Causal Models and External Validity of Randomized Experiments

A dissertation submitted for the degree MSc Philosophy of Science
London School of Economics and Political Science
Candidate Number: 12355
Supervisor: Johanna Thoma
Word Count: 9’260

List of contents

1. Introduction

Scientific evaluations of policy-effectiveness are an important part of economics and the social sciences in general. Their aim can be to inform policy-makers and donors about the costs and benefits of policy programs. A considerable amount of research in this field has been conducted using randomized experiments, also called randomized control trials (RCTs).[1] Credible RCT results about policy interventions are considered to reliably inform development organizations about policy programs, specifically about their cost-effectiveness (Duflo & Kremer, 2008). The *Abdul Latif Jameel Poverty Action Lab* (J-Pal) is an example of a research centre which aims to inform policy based on scientific evidence by means of RCTs. Randomized experiments are widely acknowledged to be the ideal method to establish causal claims. They require little or no prior information, making them largely independent of so-called *expert-knowledge*, which could have troublesome motivations (Deaton & Cartwright, 2018).

The issue this paper will focus on is the *use* of experimentally evaluated causal claims. Randomized experiments have been criticised, prominently by Angus Deaton and Nancy Cartwright (Deaton, 2010; Cartwright, 2012; Deaton & Cartwright, 2018) inter alia for not providing sufficient grounds to extrapolate the findings to new environments, which is also referred to as lack of external validity. My concern is how such extrapolations beyond the study environment can be justifiably done.

An example of the use of experimental findings to inform a large-scale policy program is the school-*deworming* program the *Deworm the World Initiative*, initiated by J-Pal. The program treats hundreds of millions of schoolchildren for soil-transmitted helminths (STH) in India, Ethiopia, Vietnam and Nigeria. J-Pal's recommendation for deworming was based on a comparison of RCT-studies of different policies, all designed to foster education by increasing student participation. The results for deworming in Busia, Kenya showed the highest number of additional years of education per 100 US dollars spent (Duflo & Kremer, 2005) of all the compared policies, each carried out in a different locations.

[1] The terms randomized experiment and RCT are here used interchangeably.

These cost-effectiveness comparisons, however, only make sense if the results can be reasonably believed to have the same effects in the target population as they did in the study sample. Esther Duflo and Michael Kremer, two of the founders of J-Pal, argue that, since policy programs "were conducted in similar environments, cost-effectiveness estimates from numerous randomized evaluations can be readily compared" (Duflo & Kremer, 2005, p.213).

This paper argues that instead of simply extrapolating treatment effects across environments based on heuristics, as in the comparison of the cost-effectiveness of deworming, policy decisions can and should be made in a more rigorous way. Causal models can be used to predict the effectiveness of policy interventions in new environments based on experimental data and causal assumptions, themselves informed by theory and data. It is also shown by means of a case study that using causal models is a feasible method for normal economic research.

In section 2, a juxtaposition of experiments and causal models makes clear that while both methods serve to evaluate counterfactuals to assess causal claims, the former establishes a causal claim with a higher internal validity[2], whereas the latter can account better for the external validity.

Section 3 introduces the causal semantics developed by Judea Pearl, a framework which represents a useful tool to handle causality.

In section 4 it will be shown that the Possible Outcome framework, using which randomized experiments are usually dealt, can be embedded in a causal model using Pearl's causal framework. This technical feature yields the possibility of explicitly expressing the causal assumptions, underlying a randomized experiment.

In section 5, it will be shown that the findings of a randomized experiment can be extrapolated, or *transported*, to a different population, by means of a causal model and extensive knowledge about the target population. This will be illustrated with a case study that uses causal calculus (Pearl & Bareinboim, 2014) for a hypothetical transportation of Miguel and Kremer's (2004) results to a target population. The assumptions required for this extrapolation will also be discussed.

Section 6 concludes.

[2] See section 2 for further explanation of these terms.

2. Evaluating counterfactuals: Experimentation vs. modelling

Claims about the effectiveness of a policy program are essentially counterfactual claims (Cartwright & Stetenga, 2011) – they assert the causal effect of the policy intervention in terms of the difference between a counterfactual in which the policy was implemented, and a counterfactual in which the policy was not implemented. Counterfactuals here are possible outcomes, induced by a certain intervention. The corresponding view on causation is formally defined by the manipulability theory. It states that X causes Y if and only if were the value of X to be changed as a result of an intervention, then the value of Y would change (Woodward, 2003, p.15).

This definition of a causal relation demands both of the counterfactuals to be determinable, one induced by a change in X and one where X experiences no change. It raises a problem for the scientist who wants to evaluate the causal effect of X on Y, since it is in principle impossible to observe both counterfactuals simultaneously on the same unit. This problem is referred to as the *fundamental problem of causal inference*. It can be tackled by evaluating causes using the logic of controlled variation and comparison (Guala, 2006). By means of *surgical interventions* conducted on a system, changes are induced in exactly one factor of that system leaving the rest intact. The outcomes in the system from different interventions can be compared to each other, yielding causal claims.

I will consider two approaches that fulfil the task of evaluating counterfactuals, the statistical and the causal approach. The statistical approach observes different outcomes induced through experimentation that simulate the counterfactuals (Heckman, 2005). The experimenter performs different interventions on a system and compares the induced outcomes. The variation of one factor at a time ensures that when the outcomes induced by the different interventions are compared to each other, any correlation between intervention and outcome can be attributed to a causal relation. Hence, perfectly controlled experiments are the ideal setting to identify causes and effects (Guala, 2006).

Take an example to illustrate a perfectly controlled experiment. Say, to save time to write on my dissertation, I wanted to know whether it is faster to cook a lentil stew with a pressure cooker or with a normal pot. To investigate this experimentally, I could conduct an experiment on the pressure cooker – once closing the valve and once without closing the valve, so that it works like a standard pot. Closing the valve cooks the lentils in a fraction of the time than when the valve is open. This suggests a causal effect of closing the valve on cooking time. The counterfactual outcomes were experimentally brought about by holding all factors fixed except the opening/closing of the valve. But *what* factors exactly were relevant, and *how* the effect took place can be ignored.

The causal approach on the other hand, aims to determine counterfactual outcomes by means of a causal model.[3] The model explicitly defines the background of a causal effect, which includes causally relevant factors and causal relations between them representing causal mechanisms. The modeller is informed by scientific theory, which can be drawn from different sources and scientific disciplines (Heckman, 2005). A surgical intervention on a factor is represented in a model by replacing the mechanism that determines the factor with a certain value, leaving the remaining mechanisms intact. The resulting outcomes of that the other factors determine the counterfactual outcome, induced by the intervention.

