'analogical reasoning' approach advocated in philosophy by Guala (2005). As they acknowledge, however, this is - by economists' standards at least - 'very loose and highly subjective'. The second, more objective, approach is to replicate studies across different contexts. The authors argue that this indicates whether results generalise and allows knowledge to accumulate on specific kinds of interventions. Duflo et al. (2006b) make a similar argument, but in addition recognise that "as we cannot test every single permutation and combination of contexts, we must also rely on theories of behavior that can help us decide whether if the program worked in context A and B it is likely to work in C" (Duflo et al., 2006b).

The relevance of covariates to external validity concerns further reinforces the sense that, as has already been noted, the definition of simple external validity in (5) is too strong to be useful. Before turning to the forecasting and decision-theoretic definitions that we discuss below, and which are focused on the final use of estimates, it is important to note that a more subtle statistical definition has been developed in the programme evaluation literature. This states that an estimate has external validity if it can be used to predict the average treatment effect, which may be different, in another population *given a set of observable covariates*. In econometrics this definition has been formalised by Hotz, Imbens, and Mortimer (2005), who refer to it as *conditional external validity*. Define the treatment as before (T) and the relevant covariate as W, then:

Definition Conditional external validity

$$E[Y_i(1) - Y_i(0)|D_i = 1] \\ = E_W[E[Y_i|T_1, D_i = 0, W_i] - E[Y_i|T_0, D_i = 0, W_i]|D_i = 1] \quad (6)$$

In words, this second definition states that: the average treatment effect in the population of policy interest (on the left-hand side) can be expressed in terms of an expectation of the covariate-varying treatment effect in the experimental sample ($D_i = 0$) taken across the covariate (W) distribution in the population of interest ($D_i = 1$).

Hotz et al. (2005) show that given independence of treatment assignment and outcomes in the experimental sample ($T_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|D_i = 0$), two further conditions are sufficient for (6) to hold. First, independence of 'location' from outcomes conditional on a set of covariates:

Assumption 2.1. *Location independence*

$$D_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|W_i \quad (7)$$

Second, *overlapping support* of the relevant controls/covariates:

Assumption 2.2. *Overlapping support*

$$\text{For all } w,\ \delta < Pr(D_i = 1|W_i = w) < 1 - \delta, \\ \text{for some } \delta > 0 \text{ and for all } w \in W \quad (8)$$

Location independence states that potential outcomes (under treatment or control) do not vary across locations except as a result of differences between individuals in values of the covariates in W. Assumption 2.2 states that there is a non-zero probability of being in either location for any realised values of the covariates ($W_i = w$). Within these two conditions are a number of implicit assumptions, discussed by Hotz et al. (2005), such as the assumption of identical treatment across context and no macro effects (existence of important factors that have little or no variance *within* the populations).

While (6) is simply a formal result, Hotz et al. (2005) make it clear that the intention is to show how a researcher might go about estimating the likely effect of treatment in a population of interest based on estimated treatment effects in an experimental sample. From this perspective, the expression implies that to proceed non-parametrically one would estimate the treatment effect across the

distribution of the covariate (W) in the experimental sample and reweight this to account for the distribution of W in the population of interest. In the next section we expand on this point and suggest that such an approach provides a set of very clear formal requirements for obtaining external validity, comparable to the well-known sets of alternative assumptions that must be satisfied to obtain *internal* validity.

A related contribution to the literature is the analysis by Angrist and Fernandez-Vál (2010, 2013), which examines the extrapolation/external validity problem when estimating a local average treatment effect (LATE). What separates that analysis from Hotz et al. (2005) is, following the LATE-ATE distinction - that observed covariates are assumed to capture the characteristics that determine compliance.

While Hotz et al. (2005) and Angrist and Fernandez-Vál (2013) describe some ways in which an empirical analysis can be based on systematic comparisons across populations, no detailed analysis is provided of the implications of the above criteria for empirical practice in general. We will expand on that issue in the next section but first we examine alternative approaches to, and conceptualisations of, external validity within econometrics.

2.5 The structural approach to programme evaluation

While Samuelson (2005) advocates the use of theoretical models to guide extrapolation of results in experimental economics, within the programme evaluation literature there already exists a well-developed body of work with a similar motivation. This builds on the earlier structural econometrics literature discussed previously. Heckman and Vytlacil (2007a,b) and Heckman and Abbring (2007) provide an unparalleled overview and development of this body of work and therefore we refer primarily to those surveys, which contain extensive references to specific contributions.[12] In doing this it is important to clearly distinguish between using experiments to test theories, as is often the case in experimental economics, as opposed to using theories to inform the estimation and extrapolation of parameters.[13]

[12]Heckman and Vytlacil (2005) in fact contrast the approach they develop - discussed further below - with the experimental *and* structural literatures. It is fairly clear, however, that their approach is essentially an extension of the structural literature and therefore this distinction largely disappears in the later survey papers (Heckman and Vytlacil, 2007a,b).

[13]Duflo et al. (2006b: 70-75), as one example, conflate these two issues, so that a discussion which is ostensibly about using theory to extrapolate estimated effects deals primarily with using experiments to test theoretical predictions.

Heckman and Vytlacil (2007a) make a number of pointed distinctions. The first is between econometric and 'statistical' approaches to causal inference, the first of which they characterise by the specification and estimation of structural models while the latter is described as being oriented towards experimental identification of causal relationships. The authors criticise the experimental literature for: confusing the econometric problems of identification and estimation; not systematically addressing selection into, or compliance with, experiments; largely ignoring the welfare effects of policies; neglecting, or being unable to address, the problem of forecasting policy effects; and, promoting an analytical framework in which knowledge cannot accumulate.

The primary difference between the structural approach and the one based on randomized experiments is that structural econometric models, "do not start with the experiment as an ideal but start with well-posed, clearly articulated models for outcomes and treatment choice where the unobservables that underlie the selection and evaluation problem are made explicit" (Heckman and Vytlacil, 2007a: 4835). It is precisely for this reason that - as alluded to in discussion of philosophical contributions - the conceptual distinction between internal and external validity is not as valuable in the case of structural modelling; *if* we are prepared to assume the correctness of a *full* structural model then identifying the parameter(s) of interest necessarily implies the ability to forecast the effect in other populations given data on the relevant variables. This applies also to causal analysis using directed graphs, as described by Pearl (2009).

There are close conceptual similarities between the view of econometrics advocated in Heckman and Vytlacil (2007a,b) and Cartwright (1989)'s philosophical theory of causal inference. Heckman and Vytlacil (2007a) state, following Marschak (1953), that "The goal of explicitly formulated econometric models is to identify *policy-invariant* or *intervention-invariant* parameters that can be used to answer classes of policy evaluation questions" (Heckman and Vytlacil, 2007a: 4789). Cartwright's theory, going back to Cartwright (1979), is based on a notion of stable 'capacities', the identification of which is required for reliable causal prediction. Unsurprisingly, then, there is an appreciable amount of conceptual overlap between criticisms of the *randomizta* approach to causal inference in the philosophy and structural econometrics literatures. Arguably the key difference is that the structural literature almost uniformly proceeds on the assumption that time-, policy- and intervention-invariant parameters exist for questions of interest, whereas this is left as an open question in the philosophy literature.

It is important to note that while the basic rationale for the structural approach is premised on use of full economic models of the phenomena of interest, such models rarely exist and when they do are not - in the form in which theorists specify them - estimable. Structural econometrics therefore typically uses pared-down models of economic relationships and optimising behaviours, which in turn are adapted in such a way as to make them relevant to estimation. The structural econometric approach begins with the specification of an explicit econometric model of individual choice, often referred to as a *latent index model* - also called 'the Roy model' (Roy, 1951). Heckman and Robb (1985) is an important early discussion of this model in the context of programme evaluation. In its full specification that allows, in fact requires, the specification of individual constraints (broadly defined), utility function and characteristics affecting the outcome of interest. This in turn allows, theoretically, analysis of *ex ante* versus *ex post* outcomes of an intervention, selection effects, welfare analysis and behavioral responses to interventions.

The general latent index model extends the standard treatment effect framework by simply modelling the participation decision explicitly.[14] Note that we now introduce a set of variables Z, such that $X \subseteq Z$.[15] First, assume there exists some cost of receiving treatment: $C = \mu_C(Z) + U_C$. An individual's gain from treatment is then: $Y_1 - Y_0 - C$. They will then select into treatment if this is positive. We can rewrite the potential outcomes as:[16]

$$Y_1 = \mu_1(X) + U_1$$
$$Y_0 = \mu_0(X) + U_0$$

One can write the generic index model of participation as:

$$T_i = \begin{cases} 1 & \text{if } \mu_T(Z) - U_T \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Where for the preceding version of the 'Roy model': $\mu_T(Z) = (\mu_1(Z) - \mu_0(Z) - \mu_C(Z))$; and, $U_T = (U_1 - U_0 - U_C)$.

[14]There are various ways of presenting this, with minor notational differences and differing levels of generality, but here we follow the presentation of Heckman and Vytlacil (2005).

[15]In this literature Z is sometimes referred to as a set, or vector, containing the variables in X and at least one additional variable, while at the same time Z is used to denote the additional variable(s) in question without any change in font or notation. We follow this, occasionally confusing, convention.

[16]More general representations write these in nonseparable form.

This extension has two significant implications. First, *if* we allow for the possibility of selection into treatment and control groups, or selective compliance with assignment, then the latent index model provides a basis - in economic theory - for analysing the implications of different kinds of selection for various estimators of treatment effects. Second, explicitly recognising the relationship between individual choice and treatment may lead to a reconsideration of what it is that researchers wish to estimate. As discussed in Heckman and Abbring (2007), such models can - theoretically at least - be extended further to account for social interaction and general equilibrium effects. In contrast, most analysis conducted within the treatment effect framework requires that individual treatment effects are not influenced by the treatment receipt of others - Rubin (1980)'s 'stable unit treatment value assumption' (SUTVA).

