CENTER FOR LABOR ECONOMICS
UNIVERSITY OF CALIFORNIA BERKELEY
WORKING PAPER NO. 55

ASSESSING COUNTERFACTUALS WHEN TREATMENT IS MULTIVALUED*

Jochen Kluve**

September 2002

Abstract. Matching estimators have received substantial attention in the recent literature on causal inference in observational studies. While causal inference has traditionally focused on the case of binary treatment, recent results regarding the generalized propensity score have made possible explicit extensions to cases where treatment is multivalued. This paper goes back to the foundations of the causal model that underlies this approach, and gives a characterization of the causally meaningful counterfactuals that can be assessed in the multivalued framework. The theoretical analysis shows that specification of the no-treatment state plays a particularly important role, and that treatment-worlds are not necessarily equidistant, as the model assumes. An application from program evaluation using propensity score matching illustrates these points.

JEL classification: C10

Keywords: Matching estimators, generalized propensity score, causal inference, counterfactuals, possible worlds.

---

* I thank David Card, Hans Gersbach, Ruth Miquel, Christoph Schmidt, and participants of the 12th $(EC)^2$ conference on "Causality and Exogeneity in Econometrics", held December 2001 in Louvain-la-Neuve, for valuable comments. Financial support from IZA and the VolkswagenStiftung is gratefully acknowledged.
** UC Berkeley, AWI Heidelberg, and IZA Bonn. Address for correspondence: Jochen Kluve / University of California Berkeley / Center for Labor Economics / 549 Evans Hall #3880 / Berkeley CA 94720 / kluve@econ.berkeley.edu

1. Introduction

Matching estimators have received substantial attention in the recent literature on causal inference in observational studies. The basic idea of matching is to mimic a randomized experiment ex post. In economics, the majority of applications has been in the context of the evaluation of so-called active labor market policy, such as employment and training programs. The main ingredient to matching is adjustment for pre-treatment variables, which are commonly referred to as covariates. In this connection, the propensity score has received particular attention. The propensity score, i.e. the probability of receiving the treatment given pre-treatment variables, is an alternative method of adjustment, and was initially proposed by Rosenbaum and Rubin (1983, 1984). In recent years, the economics literature has increasingly adapted this statistical procedure and has discussed matching methods extensively in both empirical and theoretical work.[1]

Matching methods are cast into the framework of a specific causal model that has become known as the Potential Outcome Model (POM). The POM for causal inference describes a setting in which units are potentially exposed to a set of treatments, and have corresponding outcomes or responses associated with each treatment. The causal connection of interest is the effect on the outcomes of some particular treatment relative to some other particular treatment, often called "control" treatment. Since in reality each unit can only be exposed to one treatment, the other treatment states and associated potential outcomes for the single unit are counterfactuals.

In its essence the POM dates back to the work of Neyman (1923 [1990], 1935) and Fisher (1935). Fisher is commonly credited for the invention of randomized experiments, while Neyman was probably the first one to use a model for a treatment effect in which each unit has two responses. Contributions to the development of the model include Cox (1958), Cochran (1965), and above all Rubin (1974, 1977), who was the first to apply the potential outcome framework to observational studies (cf. Rosenbaum, 1995, for further discussion, and Freedman, 1999, for some reflections on the history of association and causation in statistics). Due to Rubin's contributions the POM is frequently referred to as the "Rubin Model". Related work in economics are

[1] Cf., for instance, Hahn (1998), Angrist and Hahn (1999), Hirano, Imbens and Ridder (2000) for efficiency issues. Matching applications using the propensity score are manifold, cf. in particular Lechner (1999, 2000), Dehejia and Wahba (1999), but also Heckman, Ishimura and Todd (1997), Smith and Todd (2002), or – for an application using exact matching – Kluve, Lehmann and Schmidt (1999).

models for switching regressions (Quandt 1958, 1972) and the earnings model of Roy (1951).

In general the POM allows for a finite number of treatments, but both theory and practice have focused on merely two elements, "treatment" and "control". This is intuitively appealing, as a causal effect can only be inferred for one treatment relative to some other treatment. However, recent results have made explicit extensions to multivalued treatment settings in observational studies possible (Imbens 2000, Lechner 2001), and some papers have already applied these methods (Larsson 2000, Gerfin and Lechner 2000). This paper goes back to the foundations of the causal model, and gives a characterization of the causally meaningful counterfactuals that can be assessed within the multivalued framework. The theoretical analysis shows that specification of treatment states, especially the no-treatment state, plays a particularly important role, and that treatment-worlds are not necessarily equidistant, as the model assumes. An application from program evaluation using propensity score matching confirms these points.

The remainder is organized as follows. Section 2 presents the basic causal model, and its applicability using the propensity score for binary treatment and multivalued treatment. Section 3 assesses the various counterfactuals arising in the multivalued case, and discusses both definition of treatment states, specifically the no-treatment state, as well as distance between worlds. Section 4 illustrates these points with an application from program evaluation. Section 5 concludes.

2. The Causal Model

2.1 Basic Setup

The logical elements of the POM are a quadruple of the form {U,T,D,Y}.[2] These elements constitute the primitives of the model. U is a population of N units $[u_1,...,u_n]$, T is a set of M treatments $[t_1,...,t_m]$ to which each one of the units u may be exposed, D(u)=t

[2] This section follows the discussion of the POM in Holland (1986, 1988), and uses what could be called "statistical notation" of a model for causal inference. This terminology has been frequently criticized by econometricians (Leamer 1988, Heckman 2000). On the other hand, statisticians have continued to hold that this approach to causal inference is the most lucid one (Rubin 1990). Despite the controversy, I find it safe to assume – in particular given the substantial recent convergence on issues of causal inference across disciplines (documented in Greenland 2000) and the logical equivalence of the POM and recursive structural equation models (established in Galles and Pearl 1998) – that this notation is comprehensible for everyone.

indicates that unit u is actually exposed to a particular treatment t out of T, and Y(u,t) equals the value of the outcome that would be observed if unit $u \in U$ were exposed to treatment $t \in T$. U and T are sets, D is a mapping of U to T, and Y(.) is in general a real-valued function of (u,d).

The response variable Y depends on both the unit u and the treatment t to which the unit is exposed. If u were exposed to some $t_1 \in T$, the observed value of the outcome would be $Y(u,t_1)$, and if u were exposed to some $t_2 \in T$, the observed response would be $Y(u,t_2)$. The meaning of Y to be a function of pairs (u,t) is that it represents the measurement of some characteristic of u after u has been exposed to $t \in T$, implying that it must be possible for any unit in U to be potentially exposed to any treatment t out of T. This condition entails a certain notion of what is a cause, preventing us from interpreting associational relations as causal ones, like, e.g., associations between sex and income or between race and income.

Call Y the outcome function and let $Y_t(u)=Y(u,t)$. The mapping D is called the assignment rule because it indicates to which treatment each unit is exposed. The observed outcome of each unit $u \in U$ is given by $Y_D(u)=Y(u,D(u))$, which is the value of Y that is actually observed for unit u. Therefore, the pair $(D(u),Y_D(u))$ – where D(u) indicates the treatment in T to which u is actually exposed – constitutes the observed data for each unit u. Note the distinction between $Y_D(u)$ and $Y_t(u)$: While the former is the outcome actually observed on unit u, the latter is a potential outcome being actually observed only if D(u)=t. The basic causal parameter of interest is

(1) The unit-level treatment effect (UTE):

The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$.[3]

There are three things to note about this definition. First, the causal effect $UTE_{tc}(u)$ is

[3] Note that the two treatments in this definition are denoted with t (like "treatment") and c (like "control"). This hints at the idea of randomized assignment of units into an experimental treatment or control group. It also gives a particular flavor to the definition of an effect of one treatment relative to a control treatment, where the term "control" usually implies "no treatment". Moreover, note that the notation E_{tc} is meant to indicate the causal effect of "t relative to c".

defined at the individual-unit level. Second, $UTE_{tc}(u)$ is the increase in the potential value of $Y_t(u)$ over the potential value of $Y_c(u)$. Third, $UTE_{tc}(u)$ is defined as the causal effect of t relative to c. Since it is impossible to simultaneously observe $Y_t(u)$ and $Y_c(u)$, the causal effect $UTE_{tc}(u)$ is never directly observable. Holland (1988) emphasizes how the POM makes the unobservability of the causal effect explicit in separating the observed pair (D,Y_D) from the function Y.

2.2 Applicability

Following the exposition of the basic model, I will first discuss one crucial assumption and subsequently review the conditions under which it is possible to assess potential outcomes and infer meaningful causal statements.