Going back to my lentil stew, if I wanted to model the counterfactuals that would result from cooking with a pressure cooker, respectively without it, I would have to know a great deal more about the relevant factors and mechanisms that take place when cooking. The model would need to contain all the factors relevant to cooking time, such as the container's volume, its form and material, the temperature etc., and also the causal mechanisms that cause to the lentils to cook: heating up the container brings the liquid inside to boil, closing the valve hermetically seals the container and prevents the steam from escaping; this increases the internal pressure, which in turn permits much higher cooking temperatures, so that the lentils are quickly cooked. On the other hand, when the valve is left open the steam can escape so that no high pressure is build and the temperature is not as high– the lentils cook slowly. A fully specified model (when all parameters of the functions that represent the mechanisms are defined) can calculate the cooking time with a closed, respectively an open valve and so identify the causal effect of using a pressure cooker on cooking time. In fact, the model could calculate counterfactuals for a range of different interventions on the causal factors, e.g. different containers, varying temperature etc.

[3] Heckman (2005) refers to it as the *scientific approach* instead. I find the term *causal* more appropriate, as it does not undermine the statistical approach as unscientific. I thus take a more moderate position than Heckman towards RCTs.

Both approaches, the statistical and the causal, rely on two backdrop assumptions, first, that the underlying causal mechanisms are invariant across space and time. And second, that each mechanism is modular – it can be altered without affecting the overall structure. Comparison of counterfactuals supposes that mechanisms are invariant to different interventions. Further, conducting surgical interventions requires the mechanisms to be modular – changing one mechanism should not change the whole system (Cartwright, 2004). These two assumptions imply the intervention can be assumed to bear constant results, given the that same underlying *causal structure* that organizes the mechanisms is present, and that the *background factors*, unknown causes that directly or indirectly (through other factors) influence the outcome, stay constant.

Observed variances in the outcome do not undermine the invariance of the mechanisms, but are manifest *disturbances* caused by changes in the background factors or in the causal structure (Cartwright, 2007b). The resulting randomness in the outcomes is thus only due to our ignorance about the underlying causal background. This understanding of mechanisms goes back to the Laplacian quasi-deterministic conception of laws of nature. According to Laplace, the stochastic elements of natural phenomena come from our ignorance about the boundary conditions (Pearl, 2009a, p.26). Hence, did we know the fundamental laws of nature, the initial conditions and the boundary conditions of a system we could derive and predict actual phenomena, much like in the deductive-nomological account of scientific explanation (Hempel & Oppenheim, 1948). This is at odds with the understanding of modern physics that all natural laws are indeterministic and determinism is just a practical approximation. But remember, we find ourselves in realm of the social sciences – not quantum mechanics – so it is reasonable to assume quasi-deterministic mechanisms.

Essentially, the statistical and the causal approach both depart from the same theoretical foundations and aim to evaluate counterfactuals, but while the former is a good *hunter* of causes, the latter is the better *user*.[4] If an experiment is properly conducted, it can identify a true causal claims. It is then said to have a high *internal validity.* This does not give us, however, any deeper understanding about the effect. Experimental results on their own are thus only applicable to environments that are

[4] The terms "hunt" and "use" are due to Cartwright (2007a), while hunting a cause refers to the identification of a causal claim, using a cause refers to the application of it for a specific purpose, in our case to evaluate policy-effectiveness in different environments.

identical to the experimental setting. Hence, external validity is not reducible to statistics (Pearl, 2009a; Heckman, 2005). Conversely, a causal model gives understanding about the workings of causal mechanisms and can give predictions about how an effect will behave in a range of environments – this is referred as the *external* validity. But a model's causal claims depend on the assumptions that were built into it. Hence, it does not attain the same degree of objectivity than an experiment does. It will later be shown, that it is possible to use both approaches, so that they balance each other out, by embedding the causal effects evaluated in an experiment in a causal model.

3. Introducing Pearl's semantics

Pearl's semantics (2009a) prove to be particularly useful not only to express formally abstract causal concepts, but also to define causal models and conduct different operations on them. Pearl defines a causal mechanism as a functional equation that determines the effect in terms of its causes – with the effect represented by the dependent variable and the causes by the independent variables. Take the functional equation (3.1); it expresses a causal relation between Y and its causes X, Z and the Y-specific *error term* that captures any disturbances (see section 2) U_Y. It says that if the values of the causes were $X = x$, $Z = z$ and $U_Y = u_Y$, this would cause Y to adopt value y.[5] Note that this definition of mechanisms perfectly reflects the previously mentioned Laplacian quasi-deterministic causality: The functional equation is composed by a deterministic part, the known causes, and a stochastic part, the error term that represents unknown causes.

$$y = f_Y(x, z, u_Y), \qquad (3.1)$$

Alternatively, we can understand a causal mechanism as a conditional probability function, that determines a probability distribution of an outcome variable, conditional on its causes. So, the effect of $X = x, Z = z$ on Y, can be expressed by the probability distribution (3.2). [6]Note, that this understanding of a mechanism also fits with the Laplacian causality, since a probability function is fully deterministic in the marginal cases $P(y|x) = 1$ and $P(y|x) = 0$, reflecting the case where there are no disturbances;

[5] For clarification: Upper case letters, X, stand for variables while lower case letters, x, stand for particular values adopted by the variable.

[6] U_Y does not occur in (3.2) since the probability distribution already captures any disturbances.

and is stochastic when $P(y|x) \in (0,1)$, where there is some disturbance. This paper will stick to the probabilistic understanding of mechanisms.[7] This makes me adapt Pearl's concepts. While he expresses nonparametric models as a special case of parametric models, this paper will only concentrate on the former, so it is not necessary to consider the latter. This will have no consequences on the content, however.

$$P(y|x,z) \tag{3.2}$$

A *causal structure* G is defined as a directed acyclic graph (DAG) on a set of endogenous variables $V = \{V_1, \dots V_n\}$ in which each node corresponds to a distinct variable V_i, and each edge stands for a causal relation (Pearl, 2009a, p.44). The directionality of causal structures reflects the asymmetry of causality. A visual representation of a causal structure in form of a diagram is a straightforward way to express causal relations. The dependency of a variable X on another variable Y is expressed with an arrow $X \rightarrow Y$. Y is then said to be a *parent* of X. For illustration, let us define a structure G^* over a set $\{X, Y, Z\}$ that includes the causal relation $X \rightarrow Y$, and where X and Y are confounded by Z, $X \leftarrow Z \rightarrow Y$, as depicted in in figure 1a).