The second point emerges most clearly from the work of Heckman and Vytlacil (2005), described also in Heckman and Vytlacil (2007a,b) and summarised in Todd (2006). Within the latent index framework it is possible to derive representations of the treatment effects estimated through the experimental or quasi-experimental approach that locate these in relation to theoretical representations of individual choice and selection processes.[17] Specifically, Heckman and Vytlacil (2005) propose a new concept they call the 'marginal treatment effect' (MTE):

$$\triangle_{MTE}(X) = E[Y_{1i} - Y_{0i}|X = x, U_T = u]$$

Heckman and Vytlacil (2005) provide a useful way of thinking about this as representing the mean gain $(Y_1 - Y_0)$ from treatment for individuals with characteristics X who would be indifferent about treatment receipt if they were exogenously assigned a value z for some (instrumental) variable such that $u(z) = u$. The average treatment effect can then be written as:

$$\triangle_{ATE}(X) = \int_0^1 E[Y_{1i} - Y_{0i}|X = x, U_T = u]dU_T$$

The average effect of treatment on the treated, as well as the local average treatment effect, can similarly by written as functions of the MTE. The dependence on unobservable factors (u) affects the interpretation of these effects. The authors argue that this approach unifies the treatment effect and structural econometric literatures, but with the advantage of using somewhat weaker assumptions than the latter. There are notable connections with the assumptions used in the treatment

[17] Heckman and Vytlacil (2007a,b) argue that since randomization is in fact an instrument of a particular sort, from a structural perspective the distinction between random and quasi-random variation is largely unnecessary.

effect literature. The framework developed in Heckman and Vytlacil (2005) also invokes an unconfoundedness assumption - phrased more generically in terms of instrumental variables (of which randomized assignment can be seen as a special case) - and an assumption of overlapping support, mirroring those in (3) and (4).

As noted by Heckman and Vytlacil (2005), where individuals comply with experimental assignment, either because they do not differ on unobservable factors or because these do not - for whatever reason - affect individuals' behaviour in relation to treatment, all the different treatment effects (ATE, MTE, ATT and LATE) are equal. Since our analysis is interested in external validity absent imperfect compliance, this simply confirms that the MTE - as with the broader literature based on latent index models - is not directly relevant to our concerns here; our interest is in external validity under perfect compliance.

This is not to say that the MTE is irrelevant to the external validity problem in general. To the contrary, it provides the basis for a much more ambitious agenda. Heckman and Vytlacil (2007a) classify the policy evaluation problem into three types:

1. Evaluating interventions that have already taken place;

2. Forecasting the effect of an intervention that has already been conducted in another context;

3. Forecasting the effect of an intervention "never historically experienced".

The authors refer to problems 2 and 3 as relating to external validity. It should be clear, however, that problem 3 is more ambitious than the traditional definition of external validity we have adopted here - characterised by problem 2. Heckman and Vytlacil (2005) and Heckman and Vytlacil (2007a: 4801) take the position of many critics that both questions are effectively "ignored in the treatment effect literature". As Heckman and Vytlacil (2005) point out, the entire treatment effect literature has been oriented toward the problem of internal validity and therefore there is little formal guidance on the assumptions or requirements to obtain external validity. In that framework one needs to make some additional assumptions, beyond those typically invoked in the treatment effect literature, about *invariance*, *exogeneity* and *autonomy*. Policy invariance, loosely speaking, refers to the stability of the causal relationships across contexts. More specifically, it means that a change in policy "does not change the counterfactual outcomes, covariates or unobservables" (Heckman and Vytlacil, 2005: 685). Exogeneity, in this case,

concerns independence of the unobservables determining choice from observable characteristics. Autonomy requires that the policy does not affect relative aspects of the environment and essentially invokes a partial equilibrium framework. Although the MTE approach makes clear which theoretical distributions need to be estimated and some of the assumptions required to do so, that literature has yet to give any empirically feasible guidance on obtaining external validity.

Presumably seeing no need to do so, Heckman and Vytlacil (2005) do not provide an actual definition of external validity. For our purposes, and comparison with the other definitions above, we may use the MTE-based definition of the average treatment effect to define what one might call a 'structural' notion of external validity:

Definition Structural definition of external validity

$$\begin{aligned} & \int_0^1 E[\triangle_i | X = x, U_T = u, D = 1] dU_T \\ = & \int_0^1 E[\triangle_i | X = x, U_T = u, D = 0] dU_T \end{aligned} \tag{10}$$

The notable difference in this definition is the dependence on *unobservables* across the populations of interest.

Given the above one may wonder why economists, or indeed any researchers, wanting to conduct programme evaluations would adopt anything other than a structural econometrics approach. While possibly the most theoretically comprehensive framework for evaluation, structural econometrics is not without problems. Two in particular stand out. The first is theoretical: formulating a structural model requires extensive theoretical assumptions many of which are not, or cannot be, empirically verified. Manski (2000) notes, in relation to the earlier literature, that latent index models have not been uncontroversial and that "some researchers have regarded these models as ill-motivated imputation rules whose functional form and distributional assumptions lack foundation" (Manski, 2000: 431). The second reason, already noted, is empirical: the information required in order to estimate structural models is often unavailable. Heckman and Vytlacil (2007a: 4810) note four types of data required: private preferences; social preferences; *ex ante* distributions of outcomes in alternative states; and, *ex post* information regarding the relevant outcomes. Although the authors note that there exist literatures on the first two, there is little convincing evidence that satisfactory empirical derivation of preferences at any level has been achieved. It therefore remains an open question whether it is feasible to obtain data on all the relevant dimensions

since this in itself rests on contested theoretical assumptions. For example, there are now a wide range of competing models of choice in the theoretical microeconomics literature and as yet no consensus on which of these ought to be employed to infer well-ordered preferences (or even whether well-ordered preferences exist for all individuals).

Both issues explain, to some extent, why despite its own limitations the 'statistical' approach to programme evaluation has gained so much popularity in economics in recent decades. The unquestionably valuable contributions of the structural literature are to locate the effects estimated using experiments within a more general model of mechanisms and economic behavior, as well as revealing the strong implicit assumptions required for treatment effects from randomization to inform a decision-making process as framed by economic theory. In the analysis of the next section we will essentially ignore the complications that arise from considering choice-based compliance with treatment assignment, not because these are unimportant in general but because our objective is to isolate what are arguably even more basic challenges for external validity.

2.6 Decision-theoretic approaches to treatment effects and welfare

A strand of the theoretical literature (Heckman, Smith, and Clements (1997), Manski (2000), Dehejia (2005)) related to structural contributions on treatment effect estimation considers the implications of treatment effect heterogeneity for optimal policy decisions, where a "planner wants to choose a treatment rule that maximizes the population mean outcome" (Manski, 2000: 417). Following Manski (2000: 423-424), an individual j in population J has a treatment response function $y_j(\cdot) : T \rightarrow Y$. The policymaker needs to specify a treatment rule for each j but only has at their disposal a set of observable characteristics for each individual, $x_j \in X$. There is then a set of functions (treatment rules), $b \in B$ where $B : X \rightarrow T$, mapping characteristics to treatment assignment. Given the emphasis on the mean outcome, the problem of interest is:

$$\max_{b(\cdot)\in B} E\{y[b(x)]\}$$

An optimal treatment rule, b^*, is one that maximises expected outcomes conditional on individual characteristics:

$$b^*(x) = \underset{t\in T}{\operatorname{argmax}}\, E[y(t)|x],\ x \in X \tag{11}$$

There are perhaps two key considerations in this literature. The first is the nature of the decision maker's welfare function as defined over the full distribution of treatment effects. The above formulation is most compatible with a utilitarian social welfare function, but others - such as the Rawlsian welfare function in which the well-being of the worst off individual is maximised - will be associated with different optimal treatment rules. Particularly challenging is that some welfare functions depend on the full distribution of outcomes. The second critical issue is the information available to the decision maker from econometric analysis. In this regard, an important consideration - emphasised in particular by Manski (2000) - is the relevance of uncertainty and ambiguity in relation to estimated effects that arises from making, often unverifiable, estimating assumptions. In a constructive vein Manski (2011, 2013a) argues for greater recognition of the salience of identifying assumptions by, where possible, reporting appropriate bounds on estimated effects rather than simple point estimates. In many instances only strong assumptions produce informative bounds.

It is interesting, given our preceding discussion of the medical literature, that Manski (2000) gives as a practical example of the generic decision problem, the case of a medical practitioner in possession of reported results from a randomized trial who is considering whether to allocate treatment to specific patients. Dehejia (2005) similarly considers a case in which there is a caseworker with individual-specific information and a policymaker who decides whether to have a uniform treatment rule (all or no individuals given treatment), or to allow the caseworker discretion to decide. Such formulations raise interesting questions about the benefits of decentralisation versus central planning. Another notable aspect of the decision problem is that while the physician in Manski's example "has extensive covariate information...for the patients", the "medical journal articles that report the findings of clinical trials, however, do not usually report extensive covariate information for the subjects of the experiment" (Manski, 2000: 433). Allocation of treatment is most simple when there is no variation in the treatment effect across covariates, but when that is *not* the case an optimal decision requires covariate-specific information.[18] One complication emphasised by Manski is that in the presence of uncertainty regarding individual response functions - in other words, variation in response exists even among individuals with the same observed covariate values - more covariate information is always weakly beneficial; additional information never leads to a less optimal choice of treatment rule. Where there is ambiguity about responses this need not be true.