The stable-unit-treatment-value-assumption (SUTVA) is the pivotal assumption ensuring that the causal framework of the POM is adequate in practice. SUTVA is advocated by Rubin (1980, 1986) to play a key role in deciding which questions are formulated well enough to have causal answers. It is the a priori assumption that the value of Y for unit u when exposed to treatment t is the same independent of (1) the mechanism that is used to assign t to u, and (2) what treatments d the other units $v \neq u$ receive, and that this holds for all n units within U and m treatments within T. SUTVA is violated when, for instance, there is interference between units that leads to different outcomes depending on the treatment other units received, i.e. Y_{tu} depends on whether $v \neq u$ received t or some other $d \in T$, or there exist unrepresented versions of treatment or versions of treatments leading to "technical errors" (Neyman 1935)[4], i.e. Y_{tu} depends on which (unintended) version of treatment t unit u was exposed to.

Unit homogeneity is a name given by Holland (1986) to the assumption that the responses of all units to a particular treatment are the same, i.e. $Y_t(u)=Y_t(v)\ \forall\ u,v \in U$, $t \in T$. This is a partial specification of Y in that it restricts the values that Y can take on

[4] For further detail cf. Rubin (1980) and Rubin's (1990) discussion of Neyman (1923 [1990]). In Neyman's work the notion of potential outcomes is based on the methodological discussion of agricultural experiments (cf. also Speed 1990). In that context, possible violations of SUTVA are apparent: How should one avoid neighboring plots treated differently (by, e.g., different fertilizers) to "interfere" given nature's powers (wind, rain etc.), or in how far can one claim that each bag of fertilizer represents exactly the same treatment as any other bag of fertilizer (Rubin 1986)? Moreover, as Rubin (1990) points out, interference between units can be a major issue when studying medical treatments for infectious diseases, or educational treatments given to children who interact with each other.

but does not specify them completely. The assumption is only likely to be justified if one can claim to be working with a homogeneous sample. Under unit homogeneity, the causal effect of treatment t relative to a treatment c is given by $UTE_{tc}(u)=Y_t(u)-Y_c(v)=Y_t(v)-Y_c(u)$ for any two distinct units u and v in U. In this case, UTE_{tc} is a constant and does not depend on the unit under scrutiny. Evidently, unit homogeneity solves the fundamental problem of causal inference in that one only needs to measure the two (observable) outcomes $Y_{D(u)=t}(u)$ and $Y_{D(v)=c}(v)$ for two units u and v to infer the causal effect of t relative to c on <u>any</u> unit within U.

Unless unit homogeneity holds, individual effects are impossible to observe. Thus, one of the most important causal parameters is the average causal effect of a treatment, as it represents a useful summary of the unit-level treatment effects[5].

(2) The <u>average treatment effect (ATE)</u> of treatment $t \in T$ relative to treatment $c \in T$ is the expected value of the unit-level difference $Y_t(u)-Y_c(u)$ over all $u \in U$: $ATE_{tc}=E(UTE_{tc})=E(Y_t-Y_c)=E(Y_t)-E(Y_c)$.

The ATE is an unobserved quantity, since expectations of Y for both t and c are taken over the full range of U. In practice it is only possible to observe D(u) and $Y_D(u)$ over U, and therefore only the joint distribution of D and Y_D rather than D and $\{Y_t: t \in T\}$. The average value of the observed outcome Y_D among all those units actually exposed to a particular treatment $t \in T$ can be written as $E(Y_D|D=t)$. For the two particular treatments t and c this becomes $E(Y_D|D=t)=E(Y_t|D=t)$ and $E(Y_D|D=c)=E(Y_c|D=c)$, respectively. These two quantities are always observed in the data, and yield the following parameter.

(3) The <u>prima facie average treatment effect (FATE)</u> of treatment $t \in T$ relative to treatment $c \in T$ is the difference in average responses between those units actually

[5] In practice, further questions arise as to whether it is e.g. the "average treatment effect on the treated", or the "average treatment effect on the population" etc. that is the causal parameter of interest, and how each of these can be estimated. These questions have given rise to an extensive "treatment effect literature" (Heckman 2000). See e.g. Heckman (1992) and Heckman, LaLonde, and Smith (1999) for discussion, and Angrist, Imbens, and Rubin (1996a) for Instrumental Variables in the POM and identification of the "local average treatment effect" (LATE).

exposed to t and those units actually exposed to c: $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)$.[6]

The distinction between FATE and ATE emphasizes the fact that the quantity that is always computable from the data (FATE) does in general not equal the quantity about which one desires to draw inferences (ATE). This results from the difference between $E(Y_t)$ and $E(Y_c)$ on the one hand and $E(Y_t|D=t)$ and $E(Y_c|D=c)$ on the other hand. The two quantities are only equal when independence holds. Suppose that the determination of which treatment a unit is exposed to is statistically independent of all other variables, in particular the response function. Following common practice using Dawid's (1979) notation of independence "$\perp\!\!\!\perp$", this can be written as $D \perp\!\!\!\perp \{Y_t: t\in T\}$.

(4) If $D \perp\!\!\!\perp \{Y_t: t\in T\}$, then the prima facie average treatment effect of treatment $t\in T$ relative to treatment $c\in T$ is equal to the average treatment effect of t relative to c: $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)=E(Y_t)-E(Y_c)=ATE_{tc}$.

The independence assumption is the key point to the applicability of the model, as it allows inference on the unobserved causal parameter of interest, the ATE, directly from the FATE, which one can always compute or estimate from the data.

Under which conditions is independence likely to hold? The most probable case in practice is a randomized experiment, in which – coarsely speaking – units are randomly assigned to different treatments, so that the initial population and the subpopulations in the treatments do not differ from each other on average. This makes (4) likely to hold, yielding the ATE from the FATE. Holland (1988) describes the relation between randomization and independence as follows: Independence is an assumption about the data collection process, i.e. about the relation of D and Y over the population U, while randomization is a physical process that gives plausibility to the independence assumption in many important cases. For instance, if U were infinite, then the law of

[6] Holland (1988) calls this parameter "prima facie average causal effect FACE". It is not to be confused with a "prima facie cause" as defined by Suppes (1970) in his probabilistic theory of causation (cf. Suppes (1970) for the original definition and e.g. Sobel (1995) or Salmon (1980) for a discussion): Given two time values t and t* with $t<t^*$, the event c_t is a prima facie cause of the event e_{t^*} if $Prob(e_{t^*}|c_t)>Prob(e_{t^*})$, i.e. c temporally precedes e and is positively relevant to it. Cf. also Skyrms (1988) for a discussion of the relation between probability and causation.

large numbers together with randomization would imply that (almost) every realization of D would be independent of $\{Y_t\}$. However, randomization does not necessarily make independence plausible in each and every case, as randomization does not assure that each and every experiment is "adequately mixed", but only that "adequate mixing" is probable (Leamer 1983). To take the simplest example, imagine that U consisted only of very few units. Then the plain physical act of randomization would not render the independence assumption plausible.

The meaning of populations that do not "differ" from each other and "adequate mixing" in randomized experiments becomes clear through introducing other variables into the model. So far Y was the only variable measured on the units u – apart from the treatment indicator D. Let us now add a variable X to the model, where X can be real-valued or vector-valued. In principle, X(u,t) is defined on $U \times T$ and depends on both u and t. However, there is a special class of X-variables that are of specific interest, as defined in Holland (1988):

(5) X is a <u>covariate</u> if X(u,t) does not depend on t for any $u \in U$.

If we consider specifically the values the X-variables take on <u>prior to treatment</u>, then the X-variables are always covariates. Randomization <u>on average</u> guarantees balancing of covariates – observable and unobservable – across subpopulations in different treatments, which in turn makes the independence assumption plausible, so that (4) holds and the ATE can be inferred from the FATE. In the words of Rosenbaum and Rubin (1983): With "properly collected data in a randomized trial", X is known to include <u>all</u> covariates both used to assign treatments and possibly related to the response $\{Y_t\}$.

Covariates are of particular importance in the model in cases in which there is no randomization and one cannot arrange the values of D(u) to achieve independence. In such an <u>observational study</u> the interest remains in inferring causal effects of treatments, but now D is not automatically independent of $\{Y_t\}$. Given (observable) covariate(s) X one could check the distribution of X for subgroups in each treatment by comparing the values of Prob(X=x|D=t) across the values of $t \in T$ (Holland 1988). If there is evidence that Prob(X=x|D=t) depends on t, then the independence assumption may not appear

plausible. Instead, in the nonexperimental setting one usually builds on a weaker conditional independence assumption which says that treatment assignment and the response are conditionally independent given a vector of covariates:

(6) [Rosenbaum and Rubin 1983:] Treatment assignment is strongly ignorable if the response $\{Y_t: t \in T\}$ is conditionally independent of treatment assignment D given the observed covariates X, i.e. $\{Y_t\} \perp\!\!\!\perp D|X$, and $0<Prob(D=t|X=x)<1$.