The causal structure is the blueprint of the *causal model* that encodes the relevant background factors and the structure of the causal mechanisms that bring about of the causal effect. A causal model M is defined as the pair $\langle G, P(v) \rangle$, consisting of a causal structure G and a joint probability distribution over variables in V, $P(v) = P(v_1, \dots, v_n)$.[8] To illustrate, let M^* be the (non-parametric) causal model consisting of the causal structure G^* and the joint probability distribution over the endogenous variables, $P^*(y, x, z)$.

[7] The reason is that in section 5.2 I will define such a model from a real scientific study. For more on partially specified, non-parametric models, see Pearl and Bareinboim (2014, p.582).

[8] Pearl (2009a) defines a model as the tuple $\langle V, F, U, P(u) \rangle$ consisting of set of variables V, a set of functional equations F, a set of error terms U, and a joint probability distribution over U, $P(u)$. However, I will concentrate on partially specified, non-parametric models, in which the lack of knowledge about F and $P(u)$ is expressed as a conditional probability function (see probability function (3.2)).

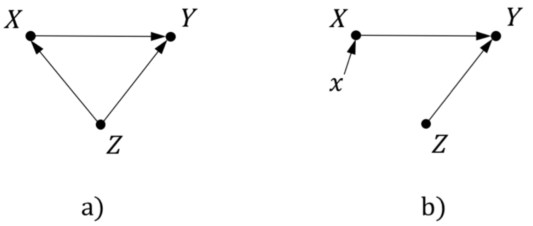

Figure 1a): Causal diagram describing the causal structure G^* . Figure 1b): Intervention $do(X = x)$, eliminates arrows coming to X from its parents.

We can now express a crucial concept of causality which makes the difference between experimentation and observation: The surgical intervention. To express an intervention on a variable $V_k \in V$ in some model M, Pearl defines an operator, the *do-operator* (Pearl, 2009a) $do(V_k = v_k)$ or simply $do(v_k)$. It detaches V_k from its parents, and sets it to a value v_k. Note that an intervention is always defined against a specific causal model, that determines the parents of the intervention variable. Model M on which the intervention has been conducted, becomes the manipulated model $M_{V_k=v_k}$, or simply M_{v_k}. This manipulated model determines the post-intervention joint distribution, which describes how the endogenous variables are distributed after the intervention on V_k, $P(v_1, \dots, v_{k-1}, v_{k+1}, \dots v_n | do(v_k))$.[9]

When an intervention $do(X = x)$ is conducted in our example model M^*, the manipulated model would be given by M_x^*, with a mutilated causal structure as it is depicted in figure 1b). The post-intervention joint distribution for M_x^* is given by distribution (3.3). It can also be written as the product of the individual conditional probabilities as in (3.5).

$$P^*(y, z | do(x)) \tag{3.3}$$

By rules of probability-calculus

$$= P^*(y | do(x), z) P^*\big(z \big| do(x)\big) \tag{3.4}$$

[9] $do(v_k)$ abbreviates $do(V_k = v_k)$.

By the independence of Z from X we can eliminate the do-operator from the last term in (3.3)

$$= P^*(y|do(x), z)P^*(z) \tag{3.5}$$

Based on the post-intervention joint distribution, the outcome of different interventions can be calculated. For instance, the effect that an intervention $do(x)$ causes in Y, $P^*(y|do(x))$, can be calculated from (3.5) by adding up all the weighted Z-conditional post-intervention distributions, as in (3.6).

$$P^*\big(y\big|do(x)\big) = \sum_z P^*(y|do(x), z)P^*(z) \tag{3.6}$$

Given the causal effect (3.6), we can calculate for example the expected value that Y will adopt due to $do(x)$, which is of course the weighted average of $P^*(y|do(x))$.

$$E\big(Y\big|do(x)\big) = \sum_y y\, P^*(y|do(x)) \tag{3.7}$$

4. A causal understanding of randomized experiments

In social contexts, it is practically unfeasible to control for all relevant background factors that could affect the outcome. However, experiments can be conducted by randomizing the treatment assignment and thus controlling for all possible confounders that would lead to biased results. The standard framework to deal with randomized experiments is the possible outcome framework (PO) (Rubin, 1974). In line with the manipulability theory, the PO defines the treatment effect as the difference in possible outcomes induced by different interventions on an experimental unit. Let $X \in \{0,1\}$ be an indicator for a treatment, and Y stand for the outcome. The treatment effect on the experimental unit u is then given by difference (4.1).

$$Y_{X=1}(u) - Y_{X=0}(u) \tag{4.1}$$

Obviously, both outcomes cannot be observed simultaneously on the same unit (see section 2). This problem is solved by randomly assigning a sample of individuals S into two disjunctive sets: a treatment group T, and a control group C. Randomization eliminates any selection bias between T and C. It ensures that individuals $t \in T$ and $c \in C$ are on average the same with respect to all factors other than the treatment, so that confounders are equally distributed across comparison groups. The average outcome approximates the expected value, provided the groups are large enough (by the law of large numbers) and the treatment assignment on one individual has no effect on the outcome of others (SUTVA). [10] Moreover, the counterfactual outcome the treatment group would have adopted had it not undergone the treatment intervention is observable in the outcome of the control group. So, the difference between the average of the observed outcomes, (4.2), can be assumed to give a good estimate of the average treatment effect (ATE). The ATE is a standard measure for the effectiveness of a policy, it describes how the treatment affects the population on average.

$$E(Y_{X=1}(t)) \; - E\big(Y_{X=0}(c)\big), t \in T, c \in C \tag{4.2}$$

4.1.Clinching conclusions at the cost of scope

Cartwright (2007b) labels randomized experiments as *clinchers,* because they clinch the causal claim they make. That is, it follows deductively that the result bears the true effect, when the right conditions are met. The conditions are that the background factors are identically distributed across treatment and control group; and that the same causal structure underlies both groups (Cartwright, 2007b).[11] The virtue of randomization is that it ensures these two conditions are met without requiring any knowledge about the background factors, that help the effect to take place, or about the causal structure, that describes the relevant mechanisms.

However, the importance of causal knowledge for predictions about causal effects, and thus also for the external validity of experimental results, has been stressed in recent philosophical literature (Cartwright & Stetenga, 2011). Knowing the right causal background of a causal effect is crucial to understand how the effect is going to work in different environments. To reasonably expect that a treatment effect is going to work in a *target* population Π' exactly as it did in the study population Π, the same conditions as in the actual experiment must be given (Cartwright, 2007b) expressed in C1-C2.

[10] The stable unit treatment value assumption (SUTVA) assumes that the potential outcomes in an individual are unaffected by the particular treatment of other units (Boesche, 2019). This includes that the counterfactual outcomes are invariant to assignment mechanism of treatment (Heckman, 2005).