[18]Relatedly, while exclusion and inclusion criteria can be a downside of medical trials, they can also be (as also noted by Ravallion (2009)) desirable in as much as in some cases they reflect a tailoring of an experiment to the likely recipients.

As with the literature surveyed in the previous subsection, there is much to recommend the logic and theoretical insights of such contributions, even if they are often practically hard to implement or produce bounds on estimated effects that are very wide. In the analysis of the next section, however, it suffices to show how external validity may fail without actually formalising the policymaker's decision process in this manner. If certain factors imply that a given programme simply does not work in the population of interest, then the form of the social welfare function is obviously of secondary concern. This is not in any way to caricature what are thorough and subtle studies: both cited authors have also addressed external validity concerns as distinct from the decision making problem, as is made clear in Manski (2013a), as well as the contributions in Manski (2011, 2013b) and Dehejia (2013).

Where these contributions do have some relevance for our analysis is in *framing* the idea of external validity. The basic definition provided in (5) can be thought of as *statistical* in the sense that it is based on any numerical deviation in the ATE in the target population from that in the experimental sample.[19] From a policy perspective, it may make more sense to utilise an *operational* definition of external validity in which an estimated effect has external validity if an *ex ante* decision based on that effect would not change if the policymaker knew the extent to which it would differ in the population of interest. Conceptually one can think of this as a two-stage process: in the first stage a researcher obtains evidence (possibly covariate-specific) on the treatment effect in the experimental population ($D = 0$) and this is used to determine an optimal treatment assignment rule; in the second stage that assignment rule is implemented in the population of interest ($D = 1$), for which the treatment effect is not known. External validity in this instance means that the rule would not change even if we had evidence on the population of interest. Denote data on the two populations as information sets $\mathcal{I}_D,\ D \in \{0, 1\}$ and the policies chosen based on this information as $\hat{b}^*_D(x) \in \hat{B}$. We can then represent this as:

Definition External validity of policy decisions

$$\begin{aligned} \hat{b}^*_1(x, \mathcal{I}_1) &= \hat{b}^*_0(x, \mathcal{I}_0) \\ &= \hat{B} : \mathcal{I}_0, x_i \rightarrow t_i \in T \end{aligned} \tag{12}$$

Arguably the most common empirical case at present is the simple one in which the information obtained is limited to the average treatment effect in the

[19] I am grateful to Jacob Stegenga for emphasising the significance of this fact.

sample population and the policy decision is whether or not to administer treatment to the entire population of interest. The above then reduces to:

$$\begin{aligned}\hat{b}_1^*(x, \mathcal{I}_1) &= \hat{b}_0^*(x, \mathcal{I}_0) \\ &= \hat{B} : \triangle_{ATE}(D = 0) \rightarrow t \in \{0, 1\}\end{aligned} \tag{13}$$

The weaker definitions in (12) and (13) may be satisfied in many more cases than the stronger one in (5), since it is possible - for example - that $\triangle_{ATE}(D = 0) \neq \triangle_{ATE}(D = 1)$ but that nevertheless $\hat{b}_1^* = \hat{b}_0^*$. The former definition also captures the underlying interest in external validity as something more than a statistical artefact.

2.7 Forecasting for policy?

The basic challenge of external validity - whether an estimate in one population can be used to determine the likely effect in another - appears analogous to the problem of *forecasting*, which has preoccupied many econometricians working with time series data. Indeed, the conceptual similarity is so striking that it seems sensible to ask whether there is any meaningful distinction between the two concepts.

In the structural literature, Heckman and Vytlacil (2007a: 4790-4791) in particular have recognised this in their descriptive typology of three policy evaluation questions: evaluating the impact of "historical interventions"; forecasting the impact of interventions conducted in one environment in different ones; and, forecasting the impact of interventions "never historically experienced". As already noted, they refer to the first problem as internal validity and the second as external validity, although it is unclear whether the authors intend to thereby assert that external validity can be obtained despite a failure of internal validity (a point discussed further below). The appendix to that review outlines the structural approach to policy forecasting, noting that parameter invariance is necessary for all methods, overlapping support of relevant variables is necessary for non-parametric methods and that additive separability "simplifies the extrapolation problem". These issues hint at a fundamental obstacle to external validity that we address in the next section. One may also note that the issues of exogeneity, autonomy and invariance that are referred to by Heckman and Vytlacil (2005) and Heckman and Vytlacil (2007a) have been developed in some detail in the time series econometric literature - see for instance the extensive discussion and analysis in Hendry (1995).

The previous review of the optimal policy approach to programme evaluation emphasises that what matters for policy is the accuracy of the estimated treatment effect as an indicator of the likely policy effect, with the importance of deviations depending on the policymaker's welfare function. Similar considerations apply when using the rather less sophisticated approach of cost-benefit analysis: the question that arises is whether deviation of the effect in the population of interest may be of magnitude large enough to reverse the conclusions reached in a cost-benefit analysis. An identical concern has been investigated in a recent literature concerning forecast optimality and the definition of this relative to loss functions with different properties - see the review by Elliott and Timmermann (2008). Those authors note three key considerations in evaluating forecast success: the relevant (policymakers') loss function; the nature of the forecasting model (parametric, semi-parametric or nonparametric); and, what aspect of the outcome of interest is being forecast (point or interval). Given data (Z), an outcome of interest (Y) and a forecasting model/rule ($f(Z, \theta)$) defined over the data and set of parameters one can define the 'risk' (R) to a policymaker, with loss function $\mathcal{L}(f, Y, Z)$, associated with a particular forecast model as (Elliott and Timmermann, 2008: 9):[20]

$$R(\theta, f) = E_{Y,Z}[\mathcal{L}(f(Z, \theta), Y, Z)] \tag{14}$$

This representation assumes a point forecast and one way that literature differs from its programme evaluation counterpart is the use of a relatively simple loss function defined over only a single forecast and realisation for a given time period. By contrast, social welfare considerations require that the programme evaluation literature pays more attention to the distribution of outcomes across a population, even if in practice this is typically summarised in an average treatment effect and simple welfare function defined over this. Regardless, the above representation can, in theory, be used to derive an optimal forecast as one that minimises the risk (expected loss).

The most important differences between the forecasting and programme evaluation literatures are not so much related to underlying motivation but rather to data availability and method. The literature on forecast optimality typically makes no distinction between models based on their plausible identification of causal relationships. This agnosticism about the extent to which successful forecasting models need to capture the underlying causal relationships is a well-established position in time series econometrics (Hendry, 1995). As Elliott and Timmermann

[20]The loss function, $\mathcal{L}(f, Y, Z)$, is envisioned as a function "that maps the data, Z, outcome, Y, and forecast, f, to the real number line".

(2008: 4) put it: "Forecasting models are best viewed as greatly simplified approximations of a far more complicated reality and need not reflect causal relations between economic variables." While it is often claimed in the randomized evaluation literature that internal validity (unbiased estimation of causal relationships) is necessary for external validity (generalising results to other populations) the forecasting literature suggests that this assertion is not as obvious as is often suggested. Most forecasting models estimate the parameters of a model, in which variables are related across time, using historical data and then use the parameterised model to predict a future outcome even though it is recognised that the parameters are unlikely to represent unconfounded causal effects. External validity in the strong sense defined in (5) may not be possible, but even the most vocal advocates of RCTs do not appear to expect that condition to be satisfied (see for instance Angrist and Pischke (2010)). Weaker versions that resemble minimisation of criteria like (14) may, under certain circumstances, allow studies that lack internal validity to outperform those that do in forecasting outcomes in new populations.

As should be evident from the discussion in section 1, approaches that neglect the relationship between estimated models and the data generating process ('true structural equation') are considered untenable in microeconometrics. This is quite understandable given that the concern of much of the applied microeconometrics literature has been in identifying the relationships between specific variables, net of confounding by others. That need not, however, provide the best methodological basis for addressing the challenge of predicting the effects of policy interventions and some researchers outside economics (Pearl, 2009) have argued forcefully that a different paradigm is required. However, the contrast with the time series forecasting literature indicates also that the limited time periods available in most microeconometric datasets constrain prospects for a similar approach to developing forecasts; there is too little data available over too widely-spaced intervals to calibrate models based on forecasts. As a partly related issue, one may note that the question of parameter stability has been directly addressed - albeit not in any way resolved - in the forecasting literature, whereas even the most advanced theoretical literatures in microeconometrics have yet to tackle this problem in any meaningful way.

This comparison suggests that from a policy perspective there is no meaningful conceptual difference between external validity and forecast accuracy. The academic distinction arises from data availability, the definition of the welfare function over a population rather than a single outcome and the established focus of microeconometrics on identifying causal relationships.

2.8 Summary

The approaches to the basic external validity question in the areas surveyed each have their own favoured emphasis and, particularly within economics, formal frameworks for addressing the problem of transporting estimated effects from one population to another. In some instances these differences in emphasis draw attention to different possible definitions of the concept. Nevertheless, a number of common themes are discernible. First, that the vast majority of contributions consider it highly unlikely that simple external validity will hold for most questions and populations of interest. Second, that *similarity* between populations is fundamental to the extrapolation problem. Such similarities might be determined qualitatively, as in the method of 'analogical reasoning' advocated by some philosophers. What the issue of similarity brings to the fore in formal frameworks is the relevance of covariates and the characteristics of individuals. External validity can then be linked to assumptions of, and requirements for, overlapping supports of variables across populations. This is particularly interesting because similar assumptions are required for obtaining internal validity, but where the populations are the 'recipients' and 'non-recipients' of the treatment of interest. A final theme is the importance of *structure* for extrapolation, whether in the form of fully developed models or, at least, more detailed information on the nature of causal relations besides only estimated mean effects.

