Martel García, Fernando; Wantchekon, Leonard

Working Paper

A graphical approximation to generalization: Definitions and diagrams

WIDER Working Paper, No. 2013/082

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Martel García, Fernando; Wantchekon, Leonard (2013) : A graphical approximation to generalization: Definitions and diagrams, WIDER Working Paper, No. 2013/082, ISBN 978-92-9230-659-5, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki

This Version is available at:
https://hdl.handle.net/10419/80913

WIDER Working Paper No. 2013/082

A graphical approximation to generalization

Definitions and diagrams

Fernando Martel García[1] and Leonard Wantchekon[2]

September 2013

Abstract

The fundamental problem of external validity is not to generalize from one experiment, so much as to experimentally test generalizable theories. That is, theories that explain the systematic variation of causal effects across contexts. Here we show how the graphical language of causal diagrams can be used in this endeavour. Specifically we show how generalization is a causal problem, how a causal approach is more robust than a purely predictive one, and how causal diagrams can be adapted to convey partial parametric information about interactions.

Keywords: generalized causal inference, external validity, causal diagrams, policy experiments, non-parametric methods, econometric methodology
JEL classification: B0, C1, C180, C99

[1]independent researcher, fmg229@nyu.edu; [2]Princeton University, lwantche@princeton.edu

This study has been prepared within the UNU-WIDER project 'ReCom–Foreign Aid: Research and Communication', directed by Tony Addison and Finn Tarp.

UNU-WIDER gratefully acknowledges specific programme contributions from the governments of Denmark (Ministry of Foreign Affairs, Danida) and Sweden (Swedish International Development Cooperation Agency—Sida) for ReCom. UNU-WIDER also gratefully acknowledges core financial support to its work programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

ISSN 1798-7237 ISBN 978-92-9230-659-5

Acknowledgements

We would like to thank Rachel Gisselquist, Miguel Niño-Zarazúa, and participants in the UNU-WIDER project workshop on 'Experimental and Non-Experimental Methods in the Study of Government Performance', New York University, 22-23 August 2013, for useful comments and suggestions. Any errors are ours.

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by the authors, and Lorraine Telfer-Taivainen at UNU-WIDER.

1 Introduction

The notion of external validity remains highly ambiguous. In their seminal work Shadish, Cook and Campbell (2002, 38) define external validity as the validity of inferences about whether a causal-effect relationship holds over variation in treatments, outcome measures, units, and settings. By variation they mean variation that is within the bounds observed in the original study, as well as variation outside those bounds (Shadish, Cook and Campbell 2002, 83–84). By *validity* they mean "the approximate truth of an inference", adding that validity "is *not* a property of designs or methods" (their emphasis Shadish, Cook and Campbell 2002, 34).

But how are we to judge the approximate truth of an inference (Shadish, Cook and Campbell 2002, 35)? Is external validity a function of constant effect sizes (Manski 2007, 26), or of constant causal direction (Shadish, Cook and Campbell 2002, 91)? Is external validity a concern only when it involves extrapolation (Manski 2007, 26–28)? Finally, should we interpret external validity as claims about the robustness of particular inferences or the generalizability of a theory (Martel García and Wantchekon 2010)? All social scientists are familiar with the quip that 'experiments lack external validity". Yet on what basis is this claim made? On the basis that a particular study has not been replicated, or on the basis of theoretical insights connecting features of the original study to new populations of interest?

External validity is about theoretical generalization, or the ability to explain and predict outcomes across variations in treatments, outcome measures, units, and settings. In this study we first make the case for a causal approach to external validity. Implicit in this causal notion of generalization is the idea that *all* systematic heterogeneity has a causal explanation. That is, asymptotically, once we remove chance variation, all remaining variation in effect sizes is causal in nature. Consequently generalization is but the process of postulating and inferring the causes behind systematic variation in causal effects.