[11] This was argued in section 2.

Let $K_1, \dots, K_n$ represent the relevant background factors – causes of outcome Y, other than the treatment X – and $P(k_1, \dots, k_n)$ stand for the probability distribution of the background factors.[12]

Condition C1 (Identical background factors). $P(k_1, \dots, k_n) = P'(k_1, \dots, k_n)$.

Condition C2 (Same causal structure). $G = G'$.

C1-C2 underline that simple extrapolation of experimental results to different environments cannot be assumed without accounting for the causal structure and the background factors. Causal dependencies are highly context-dependent, so to make claims about the external validity of some experimental results C1-C2 should be accounted for. This chimes with Cartwright's (1989) slogan "no causes in, no causes out" – without making causal assumptions we cannot make causal statements.

Although randomized experiments do not require any conditions to be tested for, when it comes to applying the conclusion in an environment outside of the study (external validity), causal assumptions matter. Heckman argues that the lack of external validity of randomized experiments is inherent to the potential outcome framework. The PO only focuses on the observed population level ATE and black-boxes the causal mechanisms that bring about the evaluated causal effect. It offers no framework to extrapolate causal claims to new environments. Causal models, on the other hand, have the advantage of encoding the causal assumptions explicitly (Cartwright, 2007a). The causal approach evaluates counterfactuals based on rigorous causal analysis, accounting for the sources of variance (Heckman, 2005). This is required when it comes to achieving external validity.

4.2. The first law of causal inference

Essentially, randomization mimics a surgical intervention, in that it eliminates selection bias and thereby blocks any lurking confounders (Pearl & Mackenzie, 2018). [13] Any bias between treatment and control group can be attributed to confounders (Pearl, 2009a). Let $P(y|x)$ stand for the probability distribution

[12] Pearl, Glymour, & Jewell (2016) account for background variables with the distribution $P(u_1, \dots, u_n)$, of the error terms, $U_1, \dots, U_n$ of the endogenous variables. In a fully specified model, in which the parameters of the functional equations are known, the value for every variable in the model can be calculated given the distribution of the error terms.

[13] I use the term confounder for factors that affect treatment and outcome at the same time. For a visual representation of a confounder, take Z in figure 1a). Pearl (2018) suggest that confounders should be defined as anything that lead to $P(y|x) \neq P(y|do(x))$.

of Y conditional on the observation $X = x$. In an experiment, confounders are responsible for the variation in treatment-conditional distributions between treatment and control group, $P_T(y|x) \neq P_C(y|x)$. However, when the treatment assignment is randomized, the treatment value is not affected by confounders. The same is true for the do-operator; it replaces the functional equation that determines X with a value x. Hence, randomization conducts the same operation as the do-operator: it blocks incoming arrows to the treatment, most importantly it blocks those coming from confounders.[14] As a result, the treatment-conditional distributions of the outcomes are equalized, which implies that on average the possible outcomes are equal across groups. This implication is expressed in (4.3). The realization that randomization can be expressed by the do-operator indicates that the PO can be expressed in causal terms using Pearl's semantics.

$$P_T(y|do(x)) = P_C(y|do(x)) \Rightarrow E\big(Y_x(t)\big) = E\big(Y_x(c)\big) \tag{4.3}$$

In his *first law of causal inference* (see equation (4.4)) Pearl (2014) subsumes the PO to his causal framework. He expresses the PO's primitive unit, the possible outcome $Y_x(u)$, in causal terms. Namely, as the outcome of Y that is induced by an intervention $do(x)$ on a causal model M. So, to predict the counterfactual $Y_x(u)$, we can take the modified model M_x and solve it for Y. Hence, according to this law a possible outcome in the PO is the outcome of an intervention on an undefined causal model.

$$Y_x(u) = Y_{M_x}(u) \tag{4.4}$$

This implies that a causal model could be defined, that could evaluate the same counterfactuals which were evaluated in the experiment. Hence, the model would yield the experimentally evaluated causal effect. Furthermore, since a causal model is modular, tweaking the causal factors and relations could give predictions about counterfactuals in new, different environments without conducting actual experiments. Causal effects can thus be inferred based on experimental data and a theoretical causal model.

[14] See Figure 1b) for a visualisation of an intervention $do(X = x)$ that blocks the causal links from confounder Z to X.

5. Causal models and transportability

This paper argues that causal models can be used to account for the external validity of an experimentally evaluated treatment effect of a policy intervention. For this purpose it is considered to embed experimental results in a causal model. When done in a rigorous and accountable way, the causal model can be used to make predictions about how a policy intervention is going to fare in a specific target population. This can be of great use to compare the cost-effectiveness of different policies. Hence, causal models could help us make more well-grounded decisions.

Judea Pearl and Elias Bareinboim (2014) develop a type of causal calculus that calculates the *transportability* of an effect to a target population other than the study population.[15] In a causal model, the scientist can explicitly encode the causal assumptions about the structure and the background factors that led to the effect in the study population. By comparing observational data from the study population and from the target population, the scientist can assess differences in the background factors between populations. Further, building on qualitative knowledge about the target population, the underlying causal structure it can be identified. Note that Pearl and Bareinboim's method fits nicely into Cartwright's analysis of external validity, as it takes her two conditions (C1-C2) into account. Moreover, their method not only accounts for the external validity of the causal effect as it is identified in the experiment, but can also make predictions about how the effect will behave in new populations.

The backdrop assumption that allows Pearl & Bareinboim to transport an effect beyond the study population is the same that allows for experimental outcomes to be compared to each other: invariant mechanisms (see section 2). The invariant mechanisms are the "licensing assumptions" for transportability, while possible changes in the causal structure and in the background factors are "threats" (Pearl & Bareinboim, 2014, S. 580).

5.1. How to transport causal effects

A causal model that defines the causal background of an experimentally evaluated treatment $do(x)$, consists of a causal structure on a set $V = \{X, Y, Z\}$, a pre-treatment joint probability distribution,

[15] Transportability is a type of external validity. It refers to the ability to transport experimental findings to a new environment, adjusting for possible differences.

$P(x, y, z)$, and a post-treatment probability distribution, $P(y, z \mid do(x))$, where X stands for the treatment, Y for the outcome and Z stand for some background factor, or *covariate* as they are usually called in statistics and econometrics (see section 3).