In the next section we present a simple framework in which to further consider these issues and attempt to draw some implications for making policy claims using estimates derived within the experimental tradition. By assuming perfect compliance with treatment assignment we remove the many complications introduced in the instrumental variables literature, including new developments there relating to selection and marginal treatment effects, and yet still find substantial barriers to extrapolation. Our analysis builds on Hotz et al. (2005), as do Allcott and Mullainathan (2012) who reach some similar conclusions - albeit with a rather more optimistic emphasis.

3 Interacting factors, context dependence and external validity

> *it is a very drastic and usually improbable postulate to suppose that all economic forces [produce] independent changes in the phenomenon under investigation which are directly proportional to the changes in themselves; indeed, it is ridiculous* Keynes (1939: 564)

One particular issue that remains neglected in the empirical literature utilising random or quasi-random variation to estimate policy-relevant causal relationships, is the connection between functional form and external validity. Theorists and practitioners have been well-aware of the generic challenge posed by *ex ante* ignorance of the form of the relationship between the explanatory and dependent variables since the founding of econometrics. However, as Heckman (2000: 55) notes, the *convenience* of separable econometric models meant that these have been the predominant focus even of structural econometricians. While important advances have been made in non-parametric estimation methods in recent decades (Matzkin, 2007), these address the internal validity issue and have little direct relevance - for reasons we discuss below - to the external validity problem. As we noted in section 1, critics of randomized evaluations - see for instance Keane (2010a) - have emphasised the possible importance of functional form for extending estimated treatment effects to instances where either the base level of a non-dichotomous treatment variable, or the magnitude of the change induced by treatment, is different. This is of course an important issue, but falls outside the focus of this study which is on external validity of the same policies across different environments. Holding the policy intervention constant, where does functional form matter for external validity?

The answer is that functional form matters where it connects other variables ('covariates') to the effect of treatment. Specifically, where the treatment variable interacts with other variables in producing variation in the outcome of interest, the values of those variables become important for external validity. Although many of the contributions surveyed in section 1 reference Campbell and Stanley (1966) or Cook and Campbell (1979), few - if any - note that those authors conceptualised threats to external validity as problems, first-and-foremost, of *interaction*. In their words:

> Since the method we prefer of conceptualizing external validity involves generalizing across achieved populations, however unclearly defined, we have chosen to list all of the threats to external validity in terms of statistical interaction effects (Cook and Campbell, 1979: 73)

The authors identify three different forms of interaction. The first, which they refer to as 'interaction of *selection* and treatment', concerns the possibility that the characteristics of those in an experimental sample are affected by the demands of participation. This to some extent captures the intuition of the choice-based latent variable approach discussed above in relation to structural econometric models. The second is 'interaction of *setting* and treatment', by which the authors seem to mean in particular the institutional environment (contrasting a bureaucracy with a university campus or military camp). The third possibility they consider is that *historical context* may interact with treatment to affect the outcome. In some sense, each of these 'threats' reflects a different mechanism by which an experimental sample may become unrepresentative along dimensions that have some bearing on the effect of treatment. This is most clear from the fact that the *solutions* Cook and Campbell (1979) propose to avoid, or remedy, failures of external validity primarily concern sampling methods - an issue we address further in section 3.3.

The remainder of this review utilises the idea of interaction between the treatment variable and other factors as a basis for structuring what we believe to be the basic challenges for external validity and providing an alternative perspective on other analyses that have identified those challenges.

3.1 Interactive functional forms and external validity

To represent the above concerns in econometric form we might simply extend the standard representation of potential outcomes provided in section 1 as follows. A dichotomous treatment variable, $T \in \{0, 1\}$, is associated with average effects τ_0 and τ_1 that are independent of covariates.[21] Consider two sets of covariates, X and W, which we assume for simplicity are independent of treatment assignment.[22] Furthermore, the effect of some covariates (W) on potential outcomes is itself dependent on treatment.

$$\begin{aligned} Y_{0i} &= \tau_0 + X_i\beta + W_i\gamma + u_{0i} \\ Y_{1i} &= \tau_1 + X_i\beta + W_i(\delta + \gamma) + u_{1i} \end{aligned}$$

Then we can write the average treatment effect as:

$$E[Y_{1i} - Y_{0i}] = (\tau_1 - \tau_0) + E[W_i|T = 1]\delta \tag{15}$$

[21] It is fairly straightforward to extend the analysis to the case where the treatment variable is not dichotomous, taking on values T_0 in the 'control group' and T_1 in the 'treatment group'.

[22] Note that this is not the same as the unconfoundedness conditions mentioned previously, which assume that assignment is independent of potential outcomes conditional on covariates.

That effect now depends, at least in part, on the mean value of the covariates (W) in the population. Similar formulations have recently been used in the context of discussions of external validity by Allcott and Mullainathan (2012) and Pritchett and Sandefur (2013), although without explicit recognition of the key role played by interactions that we develop here.

As a variation in the econometric model deployed there is nothing particularly remarkable about interactive functional forms. The simple functional form outlined above is a special case of the 'random coefficients model' (Hsiao (1992), Hsiao and Pesaran (2004)) and Angrist and Pischke (2009) describe it as "a straightforward extension" of the model in which the treatment effect is constant across individuals.[23] Following the same procedure as in section 1.1 we can write a conditional regression function that is simplified by the assumption of random assignment:

$$E[Y|T] = \tau_0 + T(\tau_1 - \tau_0) + TE[W|T]\delta + E[X|T]\beta + E[W|T](\delta + \gamma) \quad (16)$$

An estimate of the correct average treatment effect can be obtained by regressing Y on T and the covariate(s) W.

While the extension itself may be technically straightforward and have no insurmountable, or at least unknown, implications for *identification* of the average treatment effect, this is not true for extrapolation of such effects. To see this, consider taking the difference in the average treatment effects from the two populations:

$$E[\triangle|D = 1] - E[\triangle|D = 0] = (E[W|D = 1, T = 1] - E[W|D = 0, T = 1])\,\delta \quad (17)$$

Note that the preceding representation of potential outcomes implicitly assumed a 'basic' treatment effect (one that does not vary with values of covariates), $\tau_1 - \tau_0$, that is independent of population.

The expression in equation (17) implies a failure of the simple (non-conditional) definition of external validity in (5) if the mean of the covariate differs across the experimental ($D = 0$) and policy ($D = 1$) populations. Leamer (2010) makes essentially the same point, referring to W-type variables as 'interactive

[23] That model is sometimes referred to as 'the constant effects model' but this term has a different meaning in the context of panel data models.

confounders'.[24] In some of the broader social science literature, W variables are referred to as 'mediating' the causal effect of T. For the primary concerns of the two empirical studies mentioned above - Allcott and Mullainathan (2012) and Bold et al. (2013) - one could conceive of the interacting factor as either a dummy for implementer type, or a vector of partner organisation characteristics. And that does not, of course, exclude the possibility that many other factors - including some which are unknown or unobserved - may be relevant.

The basic scenario is therefore not encouraging. In some situations, however, it may be possible to obtain information on the distribution of the treatment effect *across* the values of W. Consider the simplest case where there is one, dichotomous interacting variable $W \in \{0, 1\}$ and the experiment allows us to identify $E[\triangle|W = 0, D = 0]$ and $E[\triangle|W = 1, D = 0]$, where:

$$\begin{aligned} E[\triangle|D = 0] =& Pr(W = 0|D = 0)E[\triangle|W = 0, D = 0] \\ &+ (1 - Pr(W = 0|D = 0))E[\triangle|W = 1, D = 0] \end{aligned} \tag{18}$$

If we then know the distribution of W in the target population, the average treatment effect of policy interest can be expressed in terms of these estimated values:

$$\begin{aligned} E[\triangle|D = 1] =& Pr(W = 0|D = 1)E[\triangle|W = 0, D = 0] \\ &+ (1 - Pr(W = 0|D = 1))E[\triangle|W = 1, D = 0] \end{aligned} \tag{19}$$

As some readers may already have noticed, (19) is simply a specific case of the result in Hotz et al. (2005), shown previously in (6). That result can therefore be seen as proposing a solution to the problem interactive relationships pose for external validity, as originally discussed by Campbell and Stanley (1966) and Cook and Campbell (1979). The specific requirements that emerge from this presentation of the problem and possible solution are listed in Table 2.

Requirement R3.2 corresponds to the overlapping support condition of Hotz et al. (2005), while R5 refers to their 'unconfounded location' assumption.[26]

[24] In the philosophy literature related issues have sometimes been referrred to as 'causal interaction' - see for instance Cartwright (1989) and Eells (1991).

[26] The authors also refer to R5 as the 'no macro-effects' assumption, but their explanation of macro effects suggests that these effects are only one reason why unconfounded location may fail rather than being equivalent to that assumption. Most obviously, differences on unobservable components of W would violate the unconfounded location assumption, but that has nothing to do with the variation in such variables within the relevant populations. One might add that Garfinkel, Manski, and Michalopolous (1992) use the term 'macro effects' differently to refer to issues such as the impact of social interaction on treatment.