This study introduces a set of structural definitions to better conceptualize and understand generalization. We illustrate how two classes of causal explanations, effect modulation and effect modification, can in principle explain all causal heterogeneity. We also define causal mechanisms, showing how interaction is a functional form property of such mechanisms. And we show that causal generalization is more robust than predictive generalization, or generalization based on correlations devoid of a causal justification. Our humble goal is simply to introduce practitioners to the use of graphical language for generalizability.

Generalization is of great policy relevance, and is central to the scientific enterprise. Given a budget constraint and significant sunk costs most policy makers

want to make sure policies shown to be successful elsewhere will also be successful at home. This process might involve meetings of experts to discuss the reasons why the policy may or may not work in the new context, paying special attention to the circumstances where the policy proved successful, how these might differ in the present context, why these differences may modify the effect, and if so how. In effect this amounts to a discussion of the various causes of the outcome. A times the context of other policy interventions might be judged to be so different from the target environment that previous policies are almost irrelevant, leading to what Manski (2007, §11) refers to as predictive ambiguity. But how are those judgements made, and what sort of information would be needed to avoid ambiguity. For example, using selection diagrams Pearl and Bareinboim (2011) and Bareinboim and Pearl (2012) have shown how predictive ambiguity can often be avoided by gathering additional information from the target environment using observational studies.

To advance our understanding of external validity and generalization, and in order to avoid ambiguities, this study relies on the structural causal language of Directed Acyclic Graphs (DAGs). The choice of language is predicated on the fact that external validity, as defined here, is essentially a causal question, and DAGs are specially useful for encoding and communicating researchers' private knowledge about causation. Indeed, it is on the basis of public causal knowledge, as encoded in a DAG, that we can begin to provide unambiguous justifications for why, when, and how a cause may have similar effects in different contexts. Absent this knowledge decisions makers face fundamental uncertainty, and it is anybody's guess whether the policy will work or not.

2 Introduction to causal diagrams and models

This section introduces basic definitions to enable unambiguous talk about generalization. The section may appear somewhat dry but it is critical that these terms be understood. Ultimately there can be no scientific progress if we don't know what we are talking about when we are talking about generalization.[1]

Definition 1 (Graph)**.** A graph is a collection $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ of nodes $\mathbf{V} = \{V_1, \ldots, V_N\}$ and edges $\mathbf{E} = \{E_1, \ldots, E_M\}$ where the nodes correspond to variables and the edges denote the relation between pairs of variables.

Definition 2 (Directed Acyclic Graph)**.** A directed acyclic graph (DAG) is a graph that only admits: (i) *directed* edges with one arrowhead (e.g. $\rightarrow$);

[1] In introducing these definitions I follow closely the presentation in Pearl (2009, §1.2), as described in Martel García (2013). Some of these definitions are a direct quotation from Martel García (2013).

(ii) *bi-directed* edges with two arrowheads (e.g. $\leftarrow\!\!\dashrightarrow$); and (iii) no directed cycles (e.g. $X \rightarrow Y \rightarrow X$), thereby ruling out mutual or self causation.

A *path* in a DAG is any unbroken route traced along the edges of a graph – irrespective of how the arrows are pointing (e.g. $X \leftarrow\!\!\dashrightarrow M \rightarrow Y$). A *directed* path, however, is a path composed of directed edges where all edges point in the direction of the path (e.g. $X \rightarrow M \rightarrow Y$ is a directed path between the ordered pair of variables (X, Y)). Any two nodes are *connected* if there exists a path between them, else they are *disconnected*.