Since treatment effects vary systematically with variation of covariates, conditioning for these covariates accounts for systemic differences between populations. Stratifying and reweighting is a statistical method that identifies the ATEs that are specific to a stratum so that they can be weighted again with new weights and then added together. Stratifying the experimental effect gives us the influence of covariates on the effect. The strata-specific effects can then be weighted using the observed covariate-distributions from the target population. Hereby it plays a role how exactly the covariates are causally interrelated. The result is a *transport formula* which adapts the experimentally evaluated effect to the target population. It combines experimental results obtained in the study population Π, $P\ (y|do(x), z)$, with observational data from the target population Π', $P'(z)$, to obtain an experimental claim $P\ '\ (y|\ do(x))$ about the causal effect in Π'.

It must be noted that this transportation-method relies on the assumption that the only source of variation of a causal effect across populations, are varying covariate-distributions, which is admittedly an idealized scenario. It takes as given two assumptions: First, that the policy intervention is implemented *in the same way* in the target population as it was in the study (T1). If the treatment is administered differently, or the components of the treatment are heterogeneous to the experimental setting, it cannot be supposed that the treatment triggers the same mechanisms as it did in the experiment (Hotz, Imbens, & Mortimer, 2005).

> **Assumption T1** (Equivalent treatment implementation). The experimentally evaluated treatment is implemented in the target population in the same way and with respect to the same number of individuals as in the experiment.

The second assumption is that the underlying causal structure is the same (T2). This is a strong assumption, which may not hold easily. The INUS-account tells us that one missing factor can prevent the treatment effect from acting entirely (Mackie, 1974). It is necessary to take into consideration qualitative knowledge about the individuals, the population, the society etc., to check whether this assumption T2 is justified.

Assumption T2 (Invariant causal structure). Both populations share the same causal structure, through which the treatment effect takes place.

5.1.1. Stratifying and reweighting

Experimental results, evaluated in a study population Π, can be stratified and reweighted to adjust them to a target population Π'. First, the covariate distributions of both populations are compared, so that disparities, $P(z) \neq P'(z)$, can be identified. The treatment effect, $P(y|do(x))$, is then stratified to get the covariate-specific treatment effect $P(y|do(x), z)$[16]. In order to stratify by a covariate Z, the study sample is partitioned by the observed values z into exhaustive and disjoint subsets. For instance, if Z represented age, it can be partitioned by age-groups of babies ($z = [0,3]$), children ($z = [4,12]$) teenagers ($z = [13,18]$) young adults ($z = [18,26]$), etc. Whenever a covariate is independent of the treatment the overall treatment effect, $P(y|do(x))$, is equal to the sum of the products of the Z-specific treatment effects, $P(y|do(x), z)$, with their respective probability distribution $P(z)$, as it is shown in equation (5.2).

$$P(y|do(x)) \tag{5.1}$$

(Stratifying by Z, assuming Z is independent from X, i.e. there exists no arrow going from X to Z)

$$= \sum_z P(y|do(x), z)\, P(z) \tag{5.2}$$

(Reweighting using covariate distributions in Π')

$$\sum_z P(y|do(x), z)\, P'(z) \tag{5.3}$$

(Assuming the observed covariates capture any differences between populations, we get the predicted treatment effect in Π')

$$= P'(y|do(x)) \tag{5.4}$$

[16] Recap: $do(x)$ represents the treatment intervention that sets the value of X to x; and z stands for the value that a covariate Z might adopt.

Now, the stratified effects can be *reweighted* by replacing the covariate-distributions from Π, $P(z)$, with the covariate-distribution from Π', $P'(z)$. This way, systematic differences between populations can be captured by adjusting for the disparities in the covariate-distributions, as it is done in (5.3). Hence, the treatment effect in Π', $P'\big(y|do(x)\big)$, is calculated as the Z-specific treatment effect evaluated in Π, $P(y|do(x), z)$ conditioned on the target population's covariate-distribution, $P'(z)$, as shown in (5.3)-(5.4).

A third assumption is required to adjust the effect to the target population by stratifying and reweighting. The covariate-strata in the target population, that are conditioned for, must already have been observed in the study population. Otherwise, if the stratum-specific ATE for the specific stratum, $P(y|do(x), z)$, has not been observed in the experiment in Π, it cannot be conditioned for. The distribution in Π', $P'(z)$, must therefore be overlapped by the distribution in Π, $P(z)$. For instance, if I want to reweight the effect for people with an age of eighty, but the RCT was only conducted on twenty-year-olds, then I cannot identify the ATE for eighty-year-olds, since I do not know how the treatment affects people with age eighty.

Assumption T3 (Overlapping covariate distributions). The values z of covariate Z that are observed in the target population Π', have already been observed in the study population Π.

5.1.2. Considering causal structure

Equation (5.3) expresses in terms of Pearl's semantics how Hotz, Imbens, and Mortimer (2005) account for external validity, which Muller calls "conditional external validity" (2015, p.217). It simply consists in stratifying and reweighting of the experimental effect, by conditioning on the covariate-distributions, as it is done in section 5.1.1. This method, however, does not always give the right transport formula, since it does not consider how the covariates are causally related to the treatment.

For stratification and simple reweighting (without considering causal relations) to yield the right transport formula, the covariates must be independent from the treatment variable. Equation (5.5) expresses this condition. The step from (5.1) to (5.2) is otherwise not possible as such. This, however, is not considered by Hotz, Imbens and Mortimer, which is not surprising as they do not use causal language with which they could formulate causal assumptions explicitly. But to control for any variable that could affect the outcome, regardless of its causal relation with the treatment variable, is seriously misguided (Pearl & Bareinboim, 2014).

$$P(z|x) = P(z) \tag{5.5}$$

I will show the importance of accounting for the causal structure by means of an example. In the causal structure illustrated in figure 1a) the equation (5.5) holds true - no arrow goes from X to Z. The reverse must not necessarily be the case. If our structure resembles the one in figure 1a), then equation (5.3) is indeed the right transport formula. Hence, simple stratifying and reweighting suffices for this case.

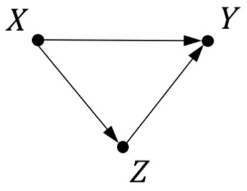

Figure 2: Causal structure in which Z is affected by the treatment.

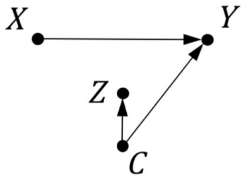

Figure 3: Causal structure in which Z has a common cause with Y.