Table 2 – Minimum empirical requirements for external validity
(assuming an ideal experiment, with no specification of functional form)

R1	The interacting factors (W) must be known *ex ante*
R2	All elements of W must be observable *and* observed in both populations
R3.1	Empirical measures of elements of W must be comparable across populations
R3.2	Where the interacting variables are discrete, all values and combinations of values of W in the policy population must be represented in the experimental sample[25]
R4.1	The researcher must be able to obtain unbiased estimates of the conditional average treatment effect ($E[\triangle \vert D = 0, W]$) for all values of W
R4.2	The size of the experimental sample should be large enough, and the dimension of W small enough, to enable R4.1
R5	The average treatment effect should not vary across populations for any reason not related to observed covariates

The generic importance of functional form for external validity has been noted by Leamer (2010) and Keane (2010a). In that regard, the most challenging requirements above are arguably R1 and R3.2. As we have seen, the experimental approach is often favoured by researchers who believe it implausible that unconfoundedness conditions can be satisfied simply by judicious covariate selection, or clever structural modelling. However, to know in advance what the interacting factors are must require some reliable theoretical knowledge. Worse, it is widely recognised that there is often no persuasive theoretical reason to choose one functional form over another. This has spurred the literature on nonparametric estimation but, as should be clear from the above framework, nonparametric estimation of the average treatment effect is insufficient for extrapolation. As Heckman and Vytlacil (2007a: 4855) put it, "To extend some function...to a new support requires functional structure: It cannot be extended outside of sample support by a purely nonparametric procedure". This point, more than any other, is underemphasised or unacknowledged in contributions to the experimental literature.

Relatedly, we must have good reasons to believe that *measured* variables are comparable across populations. For example, how does one compare racial categories across populations of different countries for the purposes of reweighting

treatment effects? There may be theoretical concerns that 'race' is a notion with different meaning in different societies and that therefore there is no common variable across such populations. More obviously, different categorizations may be used in different populations so that the available variables are not comparable.

The availability of information on the treatment effect across the support of interacting variables also has important implications for decision making. Manski summarises the problem as follows:

> The physician may have extensive covariate information for his own patients but the journal report of the clinical trial may only report outcomes within broad risk-factor groups...However the available experimental evidence, lacking covariate data, only reveals mean outcomes in the population as a whole, not mean outcomes conditional on covariates. Hence the planner faces a problem of treatment choice under ambiguity.Manski (2000: 419)

Interaction in itself is not an obstacle to estimating the average treatment effect in the experimental sample. In the context of estimating the ATE using a regression, even if the nature of the interactions are unknown a regression on the treatment variable that only conditions on the relevant covariates - but omits interaction terms - will produce an unbiased estimate. This follows from the general result - see for instance Wooldridge (2002: 21) - that $E[Y|X, W, f(X, W)] = E[Y|X, W]$, provided $E[Y|X, W]$ is linear in the parameters. However, predicting the average treatment effect in another population using the estimated parameters would require the original functional form to have been correctly specified.

The bottom line is that unless researchers have accurate *ex ante* beliefs about the factors that interact with the treatment variable *and* are able to collect data on these, forecasting effects of treatment in new populations will be a matter of luck. The framework based on interactive functional forms suggests that this can take three forms: that the causal effect happens to be approximately the same across individuals regardless of variation in characteristics; that the causal effect does not actually depend on the values of other variables (additive separability); or, that there is little variation in the mean values of the interacting variables across contexts.

3.2 Heterogeneity of treatment effects

Given the above one might expect guides for empirical practice to address the issue in some detail, but discussion of interaction terms is absent from some of the

main ‘manuals’ for conducting analysis based on randomized evaluations (Angrist and Pischke (2009), Duflo et al. (2006b)) - an omission also noted by Gelman (2000). To the extent that these issues have received any widespread attention in the treatment effect literature it has primarily been in relation to studies that examine ‘treatment effect heterogeneity’ within experimental populations, in other words the extent to which an estimated treatment effect varies across subgroups. In that vein, while Allcott and Mullainathan (2012) and, more recently, Pritchett and Sandefur (2013) both utilise representations of potential outcomes similar to the ones deployed above, those authors place little emphasis on the role of functional form per se, rather simply proceeding from the assumption that - for whatever reason - treatment effects vary with covariate values. We now briefly examine this heterogeneity-premised approach.

If the treatment effect were constant for all individuals in the entire population then external validity would, necessarily, hold. Variation in the treatment effect is sometimes referred to in the literature as ‘treatment heterogeneity’, but this term is not used consistently. Specifically, it is important for our purposes to distinguish between three conceptions of treatment heterogeneity. The first, and arguably more common use of the term to date, focuses on heterogeneity relating to the presence of compliers and non-compliers in instrumental variable estimation of local average treatment effects - see for instance Angrist (2004) and the discussion in Angrist and Pischke (2009). The second refers to some fundamental level of randomness that produces ‘intrinsic’ variation in the effect across individuals with identical characteristics. The third concerns the existence of empirical variation in the average treatment effect itself across values of covariates. These obviously need not be mutually exclusive, since if the characteristics of compliers and non-compliers differ then that would manifest as heterogeneity across the covariates representing these characteristics. Contributions on external validity have mirrored this distinction: Angrist and Fernandez-Vál (2010, 2013) present an analysis of the extrapolation/external validity problem focused on compliance in the case of estimating LATEs, whereas Hotz et al. (2005), Crump, Hotz, Imbens, and Mitnik (2008) and Crump, Hotz, Imbens, and Mitnik (2009) provide definitions of external validity based only on variation in the treatment effect across covariate values. In part contrast to these, structural approaches distinguish themselves in examining variation in behavioral responses to treatment across different populations - see for instance Heckman and Vytlacil (2005).

Having assumed perfect compliance, assumed-away selection based on choice and having no particular interest in intrinsic heterogeneity, what is relevant to the present review is variation across covariates. The way in which that has been

addressed in the literature is largely unsatisfactory. Authors typically conduct heterogeneity analyses across subgroups that are defined after the experiment has been completed, based on covariate data that was collected to establish success of random assignment or to justify a conditional unconfoundedness assumption. "At best, researchers have estimated average effects for subpopulations defined by categorical individual characteristics" Crump et al. (2008: 398), but this is typically ad hoc (see also Deaton, 2008, 2010). In some instances it could quite plausibly be argued that this constitutes specification searching without compensating adjustments for the statistical significance of results. Rothwell (2005b) makes similar points in relation to the medical literature and Fink, McConnell, and Vollmer (2013) provide an overview of such practices in development economics along with some, standard, suggestions regarding correction for multiple hypothesis testing.

More systematic methods have been proposed. Hotz et al. (2005) propose a method for testing the unconfoundedness assumption across two experimental populations by comparing the actual mean outcomes for controls to those predicted using data from the other population. Perhaps most notable is the contribution of Crump et al. (2008) who develop two nonparametric tests: the first is of the null hypothesis of zero average treatment effect conditional on a set of observable covariates *in subpopulations* (as defined by the covariates); the second is for the null hypothesis that the conditional average treatment effect is the same across subpopulations, in other words a test of treatment heterogeneity. Djebbari and Smith (2008) utilise a number of other methods to test for heterogeneity in data on the PROGRESA conditional cash transfer program implemented in Mexico. The authors make some effort to account for multiple tests and consider various nonparametric bounds of the variance of treatment effects. Allcott and Mullainathan (2012) also suggest a particular F-test of whether treatment effects vary within sub-groups of the experimental population as defined by covariate values. It is essentially a test for joint significance of the parameters on the interaction terms between sub-group dummies and the treatment variable. This would appear to be a version of the more general case proposed by Crump et al. (2008).[27] The authors discuss potential empirical obstacles to this approach, such as a possible lack of power caused by small samples and the fact that in-and-of itself the test provides no basis for extrapolating to a new population. As they note, the greatest value of such tests is likely to be where the null hypothesis of no significant sub-group variation is rejected.

[27] Allcott and Mullainathan (2012) appear to be unaware of the work by Crump et al. (2008).

In their comment on external validity, Pritchett and Sandefur (2013) examine variation in estimates of the effect of five different kinds of interventions and conclude that this is large enough to call into question the likelihood of external validity of such effects. In addition, they take an approach to heterogeneity that is similar to the previously-cited study by Concato et al. (2000), by comparing the mean squared error in non-experimental and experimental estimates. Their conclusion is that "policymakers interested in minimizing the error of their parameter estimates would do well to prioritize careful thinking about local evidence over rigorously-estimated causal effects from the wrong context" (Pritchett and Sandefur, 2013: 25). Concato et al. (2000) come to a complementary finding, that "summary results of randomized, controlled trials and observational studies were remarkably similar for each clinical topic we examined...Viewed individually, the observational studies had less variability in point estimates (i.e., less heterogeneity of results) than randomized, controlled trials on the same topic".

Under the assumption of unconfoundedness the heterogeneity of treatment effects across covariates is the consequence of a true causal relationship in which the treatment variable interacts with covariates to produce values of the outcome of interest. As we have seen, such interaction is the major challenge for obtaining external validity of results from ideal experiments. While *ex post*, data-driven assessment of heterogeneity may be informative about possible threats to extrapolation, the result has been that the experimental literature has largely neglected the question of why interactions would exist and whether incidentally-gathered data is adequate for a rigorous assessment.

3.3 Selection, sampling and matching

Another way to frame the external validity problem is as a case of sample selection bias: the population being experimented on has come about through some kind of selection process and is therefore importantly different from the population we are interested in.[28] That suggests, in turn, two other issues that are relevant to solving, or better understanding, the external validity problem.

Sampling

The first of these concerns the use of deliberate sampling of experimental populations. In their analysis, Allcott and Mullainathan (2012) note the possible advantages, at the experimental design stage, of "RCTs with representative samples

[28]This draws attention to the fact that an 'ideal experiment', which we have assumed in the preceding analysis, is typically defined only relative to factors *within* the experimental sample.

of the Target population of interest" (Allcott and Mullainathan, 2012: 32), and replicating experiments in locations where the support of the covariates overlaps with a portion of the support in the target population that does not exist in preceding experiments. Similar views are expressed by Falagasa et al. (2010: 11) that researchers should endeavour to "[match] the population to be included in the RCT to the respective population that is expected to be encountered in general practice".