Strong ignorability is the basis for all causal inference on covariate-adjusted treatment effects in observational studies (Holland 1988). Adjusting for covariates yields the covariate-adjusted prima facie average treatment effect (C-FATE) based on conditional expectations:

(7) $C\text{-}FATE_{tc}=E\{E(Y_t|D=t, X) - E(Y_c|D=c, X)\}$.

Just like the FATE, the C-FATE does in general not equal the desired ATE. This only holds under conditional independence.

2.3 The Propensity score (i): Binary Treatment

For the two-treatment case $T=\{0,1\}$, Rosenbaum and Rubin (1983) show that (6) also holds for a balancing score B(X) defined as a function of the observed covariates X such that the conditional distribution of X given B(X) is the same for the exposure groups (D=t), i.e. $X \perp\!\!\!\perp D|B(X)$. Rosenbaum and Rubin (1983) identify all functions of X that are balancing scores, the most trivial one being B(X)=X, and the coarsest one being the propensity score.

(8) The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables, i.e. P(x)=Prob(D=1|X=x). If treatment assignment is strongly ignorable given the covariates, then treatment assignment is strongly ignorable given the propensity score, i.e. $\{Y_t\} \perp\!\!\!\perp D|P(x) \; \forall \; t \in T$.

This result is of particular interest in practice, as it reduces the potential problem of conditioning on a high-dimensional X to conditioning on a scalar, provided that P(X) is known.

2.4 The Propensity score (ii): Multivalued treatment

Imbens (2000) and Lechner (2001) extend this result from the case of binary treatment to the case of multivalued treatment, in which T contains M treatments $[t_1,...,t_m]$.

(9) The generalized propensity score is the conditional probability of receiving a particular level of treatment given the pre-treatment variables, i.e. $R_t(X)=Prob(D=t|X=x)$. If treatment assignment is strongly ignorable given the covariates, then treatment assignment is strongly ignorable given the generalized propensity score, i.e. $\{Y_t\} \perp\!\!\!\perp D|R_t(X) \; \forall \; t \in T$.

Thus, it is possible to estimate average outcomes by conditioning solely on the generalized propensity score. This result opens up the enormous potential of propensity score matching in cases when treatment is multivalued.

3. Possible Worlds

The delineation of the POM in the previous section has shown that the model is based on causation in counterfactual terms, since for each unit all outcomes except one are not observed, i.e. they are counterfactual outcomes. A rigorous theory of causation in terms of counterfactuals has been developed for the first time by Lewis (1973b). In deriving logical properties of counterfactual conditionals, i.e. counterfactual statements, Lewis shows that causal dependence between two events a and b exists, if – given that a and b are actual occurrent events – if a had not occurred, then b would not have occurred. The framework for this analysis is given by the idea of closest possible worlds (Lewis 1973a): From the perspective of the actual world in which a and b occur, the counterfactual conditional "If a had not occurred, then b would not have occurred" is logically true if the

world in which they do not occur is the closest possible world to actuality.[7]

This (simplified) exposition of Lewis's theory extends to the POM in a straightforward fashion. In the POM, the actual world is given by some treatment t_1 (=event a) that results in an outcome Y_1 (=event b). The counterfactual conditional that describes the closest possible world translates to "If t_1 had not occurred, then Y_1 would have not occurred". What is the closest possible world that makes this counterfactual true? In the case of binary treatment, there is only one possible world, i.e. the world in which t_2, the control treatment, occurs, and yields outcome Y_2. This world is mechanically closest and the comparison between the two worlds makes the counterfactual true and therefore t_1 and Y_1 causally dependent.

In the case of multivalued treatment this is the same for any two treatments. As causal inference in the POM relates two treatments and their respective outcomes, the two worlds set in relation are always closest, and the counterfactual conditional is automatically true. Appendix A delineates this result in detail. The bottom line here is that causation in terms of counterfactuals is based on the idea of closest possible worlds (cf. also Robins and Greenland 2000), where the causal effect is defined by the difference between the two outcomes in an "actual" world and a "closest" world.

3.1 The no-treatment state

To formalize this idea, let W denote a set of M worlds, i.e. $W=\{W_0,W_1,\ldots,W_{m-1}\}$, such that each world $W_i \in W$ is defined by a particular treatment t_i and the associated outcome Y_i. Since $W=W_0 \cup W_1 \cup \ldots \cup W_{m-1}$ and $W_i \cap W_j = \varnothing$ for all $W_i, W_j \in W$, W is meant to consist of exactly M mutually exclusive treatment-worlds. This captures the notion that there is no interference between treatment-worlds, and no unrepresented versions of treatment exist.

Moreover, let Ω denote the universal set comprising all possible worlds that differ only with respect to these two elements, so that $W \subseteq \Omega$.[8] In general, W does not need to equal Ω, if we regard W as entailing just those worlds where the types of treatment t_i can

[7] "Actuality" means "world of point of view", i.e. "actual world" refers at any world W_i to that world W_i itself.

[8] This account still considers a finite number of treatments. For an extension of the model to the case where the set of treatments is not finite see Pratt and Schlaiffer (1988).

either be controlled or at least observed, i.e. W is meant to comprise those worlds with well-defined types of treatment. The complement W' of W is then given from $\Omega=W\cup W'$ and contains treatment-worlds outside the controlled or observed domain. For the complement, too, it is in principle possible to construct valid comparisons and infer causal relations, as W' as a whole can always be defined recursively as "anything that is not W".

Taking into account the complement W' illustrates the distinction between what could be called a controlled control treatment and an uncontrolled comparison treatment. Consider the case of binary treatment, i.e. M=2 and the causal effect of interest is that of treatment t relative to treatment c. In a randomized medical trial, for instance, where t is the medicament under study and c is a placebo, c represents a controlled control treatment. It is (a) controlled by the experimenter, and (b) a distinct alternative treatment in its own right, which is not merely characterized by the absence of t. In this case, $W=\{W_t,W_c\}$ and $W'=(W_t\cup W_c)'$, where W' entails some unspecified treatment(s) outside W characterized by not given the medicament and not given the placebo. Of course, W' might not be of interest in the study, or we might not even be able to obtain any information about it. On the other hand, consider the case of an observational study in labor economics, for instance, evaluating some government training program (=treatment t). In this case, the alternative treatment c is characterized retrospectively by the absence of training, so that c represents an uncontrolled comparison treatment. It is (a) not under control of the researcher, and (b) not defined on its own, but just by the absence of t. In this case, $W=\{W_t\}$ and $W'=W_c$.

Note that the distinction between "controlled control treatment" and "uncontrolled comparison treatment" does not imply that the one produces valid causal inference, and the other does not. It is well known that placebos are used to learn something about "not given the medication", and in that respect may perform better than "actually not given the medication", since with placebos the control units cannot be influenced by knowing that they are not given the medication. Use of placebos ensures that the response is to the treatment itself, not the idea of treatment. Hence, the controlled control treatment gives a well-defined no-treatment state, while the uncontrolled comparison treatment remains

more vague.[9]

This distinction highlights that it is important to know which treatment characterizes a specific world, and in particular which world gives a sufficient description of the no-treatment state. In the binary case, experiments usually guarantee a well-defined setup. If experimental data is not available, the treatment-world W_t is the only one that is specified, and the causal effect is inferred relative to the complement W'. This defines the no-treatment state as "any other alternative to treatment", and it may not always be straightforward to ensure that this no-treatment state, recursively defined, contains no confounding element. In the case of multivalued treatment, experiments also tend to ensure well-defined treatment-worlds, and W' retains the interpretation of "outside" worlds. Consider for instance a dose-response experiment, in which several treatment groups are exposed to different levels of a medicament, and one group is given a placebo. In an observational study with various treatments the no-treatment state is still defined recursively through the absence of all other treatments. Because now several treatments are well-specified, the uncertainty regarding the no-treatment state decreases, and it is less likely to contain confounders. However, due to the assumption that many treatments are now well-specified, this case does not retain the convenient interpretation of treatment t relative to anything-that-is-not-t.