Consider the causal structure illustrated in figure 2. Equation (5.5) does not hold true here. There exists one arrow going from X to Z. Hence, the transport formula (5.3) that would be calculated with simple reweighting does not apply here. The reason is that Z is not independent from X. The calculation of the transport formula differs from before, insofar that the covariate-specific treatment effects,

$P(y|do(x), z)$., must be weighted with the intervention-conditional covariate distributions, $P(z|do(x))$. We therefore get (5.6) as the correct transport formula (see Appendix A.1 for detailed calculation).

$$P'(y|do(x)) = \sum_z P(y|do(x), z)\, P'(z|x) \tag{5.6}$$

Now consider figure 3, which shows a structure in which Z does not affect Y nor is itself affected by it. Z is correlated with Y but only due to the common cause C. In this case, controlling for Z at all would be a mistake, since it is not causally relevant for the effect $X \rightarrow Y$. Assuming that C is equally distributed in the study and the target population, $P(c) = P'(c)$, in this case the right transport formula is given by the unaltered treatment effect (5.7). It is apparent that in order to attain the right transport formula, it is necessary to take the causal structure into consideration.

$$P'(y|do(x)) = P(y|do(x)) \tag{5.7}$$

5.2. Transporting the effect: From Busia to Uttar Pradesh

I claimed at the beginning of this paper that causal models can give reliable predictions which lend themselves for comparison of the cost-effectiveness of different policy interventions. To show that building a model is feasible in practice, I will run through a hypothetical case of transportation. It is based on Edward Miguel and Michael Kremer's (2004) study which evaluates the effects of school-based deworming on school-attendance. The RCT was conducted in the county of Busia, Kenya. The experimental findings of that study will be transported to a different population, for which observational data is available. The hypothetical target will be the state of Uttar Pradesh in India which is densely populated and poor like Busia and has one of the highest prevalence of STH among schoolchildren in the country. This makes a deworming policy intervention especially worthwhile, since it may significantly improve children's education. Note however, that this implementation of the policy would mean an immense scale-up from the randomized experiment in the county of Busia, that was conducted on about thirty thousand schoolchildren, to a program in Uttar Pradesh with millions of schoolchildren.

To go about this exercise systematically, I will follow Pearl's scheme of causal inference (Pearl, 2012). First, a set of qualitative causal assumptions describing the mechanism is determined, so that the causal structure underlying the effect of deworming on school-attendance can be defined. The causal information is extracted from the actual study by Miguel and Kremer. Second, the causal query is determined. We pose the question how the effect of X on Y is going to behave in the target population, given certain changes concerning the causal background. The answer will take the form of a transport formula, that adjusts the experimental effect to the new environment of the target population. Third, the necessary observational and experimental data about the pre- and post-intervention joint probability distributions of the relevant variables in the study and the target population are gathered. The causal structure together with the joint probability distributions in the structure will form the causal model, with which the transport formula can be calculated.

5.2.1. Building the causal model

Building a causal model requires a lot of knowledge and data about the population. Computational methods can do the trick. The TETRAD program for instance, (Glymour, Schemes, Spirtes, & Kelly, 1987) tests models against data to determine their accuracy using artificial intelligence. It seems reasonable, however, to assume that the scientists conducting experiments, have some idea about the causal background of the evaluated effect that might be sufficient to define a causal model, when their assumptions are backed by theory and observational data. This might not reflect the same objectivity that the internal validity that randomization achieves, but this does not mean that building a model could not be done in a rigorous way.

To avoid unnecessary complications, I will give a simplified version of Miguel and Kremer's study. The study results suggest a direct causal effect of deworming children on their school-attendance. This is consistent with existing literature on the topic that confirms a link between health and education. Children, infected with intestinal worms are often sick, listless and have a hard time concentrating. Deworming may improve their state of health, which allows them to attend school regularly. It also improves the children's ability to concentrate, which makes attending school even more worthwhile.

Additionally, the study finds that there are positive effects from deworming on school-level that carry over to children going to other, untreated schools in the region. Taking these spill-overs into account makes the positive effect of deworming on overall school-attendance even higher than previously thought. Given the worms are transmitted from person to person through soil, the local population density may affect the risk of getting an STH-infection. Moreover, since children go fishing and do other activities together, while often not wearing shoes and defecating openly (which helps the worms to spread) they are more likely to get infected from other children. A higher fraction of dewormed individuals, however, means that the local population has a lower environmental exposure to STH, leading consequently to a lower number of worm burdens. Therefore, Miguel and Kremer assume that the worm burden of schoolchildren depends on the total children-density in an area, as well as on the number of dewormed children in that same area. So, they presume that cross-school externalities are likely to increase with the ratio of treated children to total number of children in a certain area.

To account for these cross-over externalities in the RCT, a treatment set and a control set of school groups are first randomized, and then the deworming treatment assignment is randomized again on a school-level as follows. Let Π be the set of all schools in Busia. Groups of nearby schools in Π are randomly assigned into either the set of control groups, C, or the set of treatment groups, T. Within the groups in T, which schools are dewormed and which schools are not, is again subject of randomization. The dewormed schools will form a subset $D \subset \mathrm{T}$, while the remaining schools that have not been dewormed will form a subset $N \subset \mathrm{T}$. This way the outcome, due to the direct effect of deworming on the school-attendance, is observed in D, while the externalities are observed in N. Both outcomes can then be compared to the outcome in control set C, which gives us the overall causal effect.

The group-level randomization also generates an exogenous variation of the local density of treated pupils, while the school-level randomization generates an exogenous variation of the school-sizes where deworming was assigned, so that cross-school externalities can be estimated for different density-levels of dewormed children and for different school sizes. This will be important later, when the effect is transported to a different population.

To define the endogenous variables of the causal structure, the regression equation (5.8) that was used in the actual study to estimate the result of the RCT gives a good base.[17] The treatment variable X stands for the *treatment assignment of a school* and the outcome variable, Y, stands for the *local overall school-attendance*. X and Y may be confounded by *unobserved exogenous variables* U, e.g. by

[17] Regression equation (5.8)is a simplification of the actual regression equation used by Miguel and Kremer (see equation 1 in Miguel & Kremer, 2004, p.175).

socioeconomic status. Y is further affected by Q^T, which represents the *local density of dewormed schoolchildren* and by Q, which stands for the *local density of total schoolchildren.*

$$Y = \alpha + \beta X + \gamma Q^T + \delta Q + U \tag{5.8}$$

From the discussed qualitative information about the causal relations and equation (5.8) the causal structure on the variables can be defined (see figure 4). The direct effect of X on Y can be represented by a direct arrow, $X \to Y$, which is due to the dewormed children going to school more often. As expressed in (5.8) the treatment and the outcome might have unobserved confounders U that affect both – they are represented with a two-headed arrow between X and Y. The cross-school externalities work through Q^T and Q. The ratio of Q^T to Q influences the worm burden in the region which in turn affects the region's school attendance. The treatment assignment X affects Q^T by increasing the number of dewormed children in the region, which in turn lowers the local worm burden, resulting in a higher school-attendance Y. Q affects Y as well, by having a positive effect on the local worm burden. Hence, the causal effects due to externalities are depicted in figure 4 by the causal relations, $X \to Q^T, Q^T \to Y, Q \to Y$.