In fact, all these ideas can be found in Cook and Campbell (1979) who, much as they identify external validity as a problem of interaction, consider solutions as being fundamentally dependent on sampling. Those authors discuss three possible, sampling-based solutions: 'random sampling for representativeness'; 'deliberate sampling for heterogeneity'; and, 'impressionistic modal instance modelling' (Cook and Campbell, 1979: 74-80). The first two solutions are self-explanatory in the context of preceding discussions and correspond exactly to the two suggestions by Allcott and Mullainathan (2012). The third appears to refer to a process somewhat similar to that used when conducting case studies: look for instances that most closely resemble the situation of interest and conduct experiments on these. This, in turn, bears some resemblance to the idea of 'analogical reasoning' proposed in philosophy by Guala (2003) and is suggested by Cook and Campbell only for situations where relatively low generalisability is required.

In as much as representative sampling of the population of interest yields experimental estimates across the support of the relevant covariates in that population, it appears the most likely of the three approaches to lead to external validity. Representative sampling, however, assumes that a target population is known *ex ante*. Some empirical studies appear to have the more ambitious objective of estimating effects that are valid across multiple contexts, including populations that are disjoint from the original experimental sample, or for target populations that are not yet known. In that instance, deliberate sampling for heterogeneity will be required to obtain a large enough coverage of the support to be able to reweight conditional average treatment effects in the way envisaged by Hotz et al. (2005) in (6). A similar issue has been discussed in a recent paper by Solon, Haider, and Wooldridge (2013), albeit with an emphasis on the uses of weighting in empirical work rather than external validity per se.

Approached from what one might call (a la Keane (2010b)) the 'atheoretic' perspective of treatment heterogeneity, the suggestion that researchers sample for heterogeneity seems unobjectionable. However, from the perspective of interactive functional forms this injunction appears to beg the question. Besides the

requirement that it be coherent to compare these variables across contexts, a concern noted in Table 2, sampling for heterogeneity requires that researchers know in advance which variables play a role in determining the effect of the treatment variable. And yet, as we now briefly discuss, a similar assumption would suffice to justify the use of *non-experimental* methods to obtain identification of causal effects (internal validity).

Matching

A prominent method for obtaining identification of causal effects using *non*-experimental data is employed by *matching estimators*; Imbens (2004), Todd (2006) and Morgan and Winship (2007) all provide overviews of the relevant literature. As Rubin (1973) notes, the early matching literature was concerned with improving precision of estimates, whereas his interest - and much of the interest in the literature since Rubin's contributions - has been concerned with using matching to remove, or mitigate, bias. The basic process is intuitive: to ensure unconfoundedness - as in (3) - without experimental assignment, the researcher matches individuals from the 'treatment' and 'no treatment' groups based on a set of observable covariates.[29] If these are covariates that would otherwise confound estimation then it is possible to obtain an unbiased estimate of the average causal effect of interest by summing-up effects across matched individuals (or sub-groups). This is essentially a nonparametric approach which means that matching estimators "do not require specifying the functional form of the outcome equation and are therefore not susceptible to bias due to misspecification along that dimension" (Todd, 2006: 3861).

The two issues that have preoccupied the theoretical literature are: how to obtain the best matches between individuals or groups; and, how best to weight these individual- or group-specific effects. The criteria by which optimality is assessed are bias reduction and asymptotic efficiency. Empirically a number of other problems arise. The most obvious is how to choose the set of covariates upon which matches are constructed. Todd (2006: 3869) notes that "unfortunately there is no theoretical basis for choosing a particular set".[30] A second problem is that in some datasets there may not exist any matches for some subsets of the population.

[29] These terms are in quotes to indicate that there need not have been experimental assignment. Matching methods are sometimes employed, as per their original use, even with experimental data in order to improve precision - see Imbens (2004)'s discussion.

[30] Some authors in the broader causal inference literature - Spirtes, Glymour, and Scheines (1993) and Pearl (2009) - have developed algorithmic methods for identification of causal relationships and may disagree with this claim. The likely success of those methods remains contested, however, and detailed consideration of them would take us beyond the focus of the present review.

Strictly speaking this means the effect of interest cannot be estimated. However, some authors have proposed redefining the effect of interest based on the more limited support of the covariates that is used.[31] A final problem is that the dimension of the covariate vector might be large, making accurate estimation with most datasets infeasible. One solution to this problem has been to employ a version of the previously mentioned theorem by Rosenbaum and Rubin (1983) that conditioning on the propensity score is equivalent to conditioning on the covariates directly. Matching is then conducted on the basis of propensity scores.

Our interest in matching is not as an alternative per se to experimental methods, or indeed structural ones. Instead we are interested in how the assumptions required for matching estimators to be unbiased compare to the assumptions required for (conditional) external validity to hold. The key assumption required for cross-sectional matching estimators is that the set of covariates satisfies the unconfoundedness assumption in (3). An additional assumption is required to ensure that there exist matches for all individuals in the population (or treatment population of the researcher is estimating the ATT), which corresponds to the assumption of overlapping support in (4). Comparing these assumptions to those required for conditional external validity (Hotz et al., 2005) - assumption 2.1 and 2.2 - indicates that the requirements are identical, with the dummy for treatment receipt replaced by a dummy for presence in the experimental or policy population. Absent any other considerations, it would seem that if we are to believe in atheoretic external validity we should also be willing to believe in the unbiasedness of non-experimental estimators of treatment effects. That basic idea has been recognised by a number of authors, such as Deaton (2008: 44) who notes the relationship between qualitative arguments for similarities across contexts used by advocates of experimental methods to claim some level of generalisability of their experimental results, with the logic of matching estimators for deriving causal effects from non-experimental data.

The preceding discussion suggests one important caveat to the above conclusion. In comparing the problems of matching and external validity it is useful to distinguish between two kinds of variables that are relevant for matching using non-experimental data: variables that might confound the estimated treatment effect, by being correlated with the causal variable and the outcome variable; and, variables that could be independent of the causal variable of interest, but interact

[31] This appears to be one reason why many matching studies prefer to estimate the effect of treatment on the treated, which produces an asymmetry in the conditions that must be satisfied; most notably, the key concern becomes finding matches for individuals in the 'treated' population, preferably from a 'large reservoir of controls' - see Imbens (2004: 14).

with it and therefore mediate its effect on the outcome. Because matching does not require specification of functional form this distinction is not directly relevant for that theory - though one might expect that it should inform decisions about which variables to use or obtain data on, but it is relevant for external validity from an ideal experiment since that need only be concerned with interacting variables. Given this distinction one could argue that in scenarios where experimental assignment and compliance approximate the ideal, the set of variables required to satisfy unconfounded location is a subset of those required to satisfy unconfoundedness in non-experimental data. Another caveat is that the process by which individuals select, or are selected, into an experimental sample is likely to differ from the process whereby some come to receive a non-experimental intervention and others do not. Such differences, however, would require some level of theoretical modelling to distinguish.

3.4 Implications for replication and repetition

As we have seen, a popular position among proponents of randomized evaluations is that the problem of external validity is fundamentally empirical rather than conceptual: to assess if an effect holds across other populations or interventions that are somewhat different we must experimentally test that hypothesis. For instance, Duflo et al. (2006b: 71) suggest that: "it is a combination of replications and theory that can help generalize the lessons from a particular program". And Angrist and Pischke (2010: 23) state that: "a constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge...The cumulative force of...studies has some claim to external validity". The implication here is that to simply point out the limitations of experimental results is 'non-constructive' and therefore ought to be disregarded. Even scholars such as Manski (2013a) appear to temper criticism in the face of this dictum. By contrast, Deaton (2010: 30) argues that "repeated successful replications of a [typical randomized evaluation experiment] is both unlikely and unlikely to be persuasive" and Rodrik (2008: 21) states that, "Repetition would surely help. But it is not clear that it is a magic bullet."

As we have already noted, the emphasis on replication emerges from Cook and Campbell (1979)'s suggestion of 'sampling for heterogeneity'. Those authors advocate an identical position to modern experimentalists, arguing that: "in the last analysis, external validity...is a matter of replication [and]...a strong case can be made that external validity is enhanced more by many heterogeneous small experiments than by one or two large experiments" (Cook and Campbell, 1979: 80). What is striking about the analysis of Duflo et al. (2006b) and Angrist and

Pischke (2010) is that the authors provide no systematic framework for determining whether evidence across contexts is 'similar' or 'similar enough', nor how we ought to cumulate knowledge over multiple experiments in different contexts. This is in marked contrast to the detailed and careful development of arguments illustrating why randomized variation suffices to identify causal effects within a given sample. Indeed, with both proponents and critics of randomized evaluations, little specific justification is given for claims regarding replication.

By contrast, the obstacles to external validity identified in our preceding analysis of functional form and interaction provide a clear indication of the challenges to using replication, as a form of 'sampling for heterogeneity', to resolve the extrapolation problem.[32] To aid extrapolation replication should: take place in domains that differ according to the interacting variables; the values of these variables must be observed; and, for the final policy prediction, the functional form of the relationship should be known, or it ought to be possible to obtain nonparametric estimates. The puzzle, however, is that in the empirically simpler case where the functional form and relevant interacting factors are known *ex ante* then replication may not be necessary. As per Hotz et al. (2005)'s definition of conditional external validity in (6), we need only observe the relevant factors in the experimental and policy populations and reweight accordingly. The only role of replication, in that instance, would be to observe the value of the treatment effect across parts of the support of the vector of interacting variables that is in the policy population, but not in previous experimental populations. This is a more subtle point than addressed in the literature and again says nothing about how researchers will come to know which factors mediate the causal effect and which do not. In the absence of knowing what causes heterogeneity one cannot deliberately sample for it.