3.2 Counterfactuals and Causal Effects

Within the set of treatment-worlds W the causal effect of treatment t_i relative to treatment t_j (with $t_{i,j} \in T$, $T \subset W$) is given by

$$\Delta_{ij} = Y_i - Y_j \tag{10}$$

which follows from the counterfactual analysis outlined above. Let W_0 denote the world

[9] Experimental settings do not necessarily imply a well-defined control treatment. While this is usually a straightforward exercise in medical experimental studies of the type described above, it is far more difficult in experimental studies in labor economics, e.g., due to the length of treatment (several months of participation in a training program) and the difficulty of defining a proper alternative. One example is the experimental evaluation of the National Supported Work Demonstration (NSW) in the US: "Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group were left to fend for themselves." (LaLonde 1986, emphasis added)

specifying the no-treatment state, i.e. the absence of any treatment, be it controlled or uncontrolled, defined uniquely or recursively. Then, in the binary case, $W=\{W_0,W_1\}$ with $W_0=W_1'$ and

(11) $$\Delta_{11'} = Y_1 - Y_{1'} = Y_1 - Y_0 = \Delta_{10}$$

Almost all causal inference studies focus on this basic case with two treatment-worlds that differ solely by the treatment under study, and the comparison world is characterized by the no-treatment state which equals the absence of treatment. In the case of multivalued treatment there are at least two "real" treatments besides W_0. This has several important implications. First, consider particular treatments t_i, t_j, t_k and the following simple decomposition:

(12) $$\begin{aligned} \Delta_{ij} &= Y_i - Y_j \\ &= Y_i - Y_k - Y_j + Y_k \\ &= (Y_i - Y_k) - (Y_j - Y_k) \\ &= \Delta_{ik} - \Delta_{jk} \end{aligned}$$

Of particular interest is the special case where $t_k=t_0$.

(12a) $$\Delta_{ij} = \Delta_{i0} - \Delta_{j0}$$

(Of course, if $t_k \neq t_0$ in (12) we can only use the decomposition if $M>3$). Expression (12) shows that any causal comparison between two treatments is implicitly always related to any other baseline-treatment within W. The case in which the no-treatment state is the baseline (12a) is of particular interest, since causal effects are usually inferred relative to the absence of treatment. This relating of causal comparisons between two treatments – neither of which is the no-treatment state – to the no-treatment state is also necessary to

identify the level of effects. [10]

For the binary case property (11) has shown that the causal comparison of some treatment t_i relative to the absence of t_i equals the comparison of t_i to the no-treatment state. As outlined in the previous section this does not hold for a causal comparison of t_i relative to t_i' in the multivalued case. There are two aspects to the t_i-versus-t_i' relation in this context. First, we have the basic result that $\Delta_{ii'} \neq \Delta_{i0}$ because $t_i' \neq t_0$ and $Y_{i'} \neq Y_0$. This can be seen when we consider what the effect of t_i relative to t_i' actually is:

$$
\begin{aligned}
\Delta_{ii'} &= Y_i - Y_{i'} \\
&= Y_i - F(Y_0, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_{m-1}) \\
&= Y_i - \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k Y_k
\end{aligned}
\tag{13}
$$

where $\sum w_k = 1$ and, for instance,

$$
(13a) \quad w_k = \overline{w} = \frac{1}{M-1} \qquad \text{or} \qquad (13b) \quad w_k = \frac{P(t = t_k)}{\sum_{r=0, r \neq i}^{M-1} P(t = t_r)} = \frac{P(t = t_k)}{1 - P(t = t_i)} \; .
$$

The causal effect of treatment t_i relative to t_i' as given in (13) is thus the difference in outcomes under t_i and t_i' (first line), which equals the difference between the outcome under t_i and some function of the outcomes under all other treatments except t_i (second line), which could in an empirical application equal the difference between the outcome under t_i and the weighted sum of all other outcomes (third line). The function of the outcomes under all other treatments in W_i' gives the <u>absolute counterfactual</u> to treatment t_i, as it is a summary expression of all counterfactual possible worlds. Examples of weight functions for empirical work are given as (13a) equal weights, and (13b) the

[10] If the effect of t_i is positive relative to no-treatment, and the effect of t_j is negative relative to no-treatment, then the effect of t_i is strongly positive relative to t_j. Looking only at the last effect does not reveal the negative effect of t_j relative to no-treatment. Similarly, the effect of t_i relative to t_j could be positive, but still the effects of both of them could be negative relative to no-treatment.

probability of exposure to a particular program (that is not t_i) relative to the sum of probabilities of exposure to any program that is not t_i. The absolute counterfactual of (13) as a summary measure for the effect of some treatment relative to all other treatments can also be represented using a weighted aggregate of the pairwise causal comparisons between the particular treatment and all other treatments:

$$(14) \qquad \Delta_{ii'} = Y_i - \sum_{k=0,k\neq i}^{M-1} w_k Y_k = \sum_{k=0,k\neq i}^{M-1} w_k (Y_i - Y_k) = \sum_{k=0,k\neq i}^{M-1} w_k \Delta_{ik}$$

This expression retains the causal interpretation of the effect of treatment t_i relative to the hypothetical state of random exposure to any other program that is not t_i. Lechner (2002) uses this expression and calls it the composite treatment effect.

The second aspect to the t_i-versus-t_i' relation is that the complements to particular treatments cannot be used as a common baseline, i.e.

$$\Delta_{ij} \neq \Delta_{ii'} - \Delta_{jj'}$$

because clearly

$$\begin{aligned}
\Delta_{ii'} - \Delta_{jj'} &= (Y_i - \sum_{k=0,k\neq i}^{M-1} w_k Y_k) - (Y_j - \sum_{l=0,l\neq j}^{M-1} v_l Y_l) \\
&= Y_i - Y_j - \sum_{k=0,k\neq i}^{M-1} w_k Y_k + \sum_{l=0,l\neq j}^{M-1} v_l Y_l \\
&= \Delta_{ij} - (\sum_{k=0,k\neq i}^{M-1} w_k Y_k - \sum_{l=0,l\neq j}^{M-1} v_l Y_l)
\end{aligned}$$

Table 1 presents an overview of different causal queries and the corresponding counterfactuals. In the binary case (M=2), either (a) $t_0=t_1$' or (b) $t_0\neq t_1$'. The first case (a) is the usual one, and applies for observational studies. The second case (b) comprises two possibilities depending on a relevance criterion. On the one hand, if $t_0\neq t_1$', so that there

exists a W' world besides $W=\{W_0,W_1\}$, and t_1' is considered irrelevant for some reason, such as t_0 being explicitly specified, like in an experimental study, then this implies that W' is irrelevant. On the other hand, if one has reason to believe that $t_0 \neq t_1'$ and if W' is relevant, then there are two further possibilities: Either (i) one has some usable information about W', then this converts to the multivalued case, or (ii) one does not have such information, which hints at a violation of SUTVA because there exist unrepresented versions of treatment. Usually (as in the agricultural setting of Neyman 1923 [1990]) one thinks of unrepresented versions of treatment as unrepresented versions of the "actual" treatment – in this case, however, T' comprises unrepresented versions of the no-treatment state (cf. also section 3.3).

< Table 1 about here>

For the multivalued case (M>2), as there are several well-defined treatments, assume that there is a specific t_0 (possibly defined via the absence of all other treatments) and thus $W=\Omega$. Table 1 depicts some possible counterfactual comparisons. First, the causal effect of a particular treatment could be inferred relative to the no-treatment state. As in the binary case, this would usually be the causal question of interest. Second, one could construct the causal comparison of a particular treatment relative to any other treatment within W. In interpreting the effect it should then (a) be pointed out why this is considered to be a causal question of interest, and (b) be noted that any other treatment (besides the two we relate) can be used as baseline. The most relevant baseline is the no-treatment state, and should be considered in order to identify the level of the inferred effect. The third possible counterfactual for M>2 relates a specific treatment t_i to its absolute counterfactual, a function of the outcomes of all other treatments except t_i. This infers the causal effect of some treatment relative to (an appropriate combination of) all other alternative treatments. This could be a weighted average as given in (13). In a sense this is similar to the "t_1'=anything-that-is-not-t_1"-case for M=2, with the difference that now it is "everything", not "anything", expressing the fact that all alternative treatments are well-defined – and that the corresponding outcomes can therefore be appropriately weighted in an empirical study. Regarding the absolute counterfactual, it can be of

particular interest to compare the no-treatment state to the summary over all other treatments to infer whether the introduction of the overall set of treatments yielded any positive response.