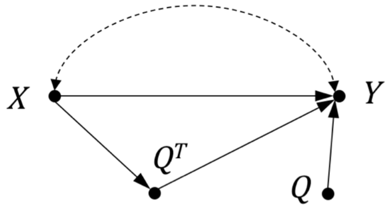

Figure 4. Causal structure describing the effect of school-deworming, X, on school attendance, Y. The dotted, two-headed arrow represents the unobserved, confounding variables U.

Now the query can be stated: *How will the effect of school deworming on school attendance evaluated in Busia behave in Uttar Pradesh?* To answer this question the pre-treatment joint probability distribution (5.9) and the post-treatment joint probability distribution (5.10) of the variables are observed in the study sample Π. The distributions can be also written as the product of the individual conditional probabilities, where the probability distribution of each variable is conditional to its parents (as it is done in (5.9) and (5.10)).[18]

$$P(y, x, q^T, q) = P(y|x, q^T, u)\ P(x, u)\ P(q^T|x)P(q) \tag{5.9}$$

$$P(y, q^T, q|do(x)) = P(y|do(x), q^T, u)\ P(q^T|do(x))\ P(q) \tag{5.10}$$

5.2.2. Calculating the transport formula

Let Π' be the set of schools in Uttar Pradesh. Having now defined the causal model describing the experimentally evaluated effect of deworming on Π, it can be used to predict the effectiveness of deworming in Π'. Observational data alerts us to the discrepancies between populations that should be accounted for. Say, a comparison of the joint distributions of the relevant variables in Π, $P(x, y, q^T, q)$ and Π', $P'(x, y, q^T, q)$, draws to our attention two discrepancies. First, the population density in Π' turns out to be on average higher than in Π (D1). A higher number of total schoolchildren may lead to lower externalities, since holding the number of dewormed schoolchildren fixed, increasing the total schoolchildren leads to an increased risk of worm infections. This leads to a smaller decrease in the worm burdens, which in turn results in a smaller increase in school attendance.

Assumption D1 (Higher population density in the target population). $P'(q) > P(q)$.

The second observation in the data is that the ratio of children to primary schools is on average lower in Π' than in Π. This means that deworming a school has a lower effect on Q^T, as fewer children are dewormed in one school. Both observed differences between the populations can be accounted for by the transport formula.

Assumption D2 (Lower effect of deworming on local density of dewormed schoolchildren in the target population). $P'(q^T|x) < P(q^T|x)$.

But before the transport formula for this task can be calculated, transportability assumptions T1-T3 must be checked for. First, it is assumed that the implementation of the intervention in Π' is

[18] Recap: parents are immediate causes in the model, see also section 3.

equivalent to the one that was conducted in Π (T1). This is crucial for the externalities to be able to take place the same way they did in Π. If, say, *all* schoolchildren in a region were dewormed, there could not be any externalities on untreated schoolchildren – since there are none. Hence, only a subset of Π' is randomly assigned to deworming, call it D'. Dewormed schools are thus equally distributed throughout the state.

Second, it is assumed that Π' shares the same causal structure with Π. This way the mechanisms that produced the effect $P(y|do(x))$ in Π, can be expected to do so in Π' too (T2). One might be concerned about T2, considering the relatively big scale-up in the number of dewormed schools. It is assumed that deworming children has no macro-effects that would influence the school-attendance. If this were not the case however, it could lead to different causal structures in Π and Π'.

The third assumption is that the probability distributions that showed discrepancies in Π', $P'(q^T|x), P'(q)$, must be overlapped by the respective distribution in Π, $P(q^T|x), P(q)$, so that the effect can be reweighted (T3).

Equipped with the causal model, the transport formula can be finally calculated, for the population discrepancies D1-D2 as it is shown in equations (5.11) - (5.14).

$$P\big(y\big|do(x)\big) \tag{5.11}$$

(Stratifying by Q and Q^T gives us the causal effect for different values of q and q^T. We assume that Q is independent from X, and that Q^T and X are unconfounded, as it is encoded in the model)

$$= \sum_{q,q^T} P(y|do(x), q, q^T)\, P(q) P(q^T|x) \tag{5.12}$$

(Reweighting using density distribution in Π')

$$\sum_{q,q^T} P(y|do(x), q, q^T)\, P'(q) P'(q^T|x) \tag{5.13}$$

(If the observed covariates capture all differences between populations, we get the predicted treatment effect in Π')

$$= P'\big(y\big|do(x)\big) \tag{5.14}$$

Transport formula (5.13) gives us an estimation of how the treatment effect will perform in Π'. Based on it, effectiveness indicators, like the ATE can be calculated, as it is shown in (5.15).

$$E\big(Y\big|do(x)\big) = \sum_{y} y\, P'(y|do(x)) = \sum_{y,q,q^T} y\, P(y|do(x), q^T, q)\, P'(q) P'(q^T|x) \qquad (5.15)$$

The predicted ATE (4.15) is arguably a qualified prediction of the effectiveness of deworming in Uttar Pradesh. It has been shown that the transport formula of the experimental effect is a valid conclusion using and Pearl and Bareinboim's causal calculus and given the causal assumptions embedded in the causal model, the basic assumptions for transportability T1-T3, and the assumptions about the population disparities D1 and D2. An important question, therefore, is whether the assumptions made are reasonable.

5.3. Discussion of assumptions

To evaluate counterfactuals using models, certain assumptions need to be made (see sections 5.1 and 5.2). When the assumptions are wrong our predictions are will be wrong as well. But there is no getting around making those assumptions. If we are not willing to make assumptions to account for external validity, we are doomed to use heuristic guidelines instead, which is less than optimal. However, possible threats to the assumptions should be considered, so that they can be reasonably held.

T1 concerns the implementation of the policy, which depends on the specific circumstances and is mainly a practical issue. For T2, extensive knowledge about the study and target population must be available, so that it can be identified which causal structure was present in the study and which causal structure can be expected to be present in the target population. When a policy program is implemented in the target population on a bigger scale than in the experiment which evaluated the policy's effect, as it is the case in the Deworm the World Initiative, this scale-up could affect the causal mechanisms. Assumption T2 might be the weakest link in a transportation inference, as it is the most complicated to establish and the most prone to error, so it should to considered in-depth. T3 is relatively straightforward to check for by observational data. A study could even be designed such that it purposely identified the ATE for a variety of strata, so that assumption T3 would be given.