The closest to an explicit method for using replication to aid prediction is discussed by Imbens (2010, 2013). The former advocates the use of repeated experiments to *semi-parametrically* estimate the functional relationship between a causal variable and the outcome of interest without estimating causal parameters. That, in turn, relies on the existence of 'detailed information' on the characteristics of the relevant populations. It should be clear that this is a direct extension of Hotz et al. (2005), with the exception of assuming that some parameteric struc-

[32] As per Cook and Campbell (1979)'s sampling-based solutions to the external validity problem, an alternative would be to use replication as a way of obtaining a random sample from the population of interest. This is not, however, the standard justification for replication in the literature.

ture can be imposed on the relationship.[33] Imbens (2013: 406) makes the more modest proposal of using differences in the average value of covariates to assess the possible difference in average treatment effects across two populations. This, too, relies on the assumption that the relevant variables are observable and says nothing about how to identify these. The problems implicit in this latter approach therefore carry-over to the replication case. Most obviously, researchers must somehow know *and* be able to observe all relevant interacting factors in al populations. In addition it must be possible for practically feasible levels of replication to obtain information on the support (joint distribution) of all such interacting factors present in the target or policy population of interest. This, unfortunately, somewhat undermines one of the primary motivations for emphasising randomized evaluations - discussed in section 1 - that researchers need not know the underlying model or observe other causal factors in order to identify a causal effect of interest. Lastly, one may note that there exist no studies in the literature that can claim, or have claimed, to satisfy these requirements.

4 Conclusions and implications for empirical work

Randomized trials have now been utilised in research areas as diverse as physics, biology, medicine, sociology, politics and economics, and as a consequence have become somewhat synonymous with scientific practice. Where they are able to satisfy, or closely approximate, the ideal experiment randomized evaluations allow researchers to estimate the causal effect of the intervention in the experimental population. It is important to recognise that the prospects for success with such methods is likely to vary by discipline. Specifically, the nature of problems in areas such as physics and, to a lesser extent, human biology are such that it is easier to control and manipulate factors than in economics, and the identified causal relationships are more likely to be stable over time and space. That may partly reflect stability in mechanisms, but also the stability of relevant interactive factors over contexts - something which is relatively implausible for many questions of interest in economics. For example, Ludwig et al. (2011: 33) cite the discovery that statins reduce the risk of heart disease, even though the process by which they do so is not yet understood, to justify the use of evidence from 'black box' evaluations to make policy decisions. Similarly, Angrist and Pischke (2010) argue that: "inconclusive or incomplete evidence on mechanisms does not void empirical evidence of predictive value. This point has long been understood in medicine, where clinical evidence of therapeutic effectiveness has for centuries run ahead of the theoretical understanding of disease". However, there are few - if

[33] Imbens (2010: 25) specifically refers to "fitting a flexible functional form" to the relevant conditional expectation.

any - economic processes that appear likely to possess the stability across contexts that basic human biology does and therefore such comparisons seem unlikely to be informative about external validity in economics.[34]

Our analysis of interaction, which builds upon the much-referenced but otherwise, in economics, largely neglected insights of Campbell and Stanley (1966) and Cook and Campbell (1979), examines a logically distinct problem: when causes interact with other factors, extrapolation to new contexts requires data on these factors *in both contexts* and, for most sample sizes, knowledge of the functional form of the underlying mechanism. In the absence of these, the researcher or policymaker relies implicitly or explicitly on the optimistic assumption that - if the expectation of the treatment effect is of interest - the means of any mediating factors are approximately the same across the experimental and policy populations. If that assumption is false then even where the average treatment effect of a given experimental evaluation is accurately estimated it will not generalise to other environments.

In this regard, it is our view that the work of Hotz et al. (2005)in particular and, in the realm of instrumental variable estimation, Angrist and Fernandez-Vál (2013) provide the first indications of what a systematic, formal approach to external validity might look like. In both cases variation of treatment effects across the distribution of covariates is fundamental. Therefore where there is full overlap in the supports of the relevant variables across the populations of interest and these are observed in the experiment, researchers can get some sense of external validity problems from an analysis of 'treatment heterogeneity' - which we have defined in the narrow sense to exclude issues of compliance that are nevertheless partly addressed in Angrist and Fernandez-Vál (2013). While briefly popular in the experimental evaluation literature, tests of heterogeneity typically been conducted *ex post* on data that happens to available and with the risk of falsely significant results due to multiple hypothesis testing or failure to deal with dependence across variables. Tools for more systematic approaches have recently been proposed by Crump et al. (2008). Regardless, only a very small minority of contributions to the applied literature make any attempt to extend empirical analysis of treatment heterogeneity to forecasting or extrapolation of results in new contexts and there is currently no consensus on appropriate methods for doing so.

[34]This point is acknowledged by Imbens (2010). In philosophy the influential work of Nancy Cartwright Cartwright (1979, 1989, 2007) has emphasised the importance of what she refers to as 'stable capacities' in order to predict the causal effects of interventions. And as Cartwright notes, differing opinions on the likely existence of similarly stable properties in domains of economic interest were one consideration in early debates on the merits of econometric analysis.

The above review is based on a deliberately simplified version of the extrapolation problem, assuming-away many real world obstacles to obtaining an 'ideal experiment'. This includes ignoring the consequences of individual optimising behavior, which is the starting point for the entire structural literature. Even in that pared-down scenario we identified - in Table 2 - five major obstacles to obtaining external validity in practice. The absence, to date, of any search-based method for obtaining knowledge of the relevant interacting variables somewhat undermines the oft-stated rationale for experimental methods of obtaining meaningful causal effects without committing to implausible structural assumptions. Such knowledge, in turn, is required to gather the data necessary to conduct such analyses in practice. It is also possible that for some important variables - such as the history of institutions in different countries - there is no meaningful overlap in support, rendering extrapolation in this formal framework impossible. What is perhaps most striking is that the requirements for external validity to be achieved parallel those required to obtain identification ('internal validity') from non-experimental data. This, we suggest, confirms the view expressed by authors such as Manski (2013a,b) that the external validity question deserves at least as much attention as internal validity. Perhaps this problem can be solved, as suggested by Cook and Campbell (1979) and much more recently by Allcott and Mullainathan (2012), through constructing the experimental population via random sampling of the population of interest, much as the treatment populations are constructed by random assignment. Some studies, however, appear to aspire to much greater generalisability even without such representative sampling, which is evidently problematic.

From the perspective of using econometric programme evaluations to inform policy making - whether in the area of development or elsewhere - this raises two questions. First, what should the null hypothesis of researchers be, that causal relationships are interactive or additively separable? As things stand the majority of contributions to the experimental programme evaluation literature that seek to make any claims of relevance beyond the experimental sample implicitly assume additive separability. Prudence suggests the opposite approach when informing policy: assume interactive functional forms unless there is evidence to suggest otherwise and temper policy recommendations accordingly. The second issue is how important such interaction effects are *empirically*. Where are interactive relationships important? And to what *extent* does unrecognised functional form and variation in the means of mediating variables across contexts affect external validity from a policymaker's perspective?

There is no doubt that the role of theory is, at least in part, to inform empirical analysis in this way and that 'atheoretic' replication cannot plausibly suffice to

resolve the problem. Whether that can be done using more, or better, theory to an extent sufficient to produce confident extrapolation of results from one context to another remains a wholly open question.

References

Allcott, H. and S. Mullainathan (2012). External validity and partner selection bias. *NBER Working Paper* (18373).

Angrist, J. (2004). Treatment effect heterogeneity in theory and practice. *Economic Journal 114*, C52–C83.

Angrist, J. and I. Fernandez-Vál (2010). ExtrapoLATE-ing: external validity and overidentification in the LATE framework. *NBER Working Paper* (16566).

Angrist, J. and I. Fernandez-Vál (2013). ExtrapoLATE-ing: external validity and overidentification in the LATE framework. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Econometric Society Monographs, Tenth World Congress (Vol.III)*. Cambridge University Press.

Angrist, J. D. and A. B. Krueger (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives 15*(4), 69–85.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics*. Princeton: Princeton University Press.

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Banerjee, A. V. (2007). *Making aid work*. Cambridge(MA): MIT Press.

Banerjee, A. V. and E. Duflo (2008). The experimental approach to development economics. *NBER Working Paper 14467*.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics 1*, 151–178.

Banerjee, A. V. and S. M. R. Kanbur (2005). *New Directions in Development Economics: Theory Or Empirics? : a Symposium in Economic and Political Weekly*. Working paper (New York State College of Agriculture and Life Sciences. Dept. of Applied Economics and Management). Cornell University.

Bardhan, P. (2013, 20 May). Little, big: Two ideas about fighting global poverty. *Boston Review*.

Barton, S. (2000). Which clinical studies provide the best evidence? the best rct still trumps the best observational study. *British Medical Journal 321*, 255–256.

Benson, K. and A. J. Hartz (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine 342*(25), 1878–1886.

Binmore, K. (1999). Why experiment in economics? *Economic Journal 109*, 16–24.

Bold, T., M. Kimenyi, G. Mwabu, A. Ngángá, and J. Sandefur (2013). Scaling up what works: Experimental evidence on external validity in Kenyan education. *Center for Global Development Working Paper* (321).

Campbell, D. T. and J. C. Stanley (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally College Publishing.

Card, D., S. DellaVigna, and U. Malmendier (2011). The role of theory in field experiments. *Journal of Economic Perspectives 25*(3), 39–62.

Cartwright, N. (1979). Causal laws and effective strategies. *Nous 13*, 419–437.

Cartwright, N. (1989). *Nature's Capacities and their Measurement*. Oxford: Oxford University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical studies 147*, 59–70.

Cartwright, N. (2011a). Evidence, external validity, and explanatory relevance. In G. J. Morgan (Ed.), *Philosophy of Science Matters: The Philosophy of Peter Achinstein*, pp. 15–28. New York: Oxford University Press.