Finally, it should be noted that one could of course construct many more counterfactuals within this model. For instance, one could use causal relations between treatments as a baseline for causal relations between other treatments, or construct the comparison between a particular treatment and a weighted combination of some, but not all of the alternative treatments, etc. That, however, is pure mechanics, and I suppose it might be difficult to conceive the exact causal interpretations of such counterfactuals.

3.3 Illustration

In the M=2 case, why can it can be insightful to distinguish a known or well-defined no-treatment state ($t_0 \neq t_1'$) from a no-treatment state defined merely by the absence of treatment ($t_0 = t_1'$)? Imagine a researcher evaluating some government training program in an observational study. She constructs some retrospective comparison group defined by not having participated in the program. However, training usually takes time. Assume an average of three months in this example. What did comparison group units do during that time? Remain unemployed, continue job search, do nothing, take private training course, etc.? Maybe some of that, maybe all of that, maybe none of that. In most cases, the data do not contain any information in this regard. Thus, since it is impossible to open this black box, one needs to make some assumption about the comparison treatment. It is then fairly convenient to define the no-treatment state as just that, the absence of the treatment under study. The causal effect is that of the training program relative to any other possible (but unobserved) alternative action the program participants would have engaged in had they not participated. Clearly, this is quite different from the explicit specification of the no-treatment state in an experimental medical study (t_0=placebo).

An observational study by Larsson (2000) uses the propensity score matching approach for multivalued treatment developed by Lechner (2001) to evaluate labor market programs in Sweden. In the study M=3, the treatments being Youth Practice (YP), Labor Market Training (LMT), and non-participation. In personal communication with the author the interpretation was given that the no-treatment state comprises a state of job

search rather than non-participation. This finding has several implications: (a) If one has usable information to distinguish job searchers from non-participants, this converts to a case of M=4 with treatments YP, LMT, job search, non-participation. (b) If in fact all individuals in the no-treatment state are in job search, this changes the counterfactual question, and causal inference is on the effect of YP (or LMT) relative to job search, and not relative to non-participation. (c) If the no-treatment state comprises both individuals in job search and non-participants, this hints at a violation of SUTVA.

3.4 Distance

Since the counterfactual account of causation, and therefore also the POM, are based on the notion of closest possible worlds, the question arises what can be said about the proximity of worlds. From the counterfactual logic of causation outlined above, and detailed in Appendix A, it follows that the POM constructs closeness a priori. There is no "quest for the closest possible world" to infer the causal effect of one treatment to some other treatment. Rather, in comparing two treatment-worlds, the model mechanically makes the counterfactual true and establishes closeness.

As a theoretical construct, the causal model consists of equidistant worlds, because worlds differ only with respect to treatment and associated outcome, and are otherwise identical. In an application of the model under randomization, equidistance still holds in principle. This is because randomization ensures that the subpopulations that are exposed to each treatment do not differ from each other. The logic is as follows. In an application, all treatment-worlds, each defined by a treatment and an outcome, are represented by the population exposed to that treatment and responding with the respective outcome. It is therefore the population in each treatment that characterizes the specific treatment-world. If worlds did possibly differ in respects other than treatment and outcome, it would have to be through differences in the population that define each treatment-world. Since the POM produces valid causal inference only when worlds do not differ (besides treatment and outcome), the populations that define the worlds must not differ from each other. How could populations differ from other? In terms of their covariates. If covariates do not differ, and thus populations are identical, then the

treatment-worlds defined by these populations are identical (apart from treatment and outcome), and are thus equidistant.

The distance between worlds finds expression in the distribution of covariates across populations in different treatments. If these distributions are the same (or balanced), the only possible distance between worlds could arise from distance between treatments. In the principal model there is no such distance. In applications, there may well be. Consider the case of a dose-response medical treatment with 3 levels of treatment and a no-treatment state. For instance, 4 groups of patients are exposed to the following amounts of a drug, (I) 100mg, (II) 200mg, (III) 400mg, and (IV) a placebo (0mg). Under randomization, all 4 treatment groups are identical in every respect besides treatment level. Imagine the drug is supposed to decrease systolic blood pressure, and the outcome variable takes on values (IV) 150 for the no-treatment group, and (I) 147, (II) 145, and (III) 144, respectively, for the other groups. The effects relative to the no-treatment group are $\Delta_{I,IV}=-3$, $\Delta_{II,IV}=-5$, and $\Delta_{III,IV}=-6$, which would imply that treatment (III) is the most effective. However, in this example treatments (I) to (III) represent different doses of the same drug. In order to consider the respective dose, and thus consider the respective distance of each treatment level to the no-treatment state, one could weigh each effect by the size of the dose, yielding $\Delta_{I,IV}=-3/100=-0.033$, $\Delta_{II,IV}=-5/200=-0.025$, and $\Delta_{III,IV}=-6/400=-0.015$, which would imply that treatment (I) was the most effective one.[11]

Hence, in applications, distances between treatment-worlds are likely to play a role. Under randomization, since covariates are identical across treatment populations, these distances are solely expressed in the level of treatment. If the treatments do not have an ordering as in a dose-response setting, treatment worlds will be equidistant, unless weights on the outcomes are deliberately imposed. In a randomized experiment, this could find expression in different population shares across treatments, i.e. different participation probabilities. But even if the population were randomized into treatment populations with different probabilities, there would only be the possibility but not the necessity to use these probabilities, since covariates are balanced nevertheless.

[11] Clearly, this example is not supposed to make sense in medical terms, but is only meant for illustration.

In an observational study, covariate balance is rarely given, and it is the aim of matching estimators to assure that treatment groups do not differ from each other in, at least, observable characteristics. If they achieve this goal (and selection is indeed on observables), all results from the randomized context hold, i.e. worlds differ only by treatment and are otherwise equidistant. However, using matching in the multivalued context usually achieves only pairwise congruence of the distribution of covariates, or the propensity score. If comparing one treatment to more than one alternative treatment (such as, for instance, in the absolute counterfactual), the distributions of covariates are unlikely to be balanced across all treatments. However, given the probability of participation in each treatment, i.e. the propensity score, it is possible to weigh each treatment-world accordingly, using, for instance, expression (13b). Such an empirical procedure appropriately reflects distances between treatment worlds.

4. Application

This section applies the POM to the evaluation of active labor market policy. Since the application is merely meant to illustrate the mechanisms of the POM, it reduces propensity score matching methods in the multivalued treatment case to its essentials, in order to emphasize the points made in the previous section. For a comprehensive account of detailed statistical and econometrical issues in the application of this method see Lechner (2001, 2002).

The data contain M=4 treatment states, a non-participation state (t_0), labor market training (t_1), wage subsidy scheme (t_2), and public work, i.e. direct employment in the public sector (t_3).[12] The sample comprises N=6,037 observations, of which 121 participate in t_1, 275 in t_2, 49 in t_3, and 5,592 are in the non-participation state. The latter group is only specified as being unemployed, and registered at the local labor office. The outcome variable Y captures post-treatment labor market performance, expressed by average employment rates in the 9 months succeeding treatment. The outcomes across

[12] The data come from the PLFS (Polish Labor Force Survey) 1996, which contains a special questionnaire on participation in active labor market programs and individual labor force status histories. The programs and their effectiveness have been discussed at length in Kluve et al. (1999).

the initial groups are Y_0=0.547, Y_1=0.499, Y_2=0.238, Y_3=0.361. Simple comparisons between these outcomes yield the FATE. Table 2 depicts these pairwise comparisons. Clearly, since we are just comparing sample averages, the pairwise effects are symmetric. The no-treatment state shows a positive effect relative to any other treatment. Table 2 illustrates that, when comparing for instance t_1 (training) and t_2 (wage subsidy), the resulting positive effect (a 26.1% increase in employment rates) cannot reveal that both treatments show negative effects relative to the no-treatment state.

< Table 2 about here >

Table 3 compares each treatment with a function of outcomes of all other treatments given in equation (13). This absolute counterfactual is a hypothetical comparison between a specific treatment and a synthetic state in which units would have been randomly assigned to one of the other treatments. Equal weights according to (13a) are used, since in this simple FATE comparison there is no reason to assume distances between treatment-worlds to differ across treatments. Table 3 shows that the non-participation state dominates the combination of all other treatments. A policy implication of such a finding would be that apparently the introduction of the whole set of programs did not yield any positive effects. The absolute counterfactual effect for training is also positive, while the wage subsidy program fares worse than a combination of its alternatives. For public work, the result is inconclusive.