The population disparities D1 and D2 were chosen on purpose such that a transportation of the effect could be conducted using Pearl and Barenboim's causal calculus, without getting too deep into technical details. Not every disparity between populations can be accounted for in form of a transport formula, just as not every target population is suitable for a certain causal effect to take place, since certain invariances across populations must be given. Pearl and Bareinboim define a criterion that determines whether the disparities between study and target populations can be accounted for, based on the specific causal structure (see Pearl & Bareinboim, 2014, p.589). Hence, using causal calculus it can be assessed beforehand whether the right conditions are given and the disparities are accountable.

Some see it as a downside that the required assumptions to estimate causal effects via a causal model demand a pronounced knowledge about the causal background. Criticism, voiced by Deaton and Cartwright (2018), argue that using causal models "takes us back" to the situation that was meant to be avoided by randomized experiments, since it requires for the underlying causal structure to be completely specified. Pearl (2018) responded to this criticism by stating, first, that causal models are more suited to handle the task of transporting causal effects. Put briefly, RCTs are designed to neutralize confounding between treatment and outcome, while Pearl's and Bareinboim's calculus is designed to neutralize differences between populations. And second, that since the completeness of their framework has been shown (Bareinboim & Pearl, 2012) the minimum assumptions required to establish sound estimates of causal effects can be determined.

A point I would add to Pearl's response is the apparently intuitive way in which scientists reason in causal terms. When conducting RCTs, scientists must have a quite elaborate idea about the mechanisms a causal effect relies on – even if this might not be explicitly stated. Take for instance the causal structure defined in section 5.2 (see figure 4) which has been derived in a straightforward way from information taken from Miguel and Kremer's study (2004), although this was arguably not their intention to build a causal model. Hence, I would argue that explicitly stating the causal reasoning applied by the scientists, does not "take us back". On the contrary, it takes us forward, towards the ultimate aim of evidence-based policy – reaching conclusions to inform policy-makers, based on assumptions that are open to scrutiny under scientific investigation (Pearl, 2018). To understand the causes that produce effects so that predictions can be made, the causal approach is arguably the best and most comprehensible approach to take.

6. Conclusion

To summarize, it is argued in this paper that causal models can be used to account for the external validity of causal claims evaluated by randomized experiments. This replies to concerns voiced about randomized experiments that the external validity of their results was not warranted. Causal models lend themselves to predict the effectiveness of RCT-evaluated policies in populations beyond the study sample. These predictions allow to make rigorous comparisons of cost-effectiveness between policies, based on theory and observational data so that for a specific population the optimal policy can be decided upon.

It is further argued that scientists, when conducting randomized experiments to evaluate a policy's effectiveness, have already implicit causal knowledge that can inform a causal model. Nonetheless, it must be noted that causal claims based on causal models are only as good as the model's assumptions. If the assumptions are wrong, the conclusion is going to be wrong as well. However, it is argued in this paper that making causal assumptions is necessary to do predictions about policy-effectiveness. External validity is not reducible to statistics but needs to be dealt with as a scientific process in which theories and causal considerations play an undeniable role.

7. Appendix

A.1 (Derivation of the transport formula (4.8) for the causal effect in the model of Figure 2).

$$P\big(y\big|do(x)\big)$$

(Stratification by Z, taking into account that Z is dependent from X)

$$= \sum_z P(y|do(x), z)\, P(z|do(x))$$

(Reweighting using covariate distributions in Π')

$$\sum_z P(y|do(x), z)\, P'(z|do(x))$$

(Assuming that Z and X are unconfounded, we can equate $P'(z|do(x))$ with $P'(z|x)$)

$$\sum_{z} P(y|do(x), z)\, P'(z|x)$$

(Assuming that the observed covariates capture any differences between populations, we get the predicted treatment effect in Π')

$$= P'(y|do(x), z)$$

8. References

Bareinboim, E., & Pearl, J. (2012). Transportability of causal effects: Completeness results. *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*.

Boesche, T. (2019). Reassessing Quasi-experiments: Policy Evaluation, Induction, and SUTVA. *The British Journal for the Philosophy of Science*.

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

Cartwright, N. (2004). Causation: One Word, Many Things. *Philosophy of Science , 71*(5), pp. 805-819.

Cartwright, N. (2007a). *Hunting Causes and Using Them: Approaches in Philosophy and Economics.* Cambridge: Cambridge University Press.

Cartwright, N. (2007b). Are RCTs the Gold Standard? *BioSocieties, 2*, pp. 11–20.

Cartwright, N. (2012). Will this policy work for you? predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), pp. 973–989.

Cartwright, N., & Stetenga, J. (2011). A Theory of Evidence for Evidence-Based Policy. *Proceedings of the British Academy, 171*, pp. 289–319.

De Regt, H. W. (2017). *Understanding Scientific Understanding.* New York: Oxford University Press.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature, 48*(2), pp. 424-455.

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

Duflo, E., & Kremer, M. (2005). Use of Randomization in the Evaluation of Development Effectiveness. In G. e. Pitman, & W. B. development (Ed.), *Evaluating development effectiveness* (Vol. 7). New Brunswick, N.J.: Transaction Publishers.

Duflo, E., & Kremer, M. (2008). Use of randomization in the evaluation of development effectiveness. In W. Easterly (Ed.), *Reinventing Foreign Aid* (pp. 93–120). Washington, DC: Brookings.

Glymour, C., Schemes, R., Spirtes, P., & Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling.* Orlando: Academic Press.

Guala, F. (2006). *The Methodology of Experimental Economics.* New York: Cambridge University Press.

Heckman, J. J. (2005). The Scientific Model of Causality. *Sociological Methodology, 35*, pp. 1-97.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science, 15*, pp. 567-579.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics, 125*(1), pp. 241–270.

Mackie, J. (1974). *The Cement of the Universe: a Study of Causation.* Oxford: Oxford University Press.

Miguel, E., & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, 72*(1), pp. 159-217.

Muller, S. M. (2015). Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *The World Bank Economic Review, 29*(Supplement), pp. 217 – S225.

Pearl, J. (2009a). *Causality: Models, Reasoning, and Inference* (second edition ed.). New York: Cambridge University Press.

Pearl, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys, 3*, pp. 96–146.

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling.* New York: Guilford Press.

Pearl, J. (2014, November 29). *Causal Analysis in Theory and Practice: On the First Law of Causal Inference*. Retrieved August 12, 2019, from http://causality.cs.ucla.edu/blog/index.php/2014/11/29/on-the-first-law-of-causal-inference/

Pearl, J. (2018). Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science and Medicine, 210*, pp. 60-62.

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

Pearl, J., & Mackenzie, D. (2018). *The book of why.* New York: Basic Books.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer.* Chichester: Wiley.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), pp. 688-701.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* New York: Oxford University Press.