Cartwright, N. (2011b). Predicting 'it will work for us': (way) beyond statistics. In P. I. McKay, F. Russo, and J. Williamson (Eds.), *Causality in the sciences*. Oxford (UK): Oxford University Press.

Concato, J., N. Shah, and R. Horwitz (2000). Randomized controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine 342*(25), 1887–1892.

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Wadsworth.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics 90*(3), 389–405.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

Deaton, A. (2008, October 9th). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Keynes Lecture, British Academy.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. *NBER working paper* (w14690).

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics 125*, 141–173.

Dehejia, R. H. (2013). The porous dialectic: Experimental and non-experimental methods in development economics. *WIDER Working Paper* (No. 2013/11).

Dekkers, O. M., E. von Elm, A. Algra, J. A. Romijn, and J. P. Vandenbroucke (2010). How to assess the external validity of therapeutic trials: a conceptual approach. *International Journal of Epidemiology 39*, 89–94.

Djebbari, H. and J. Smith (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics 145*, 64–80.

Duflo, E., R. Glennerster, and M. Kremer (2006a). Chapter 61: Using randomization in development economics research: A toolkit. In *Handook of Development Economics Volume 4*. Amsterdam: Elsevier.

Duflo, E., R. Glennerster, and M. Kremer (2006b). Using randomization in development economics research: A toolkit. Accessed 24th January 2011 from http://www.povertyactionlab.org/sites/default/files/documents/Using

Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

Elliott, G. and A. Timmermann (2008). Economic forecasting. *Journal of Economic Literature 46*(1), 3–56.

Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing 12*(1), 77–84.

Falagasa, M. E., E. K. Vouloumanoua, K. Sgourosa, S. Athanasioud, G. Peppasa, and I. I. Siemposa (2010). Patients included in randomised controlled trials do not represent those seen in clinical practice: focus on antimicrobial agents. *International Journal of Antimicrobial Agents 36*, 1–13.

Fink, G., M. McConnell, and S. Vollmer (2013). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness*. Published online at `http://dx.doi.org/10.1080/19439342.2013.875054`.

Garfinkel, I., C. F. Manski, and C. Michalopolous (1992). Micro experiments and macro effects. In *Evaluating welfare and training programs*, pp. 253–276. Cambridge (MA).

Gelman, A. (2000). A statistician's perspective on Mostly Harmless Econometrics: An Empiricist's Companion, by Joshua D. Angrist and J orn-Steffen Pischke. *Stata Journal 9*(2), 315–320.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science 70*(5), 1195–1205.

Guala, F. (2005). Economics in the lab: completeness vs. testability. *Journal of Economic Methodology 12*(2), 185–196.

Guala, F. and L. Mittone (2005). Experiments in economics: external validity and the robustness of phenomena. *Journal of Economic Methodology 12*(4), 495–515.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*, iii–iv,1–115.

Hacking, I. (1988). Telepathy: origins of randomization in experimental design. *Isis 79*(3), 427–451.

Hadorn, D. C., D. Baker, J. S. Hodges, and N. Hicks (1996). Rating the quality of evidence for clinical practice guidelines. *Journal of Clinical Epidemiology 49*(7).

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Heckman, J. and J. H. Abbring (2007). Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics (Volume 6B)*, Volume 6B, Chapter 72, pp. 5145–5303. Amsterdam: Elsevier.

Heckman, J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics 30*, 239–267.

Heckman, J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies 64*(4), 487–535.

Heckman, J. and E. Vytlacil (2007a). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics (Volume 6B)*, Volume 6B, Chapter 70, pp. 4779–4874. Amsterdam: Elsevier.

Heckman, J. and E. Vytlacil (2007b). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics (Volume 6B)*, Volume 6B, Chapter 71, pp. 4875–5143. Amsterdam: Elsevier.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *Quarterly Journal of Economics 115*, 45–97.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review 76*, 1–27.

Heckman, J. J., H. Ichimura, and P. Todd (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J. and S. Urzua (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics 156*(1), 27–37.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: OUP.

Herberich, D. H., S. D. Levitt, and J. A. List (2009). Can field experiments return agricultural economics to the glory days? *American Journal of Agricultural Economics 91*(5), 1259–1265.

Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association 81*(396), 945–960.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*, 241–270.

Hsiao, C. (1992). Random coefficient models. In L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data: Handbook of theory and applications*, pp. 72–94. Springer.

Hsiao, C. and M. H. Pesaran (2004). Random coefficient panel data models. *IZA Discussion Paper 1236*.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton(2009) and Heckman and Urzua(2009). *Journal of Economic Literature 48*(2), 399–423.

Imbens, G. W. (2013). Book review feature: Public Policy in an Uncertain World. *Economic Journal 123*, F401–F411.

Kahneman, D. and A. Tversky (1979). Prospect theory: an analysis of decision under risk. *Econometrica 47*(2), 263–292.

Keane, M. (2005, 17-19 September). Structural vs. atheoretic approaches to econometrics. Keynote Address at the Duke Conference on Structural Models in Labor, Aging and Health.

Keane, M. P. (2010a). A structural perspective on the experimentalist school. *Journal of Economic Perspectives 24*(2), 47–58.

Keane, M. P. (2010b). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics 156*(1), 3–20.

Keynes, J. (1939). Professor tinbergen's method. *Economic Journal 49*(195), 558–577.

Kramer, M. and S. Shapiro (1984). Scientific challenges in the application of randomized trials. *Journal of the American Medical Association 252*, 2739–2745.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 96*(4), 604–619.

Leamer, E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives 24*(2), 31–46.

Levitt, S. D. and J. A. List (2007). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics 40*(2), 347–370.

Levitt, S. D. and J. A. List (2009). Field experiments in economics: The past, the present and the future. *European Economic Review 53*, 1–18.

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives 25*(3), 3–16.

Loewenstein, G. (1999). Experimental economics from the vantage point of behavioural economics. *Economic Journal 109*, 25–34.

Ludwig, J., J. R. Kling, and S. Mullainathan (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives 25*(3), 17–38.

Manski, C. (2000). Identification and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *JoE 95*(2), 415–442.

Manski, C. (2011). Policy analysis with incredible certitude. *Economic Journal 121*(554), F261–F289.

Manski, C. F. (2013a). *Public policy in an uncertain world: analysis and decisions*. Cambridge (MA): Harvard University Press.

Manski, C. F. (2013b). Response to the review of 'public policy in an uncertain world'. *Economic Journal 123*, F412–F415.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. Hood and T. Koopmans (Eds.), *Studies in Econometric Method*, pp. 1–26. Wiley.

Matzkin, R. L. (2007). Nonparametric identification. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics (Volume 6B)*, Volume 6B, Chapter 73, pp. 5307–5368. Amsterdam: Elsevier.

McKee, M., A. Britton, N. Black, K. McPherson, C. Sanderson, and C. Bain (1999). Interpreting the evidence: choosing between randomised and nonrandomised studies. *British Medical Journal 319*, 312–315.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics 13*(2), 151–161.

Morgan, S. L. and C. Winship (2007). *Counterfactuals and causal inference*. Cambridge: Cambridge University Press.

Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society II Suppl.(2)*, 107–180.

Pearl, J. (2000 (2009)). *Causality: Models,reasoning and inference*. Cambridge: Cambridge University Press.

Pritchett, L. and J. Sandefur (2013). Context matters for size. Center for Global Development Working Paper 336.

Ravallion, M. (2008, March). Evaluation in the practice of development. World Bank Policy Research Working Paper 4547.

Ravallion, M. (2009, February). Should the Randomistas rule? Economists' Voice.

Rodrik, D. (2008, October). The new development economics: We shall experiment, but how shall we learn? *Harvard Kennedy School Working Paper RWP08-055*.

Roe, B. E. and D. R. Just (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics 91*(5), 1266–1271.

Rosenbaum and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Ross, S., A. Grant, C. Counsell, W. Gillespie, I. Russell, and R. Prescott (1999). Barriers to participation in randomised controlled trials: A systematic review. *Journal of Clinical Epidemiology 52*(12), 1143–1156.

Roth, A. E. (1988). Laboratory experimentation in economics: a methodological overview. *Economic Journal 98*, 974–1031.

Rothwell, P. M. (2005a). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet 365*, 82–93.

Rothwell, P. M. (2005b). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet 365*, 176–186.

Rothwell, P. M. (2006). Factors that can affect the external validity of randomised controlled trials. *PLoS CLinical Trials 1*(1), 1–5.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*, 135–146.

Rubin, D. (1973). Matching to remove bias in observational studies. *Journal of Educational Psychology 29*(1), 159–183.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. (1980). Comment. *Journal of American Statistical Association 75*(371), 591–593.

Rust, J. (2010). Comments on: "Structural vs. atheoretic approaches to econometrics" by Michael Keane. *Journal of Econometrics 156*, 21–24.

Samuelson, L. (2005). *Journal of Economic Literature 43*(1), 65–107.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology 12*(2), 225–237.

Solon, G., S. J. Haider, and J. Wooldridge (2013). What are we weighting for? *NBER Working Paper* (18859).

Spirtes, P., C. N. Glymour, and R. Scheines (2000 (1993)). *Causation, prediction and search*. Cambridge(MA): MIT Press.

Steel, D. (2008). *Across the boundaries: extrapolation in biology and social science*. Oxford: Oxford University Press.

Sugden, R. (2005). *Journal of Economic Methodology 12*(2), 177–184.

Todd, P. E. (2006). Chapter 60: Evaluating social programs with endogenous program placement and selection of the treated. In *Handook of Development Economics Volume 4*. Amsterdam: Elsevier.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge(MA): MIT Press.