The main point is that – as shown in section 3.2 – the absolute counterfactual specifies an alternative state to any particular treatment t_i. It is the complement t_i' composed of all other particular treatments. The no-treatment state t_0 also specifies an alternative state to t_i. Whereas in the binary case these two alternatives are the same, a comparison between Tables 2 and 3 shows the substantial differences that can arise in the case of multivalued treatment. For instance, relative to the no-treatment state training displays an insignificant, possibly negative, treatment effect (Table 2). When comparing training to its complement, however, the result is that training appears to have a significantly positive effect (Table 3). This reflects the positive impact of training relative to both the wage subsidy and the public work program. Both counterfactuals – the no-

treatment state as well as the complement – have a coherent causal interpretation, but the comparison highlights that they answer expressly different questions.

< Table 3 about here >

So far this illustration has only considered the FATE parameters, i.e. naïve comparisons of sample averages. However, as section 2 has pointed out, this does in general not equal the causal parameter of interest, the ATE, unless independence holds. In an observational study, this implies the necessity to adjust for covariates based on the assumption that treatment assignment is strongly ignorable. Using the result due to Imbens (2000) and Lechner (2001) outlined in section 2.4 it is sufficient to condition on the estimated generalized propensity score. The score can be estimated using a multinomial choice model. Subsequently, the generalized propensity score is used for pairwise matching of all treatment groups, thus adjusting for differences in covariates. The algorithm applied here is a simplified version of the prototypical algorithm given in Lechner (2002).

Table 4 depicts the estimates of a multinomial logit model. The base category is non-participation. The variables "unemp1", "unemp2" etc capture pre-treatment employment histories over the four quarters preceding treatment in indicating in which quarter an individual had been unemployed. Employment histories have been identified as a particularly important determinant of selection into programs (cf. Card and Sullivan 1988, Heckman and Smith 1999, Kluve et al. 1999). Since the start of treatment is not defined for non-participant units, these units are randomly assigned a pre-treatment history from the distribution of histories for treated units. Table 4 shows that individuals with little education are less likely to participate in training, whereas a high level of education and being unemployed only in the last quarter preceding treatment is positively associated with participation in training, relative to non-participation. For selection into the wage subsidy program a history of unemployment seems important, in particular long-term unemployment (variable "unemp1234"). Moreover, men and individuals with low education are more likely to participate in the wage subsidy program. In the public

work program, men are also predominant, while the employment history does not give clear indications.

< Table 4 about here >

The coefficients from the model are used to predict the generalized propensity score, which in turn is used in pairwise matching across treatment groups. Table 5 presents the resulting average treatment effects ATE. The row/column relation indicates the direction of matching, i.e. rows indicate the "treatment" sample, columns the "control" sample. Clearly, since sample sizes differ substantially and the algorithm adjusts for that in allowing for multiple use of observations if the control sample is smaller than the treated sample, the direction of matching matters and effects are not symmetric. While, however, most of the effects are at least symmetric in qualitative terms, the negative effect of non-participation versus training could not be reproduced as a reverse positive effect in the comparison of training relative to non-participation. Usually the ATE of a treatment relative to the no-treatment state is the causal parameter of interest. Irrespective of the direction of matching, Table 5 shows that both the wage subsidy and the public work program have negative effects on the post-treatment employment rate relative to non-participation. As an implication for the model, the asymmetry of pairwise effects shows that worlds are not necessarily as identical as the model assumes.

< Table 5 about here >

Table 6 displays the absolute counterfactual treatment effects for the matched samples. The upper part applies equation (13) and compares each treatment relative to a synthetic combination of all other treatments using equal weights (13a). This implies equidistance between worlds. The lower part of Table 6 applies the probability weights given in equation (13b) and therefore reflects different distances between treatment-worlds expressed in the generalized propensity score. The results clearly show that the distance between treatment-world plays an important role in assessing counterfactual

treatment effects. While in a pairwise comparison distance does not matter, a comparison of outcomes across treatment worlds must and can reflect the fact that treatments have varying "distance" from each other. As delineated in section 3.4, this distance can be expressed in terms of the covariates, i.e. the generalized propensity score. The results using probability weights thus appear more credible, since they consider – from the perspective of each treatment group – how close the alternative treatment-worlds are.

< Table 6 about here >

As a result, this streamlined application of matching methods using the generalized propensity score has shown that many subtle issues arise for causal inference in the case of multivalued treatment. Besides well-known problems such as the specification of a multinomial choice model, this extends to the definition of the no-treatment state, or other alternative states, and to appropriate weighting schemes to express distance when assessing counterfactual treatment effects across treatment-worlds. These are fundamental issues that have to be addressed in any study using this method.

Another practical issue that should be mentioned regards the use of a binary versus a multinomial choice model. Lechner (2002) shows that in an applied observational study it does make a difference whether one assesses t_i' as $W'=W_0$ using a binary probability model or t_i' as $W'=\{W_0 \cup W_1 \cup ... \cup W_{i-1} \cup W_{i+1} \cup ... \cup W_{m-1}\}$ using a multinomial probability model. The first results in an insufficient specification of the alternative state by aggregating groups into one alternative group without taking into account the different composition of subgroups, while the second appears to correctly disentangle the desired absolute counterfactual. This finding emphasizes the importance attributed to the definition of W' in section 3.1.

5. Conclusion

It is well-known that causal inference for binary treatments can be relatively straightforward using a randomized experiment, but that it can be a formidable task when

only non-experimental data is available. Recent developments in matching estimators have made explicit extensions to observational studies on causal inference for multivalued treatment possible. Whereas this opens up a large potential for applied research, a set of further complications arises. This has been well recognized regarding identification and estimation issues involved (Lechner 2002). This paper has gone back to the foundations of the causal model that underlies these studies, namely a model of causal dependence in counterfactual terms based on relations between closest possible worlds, and has demonstrated which causally meaningful counterfactual questions can be asked, and answered. As a result, the specification of the no-treatment state, and any other alternative state, plays a particularly important role in the case of multivalued treatment. Furthermore, the principal idea of equidistant worlds, on which the model is based, does in general not hold in practice, and distances between worlds need to be taken into account. An application from program evaluation has illustrated these points.

These are fundamental issues that need to be addressed, and they point to the fact that particular care is needed in modeling causal relationships. Philip Dawid's often-cited observation that "causal inference is one of the most important, most subtle, and most neglected of all the problems of statistics" (Dawid 1979) may not be true anymore regarding the "neglect". Indeed, advances in causal inference have been quite impressive over the last two decades – and this paper has discussed only a small share of them. However, as I have tried to show, causal inference continues to demand a clear conception of the mechanisms of a causal model. The subtlety of the problem remains.

Appendix A: Counterfactual conditionals and the POM

Lewis's theory of causation employs possible world semantics for counterfactual conditionals, a semantics which provides truth conditions for counterfactuals in terms of relations between possible worlds (Lewis 1973a,b). Possible world semantics for counterfactuals are based on the idea of comparative similarity between worlds. Given a set of worlds W, one world $W_j \in W$ is closer to a given world $W_i \in W$ than another world $W_k \in W$ if W_j resembles W_i more than W_k resembles W_i. The notion of closeness is based on the idea of W_i being the actual world, and defining $W_j, W_k \in W$ with respect to their proximity to actuality. Lewis imposes two formal constraints on this similarity relation: (i) It produces a weak ordering of worlds such that any two worlds can be ordered with respect to their closeness to the actual world, where "weak" implies that ties are permitted, but any two worlds are comparable. (ii) The actual world is closest to actuality, resembling itself more than any other world does.

For any two propositions C and E, define the following counterfactual conditionals:

(A1a) $C \mathrel{\square\!\!\rightarrow} E$ "If C were (had been) the case, then E would be (have been) the case."

(A1b) $\sim C \mathrel{\square\!\!\rightarrow} \sim E$ "If C were not (had not been) the case, then E would not be (not have been) the case."

The counterfactual conditional $C \mathrel{\square\!\!\rightarrow} C$ is characterized by the following truth condition in terms of the similarity relation:

(A2) $C \mathrel{\square\!\!\rightarrow} E$ is true at a world $W_i \in W$ iff either (i) there are no possible C-worlds, or (ii) some C-world where E holds is closer to W_i than is any C-world where E does not hold.

(i) implies that the counterfactual is vacuously true. From the perspective of W_i being the

actual world, the idea of (ii) is that C $\Box\!\!\rightarrow$ E is (nonvacuously) true in the actual world if it takes less of a departure from actuality to make the antecedent true along with its consequent, than it does to make the antecedent true without the consequent. Under the assumption that there must always be one or more closest C-worlds this condition simplifies to C $\Box\!\!\rightarrow$ E being nonvacuously true iff E holds at all the closest C-worlds.

Lewis (1973b) extends this setting by pairing propositions and events: To any possible event e there corresponds the proposition O(e) that holds at all and only those worlds where e occurs. Thus, O(e) is the proposition that e occurs, and counterfactual dependence among events is simply counterfactual dependence among the corresponding propositions. We then have a definition of causal dependence:

(A3) Let c and e be two distinct possible particular events. Then e causally depends on c iff O(c) $\Box\!\!\rightarrow$ O(e) and ~O(c) $\Box\!\!\rightarrow$ ~O(e).

This condition states that whether e occurs or not depends on whether c occurs or not. The dependence consists in the truth of the two counterfactuals O(c) $\Box\!\!\rightarrow$ O(e) and ~O(c) $\Box\!\!\rightarrow$ ~O(e). Consider two cases: first, if c and e do not actually occur, then the second counterfactual is automatically true because its antecedent and consequent are true. Thus, e depends causally on c iff the first counterfactual holds, i.e., iff e would have occurred if c had occurred. Second, if c and e are actual occurrent events, it follows from the second formal condition on the comparative similarity relation that the first counterfactual is automatically true, because the condition implies that a counterfactual with true antecedent and true consequent is itself true. Thus, e depends causally on c iff, if c had not been, e never had occurred. To put it simply:

(A3a) c causes e iff both c and e are actual occurrent events and if c had not occurred then e would not have occurred.

Or, using the possible world semantics for counterfactuals:

(A3b) c causes e iff both O(c) and O(e) are true in the actual world and in the closest (to

the actual world) possible world in which O(c) is not true, O(e) is not true.

These results can be used to show that the POM is based on causal dependence in counterfactual terms. Define the following set of events:

e_k: Unit $u \in U$ is exposed to treatment $t_k \in T$, i.e. $D(u)=t_k$, and

e^*_k: Unit $u \in U$ has the value $Y_k(u)$ for variable Y,

where k=1,…,m, so that the number of events for each individual unit is $2 \times M$ (as there are N units in U, the total number of events is $2 \times M \times N$). Then:

(A4) The unit-level causal effect of treatment $t_i \in T$ relative to treatment $t_j \in T$ (as measured by Y) is defined by the difference $Y_i(u) - Y_j(u) = UTE_{ij}(u)$ iff the counterfactual conditionals $O(e_i) \mathrel{\square\!\!\rightarrow} O(e^*_i)$, $\sim O(e_i) \mathrel{\square\!\!\rightarrow} \sim O(e^*_i)$, and $O(e_j) \mathrel{\square\!\!\rightarrow} O(e^*_j)$, $\sim O(e_j) \mathrel{\square\!\!\rightarrow} \sim O(e^*_j)$ are true.

To illustrate this, consider the binary case and define the four events:

e_1: Unit u is exposed to treatment t

e_2: Unit u is exposed to treatment c

e^*_1: Unit u has the value $Y_t(u)$ for variable Y

e^*_2: Unit u has the value $Y_c(u)$ for variable Y

In this special case (A4) becomes:

(A4a) The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is defined by the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$ iff the counterfactual conditionals $O(e_1) \mathrel{\square\!\!\rightarrow} O(e^*_1)$, $\sim O(e_1) \mathrel{\square\!\!\rightarrow} \sim O(e^*_1)$, and $O(e_2) \mathrel{\square\!\!\rightarrow} O(e^*_2)$, $\sim O(e_2) \mathrel{\square\!\!\rightarrow} \sim O(e^*_2)$ are true.

If, furthermore, treatment c is merely the absence of treatment, i.e. $e_2 = \sim e_1$, and $e^*_2 = \sim e^*_1$, then UTE is defined under the simplified condition that only the two counterfactuals $O(e_1) \Box\!\!\rightarrow O(e_3)$ and $\sim O(e_1) \Box\!\!\rightarrow \sim O(e_3)$ need to be true, because $O(e_2) \Box\!\!\rightarrow O(e^*_2) = O(\sim e_1) \Box\!\!\rightarrow O(\sim e^*_1) = \sim O(e_1) \Box\!\!\rightarrow \sim O(e^*_1)$, and $\sim O(e_2) \Box\!\!\rightarrow \sim O(e^*_2) = \sim O(\sim e_1) \Box\!\!\rightarrow \sim O(\sim e^*_1) = O(e_1) \Box\!\!\rightarrow O(e^*_1)$.

References

Angrist, Joshua D., and Jinyong Hahn (1999), ”When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects”, NBER Technical Working Paper 241, Cambridge, MA.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), Journal of the American Statistical Association 91, 444-472.

Card, David, and Daniel Sullivan (1988), “Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment”, Econometrica 56, 497-530.

Cochran, W.G. (1965), "The Planning of Observational Studies of Human Populations" (with discussion), Journal of the Royal Statistical Society Series A 128, 234-266.

Cox, David R. (1958), Planning of Experiments, New York: Wiley.

Dawid, A.P. (1979), "Conditional Independence in Statistical Theory", Journal of the Royal Statistical Society Series B 41, 1-31.

Dehejia, R. and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", Journal of the American Statistical Association 94, 1053-1062.

Fisher, Ronald A. (1935), The Design of Experiments, Edinburgh: Oliver & Boyd.

Freedman, David (1999), "From Association to Causation: Some Remarks on the History of Statistics", Statistical Science 14, 243-258.

Galles, David and Judea Pearl (1998), "An Axiomatic Characterization of Causal Counterfactuals", Foundations of Science 3, 151-182.

Gerfin, M. and M. Lechner (2000) ‘Microeconometric Evaluation of the Active Labour Market Policy in Switzerland’, IZA Disc. paper No. 154, IZA: Bonn.

Greenland, Sander (2000), "Causal Analysis in the Health Sciences", Journal of the American Statistical Association 95, 286-289.

Hahn, Jinyong (1998), “"On the Role of the Propensity Score in the Efficient Semi-parametric Estimation of Average Treatment Effects”, Econometrica 66, 315-332.

Heckman, James J. (1992), "Randomization and Social Policy Evaluation", in C. Manski and I. Garfinkel (eds), Evaluating Welfare and Training Programs, Cambridge, MA: Harvard University Press, 201-230.

Heckman, James J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", Quarterly Journal of Economics 115, 45-97.

Heckman, James J., Hidehiko Ishimura and Petra E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", Review of Economic Studies 64, 605-654.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs", in O. Ashenfelter and D. Card (eds), Handbook of Labor Economics, Vol. III, Ch. 31, Amsterdam: North-Holland.

Heckman, James J., and Jeffrey A. Smith (1999), "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies", The Economic Journal 109, 313-348.

Hirano, Kei, Guido W. Imbens and Geert Ridder (2000) "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", NBER technical working paper T0251.

Holland, Paul W. (1986), "Statistics and Causal Inference" (with discussion), Journal of the American Statistical Association 81, 945-970.

Holland, Paul W. (1988), "Causal Inference, Path Analysis, and Recursive Structural Equation Models", Sociological Methodology 18, 449-484.

Imbens, Guido W. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions", Biometrika 87, 706-710.

Kluve, Jochen, Hartmut Lehmann, and Christoph M. Schmidt (1999), "Active Labor Market Policies in Poland: Human Capital Enhancement, Stigmatization, or Benefit Churning?", Journal of Comparative Economics 27, 61-89.

LaLonde, Robert J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", American Economic Review 76, 604-620.

Larsson, Laura (2000), "Evaluation of Swedish youth labour market programmes", Uppsala University, Dept. of Economics Working paper 2000-6.

Leamer, Edward E. (1983), "Let's Take the Con Out of Econometrics", American Economic Review 73, 31-43.

Leamer, Edward E. (1988), "Discussion" [of papers by Holland, Marini and Singer, and Glymour, Scheines and Spyrtes], Sociological Methodology 18, 485-493.

Lechner, Michael (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification", Journal of Business & Economic Statistics 17, 74-90.

Lechner, Michael (2000), "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany", The Journal of Human Resources 35, 347-375.

Lechner, Michael (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption", in M. Lechner and F. Pfeiffer (eds), Econometric Evaluation of Labour Market Policies, Heidelberg: Physica.

Lechner, Michael (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", Review of Economics and Statistics, forthcoming.

Lewis, David (1973a), Counterfactuals, Oxford: Blackwell.

Lewis, David (1973b), "Causation", Journal of Philosophy 70, 556-567.

Neyman, Jerzy (1923 [1990]), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.", translated and edited by D.M. Sabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923), 1-51 (Annals of Agriculture), Statistical Science 5, 465-472.

Neyman, Jerzy (1935), with co-operation by K. Iwaszkiewicz, and S. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation" (with discussion), Supplement to the Journal of the Royal Statistical Society 2, 107-180.

Pratt, John W. and Robert Schlaiffer (1988), "On the Interpretation and Observation of Laws", Journal of Econometrics 39, 23-52.

Quandt, Richard E. (1958), "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes", Journal of the American Statistical Association 53, 873-880.

Quandt, Richard E. (1972), "A New Approach to Estimating Switching Regressions", Journal of the American Statistical Association 67, 306-310.

Robins, James M. and Sander Greenland (2000), "Comment" on 'Causal Inference Without Counterfactuals' by A.P. Dawid, Journal of the American Statistical Association 95, 431-435.

Rosenbaum, Paul R. (1995), Observational Studies, New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika 70, 41-55.

Rosenbaum, Paul R. and Donald B. Rubin (1984), "Reducing Bias in Observational Studies using Subclassification on the Propensity Score", Journal of the American Statistical Association 79, 516-524.

Roy, A.D. (1951), "Some Thoughts on The Distribution of Earnings", Oxford Economic Papers 3, 135-146.

Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", Journal of Educational Psychology 66, 688-701.

Rubin, Donald B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", Journal of Educational Statistics 2, 1-26.

Rubin, Donald B. (1980), "Comment" on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test' by D. Basu, Journal of the American Statistical Association 75, 591-593.

Rubin, Donald B. (1986), "Which Ifs Have Causal Answers", Comment on 'Statistics and Causal Inference' by P.W. Holland, Journal of the American Statistical Association 81, 961-962.

Rubin, Donald B. (1990), "Neyman (1923) and Causal Inference in Experiments and Observational Studies", Comment on 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.' by J. Neyman, Statistical Science 5, 472-480.

Salmon, Wesley (1980), "Probabilistic Causality", Pacific Philosophical Quarterly 61, 50-74.

Skyrms, Brian (1988), "Probability and Causation", Journal of Econometrics 39, 53-68.

Smith, Jeffrey A. and Petra E. Todd (2002), "Does Matching overcome LaLonde's critique of Nonexperimental estimators?" Journal of Econometrics, forthcoming.

Sobel, Michael E. (1995), "Causal Inference in the Social and Behavioral Sciences", in G. Arminger, C.C. Clogg, M.E. Sobel (eds), Handbook of Statistical Modeling for the Social and Behavioral Sciences, New York: Plenum Press.

Speed, T. J. (1990), "Introductory Remarks on Neyman (1923)", Statistical Science 5, 463-464.

Suppes, Patrick (1970), A Probabilistic Theory of Causality, Amsterdam: North Holland.

Table 1. Varieties of Counterfactuals

Number of treatment worlds in W	Treatment of interest	Counterfactual treatment	Causal effect	Interpretation / Notes
M=2	t_1	t_0	$\Delta_{10} = Y_1 - Y_0$	The no-treatment state, in most cases the counterfactual of interest. Usually equals t_1', differs only if explicitly specified (as in experimental studies), or if SUTVA is violated.
		t_1'	$\Delta_{11'} = Y_1 - Y_{1'}$	Anything that is not t_1. Usually applies in observational studies, where it equals t_0.
M>2	t_i	t_0	$\Delta_{i0} = Y_i - Y_0$	The no-treatment state, again the counterfactual of interest in most cases. Relevant as baseline.
		$t_j \neq t_i$	$\Delta_{ij} = Y_i - Y_j$ $= \Delta_{ik} - \Delta_{jk}$ $= \Delta_{i0} - \Delta_{j0}$	Any other particular treatment can be used as counterfactual, for interpretation important to note that the baseline (usually the no-treatment state) is implicit.
		t_i'	$\Delta_{ii'} = Y_i - Y_{i'}$ $= Y_i -$ $F(Y_0, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_{m-1})$	Everything that is not t_i – the absolute counterfactual, the outcome of which is given as a function of the outcomes of all treatments except t_i

Notes: For M>2, as in the discussion in the text, assume that W=Ω.

Table 2. Pairwise treatment effects: FATE

"Treatment"	"Control"			
	Nonparticipation	Training	Wage subsidy	Public work
Nonparticipation	_	.048	.309***	.186***
		(.040)	(.025)	(.061)
Training	-.048	_	.261***	.138**
	(.040)		(.046)	(.072)
Wage subsidy	-.309***	-.261***	_	-.123**
	(.025)	(.046)		(.066)
Public work	-.186***	-.138**	.123**	_
	(.061)	(.072)	(.066)	

Notes: Pairwise treatment effects on post-treatment employment rates. Rows denote the "treatment" t, columns the "control" treatment c, i.e. table entries are $\Delta_{tc}=Y_t-Y_c$. Standard errors in parentheses. Significance levels are denoted ***=1%, **=5%, *=10%.

Table 3. Absolute counterfactuals: FATE

0=Non-participation	1=Training	2=Wage subsidy	3=Public work
.181***	.117***	-.231***	-.067
(.026)	(.045)	(.034)	(.063)

Notes: Absolute counterfactual treatment effects on post-treatment employment rates. Table entries are $\Delta_{ii'}=Y_i-Y_{i'}$. Control treatments weighted with equal weights. Standard errors in parentheses. Significance levels are denoted ***=1%, **=5%, *=10%.

Table 4. Multinomial Logit estimates

	Outcome		
	1=Training	2=Wage subsidy	3=Public work
Female	0.055	-0.715***	-2.669***
	(0.189)	(0.135)	(0.526)
Married	0.328	0.194	0.742**
	(0.216)	(0.155)	(0.323)
Age	0.147*	0.158***	0.325***
	(0.083)	(0.052)	(0.125)
Agesq	-0.002*	-0.002***	-0.004**
	(0.001)	(0.001)	(0.002)
Urban	0.439	0.301	0.691
	(0.373)	(0.269)	(0.732)
Higheduc	1.062***	-1.385	0.075
	(0.389)	(1.021)	(1.042)
Loweduc	-1.071***	0.559***	0.368
	(0.338)	(0.144)	(0.323)
unemp1	0.151	0.553	1.419**
	(0.433)	(0.435)	(0.628)
unemp2	0.105	0.835	-0.745
	(0.598)	(0.526)	(0.982)
unemp3	0.161	-0.725	0.708
	(0.668)	(0.678)	(1.016)
unemp4	1.735***	1.447***	0.873
	(0.292)	(0.348)	(0.670)
unemp34	-0.146	1.637**	0.178
	(0.778)	(0.796)	(1.318)
unemp234	-0.863	-1.572**	0.372
	(0.809)	(0.703)	(1.331)
unemp1234	0.193	1.144**	-0.409
	(0.652)	(0.588)	(0.975)
_cons	-7.623***	-7.937***	-12.513***
	(1.510)	(1.000)	(2.443)
Loglikelihood		-1637.5	

Notes: Base category is outcome 0=Non-participation. N= 6,037. Standard errors in parentheses. Significance levels are denoted ***=1%, **=5%, *=10%. "higheduc"=university, "loweduc"= primary school attainment or less. "unemp1"= person was unemployed in the first of the four quarters preceding treatment. "unemp2", "unemp3", "unemp4" are defined accordingly. "unemp34"= person was unemployed in the third and fourth quarter, etc.

Table 5. Pairwise treatment effects after matching: ATE

"Treatment"	"Control"			
	Nonparticipation	Training	Wage subsidy	Public work
Nonparticipation	_	-.067***	.031***	.111***
		(.008)	(.008)	(.008)
Training	-.018	_	.110**	.092*
	(.041)		(.053)	(.058)
Wage subsidy	-.187***	-.186***	_	.049*
	(.027)	(.035)		(.032)
Public work	-.232***	-.090	.116*	_
	(.067)	(.084)	(.079)	

Notes: Pairwise treatment effects on post-treatment employment rates. Rows denote the "treatment" t, columns the "control" treatment c, i.e. table entries are $\Delta_{tc}=Y_t-Y_c$. Standard errors in parentheses. Significance levels are denoted ***=1%, **=5%, *=10%.

Table 6. Absolute counterfactuals: ATE

	0=Non-participation	1=Training	2=Wage subsidy	3=Public work
Equal weights	.025***	.061*	-.108***	-.069
	(.007)	(.044)	(.027)	(.067)
Probability weights	-.003	-.011	-.184***	-.213***
	(.007)	(.041)	(.027)	(.066)

Notes: Absolute counterfactual treatment effects on post-treatment employment rates. Table entries are $\Delta_{ii'}=Y_i-Y_{i'}$. Standard errors in parentheses. Significance levels are denoted ***=1%, **=5%, *=10%.