Definition 3 (Causal Structure, adapted from Pearl (2009, 44, 203))**.** A *causal structure* or diagram of a set of variables $\mathbf{W}$ is a DAG $\mathcal{G} = \langle \{\mathbf{U}, \mathbf{V}\}, \mathbf{E} \rangle$ with the following properties:

1. Each node in $\{\mathbf{U}, \mathbf{V}\}$ corresponds to one and only one distinct element in $\mathbf{W}$, and vice versa;
2. Each edge $E \in \mathbf{E}$ represents a direct functional relationship among the corresponding pair of variables;
3. The set of nodes $\{\mathbf{U}, \mathbf{V}\}$ is partitioned into two sets:
   (a) $\mathbf{U} = \{U_1, U_2, \ldots, U_N\}$ is a set of *background* variables determined only by factors outside the causal structure;
   (b) $\mathbf{V} = \{V_1, V_2, \ldots, V_N\}$ is a set of *endogenous* variables determined by variables in the causal structure – that is, variables $\mathbf{U} \cup \mathbf{V}$; and
4. None of the variables in U have causes in common.

A causal structure or diagram provides a transparent graphical language for communicating our private knowledge about what variables we believe are relevant for a specific causal analysis, and how these variables stand in causal relation to one another. Figure 1b, adapted from Morgan and Winship (2012, fig. 6), is an example of a causal diagram. In this causal diagram variables U_i are the unobserved background variables, and all other variables are the endogenous variables in $\mathbf{V}$ (i.e. they all have at least one arrow pointing into them).[2] By convention solid nodes represent known and measured variables, whereas empty nodes depict unmeasured ones.

Causal diagram $\mathcal{G}$ represents a possible theory of causation, one where exposure to charter versus public school (C) affects test scores (Y) via feelings of self-worth (S).[3] At the same time parental education (P) and student ability

[2]Background variables are exogenous but not all exogenous variables are background variables, see Pearl (2009, §5.4.3, 7.4.5)

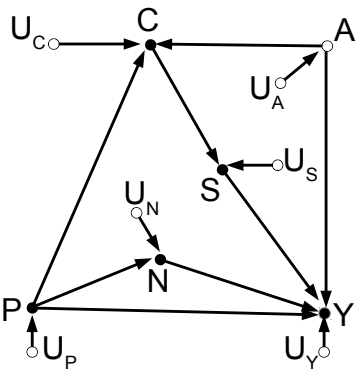
[3]We are not asking that the reader believe this story. The point is simply to have some plausible example of policy relevance.

Figure 1: Panel (a) depicts a non-parametric structural equation model explaining how test scores (Y) depend on student ability (A), feelings of self-worth (S), neighborhood (N), parental education (P), and unobserved background causes (U_Y). Exposure to charter schools (C) is caused by ability, parental education, and unobserved background causes (U_C); and it affects test scores via feelings of self-worth, a mediator. Panel (b) is the equivalent graphical representation of the non-parametric structural model in Panel (a). The causal diagram contains all the information needed for non-parametric causal identification.

(a) Causal Model

$$
\begin{aligned}
y &= f_y(a, s, n, p, u_y);\\
a &= f_a(u_a);\\
s &= f_s(c, u_c);\\
n &= f_n(p, u_n);\\
p &= f_p(u_p);\\
c &= f_c(a, p, u_c);\\
u_y &\perp u_c \perp u_n \perp u_a \perp u_p \perp u_s.
\end{aligned}
$$

(b) Causal Diagram ($\mathcal{G}$)



(A) (unobserved, notice hollow circle) are both common causes of exposure to charter schools, and of test scores. These two causes act as potential confounders. They both imply an association between charter schools and test scores even if charter schools are without effect (e.g. even if we delete the arrows in $C \to S$, or $S \to Y$ from $\mathcal{G}$). Parental education (P) affects tests scores directly, by helping with homework say, and indirectly, via the choice of residential neighbourhood (N) and school type (C).

Causal diagrams invite the use of an intuitive terminology to refer to causal relations. In a causal diagram $C \to S$ reads "C causes S". We also say that C is a *parent* of S, and S is a *child* of C, if C directly causes S, as in $C \to S$. For example, the *parents* of Y are denoted $\text{PA}(Y) = \{P, N, S, A\}$.[4] Similarly, we say that C is an *ancestor* of Y, and Y a *descendant* of C, if C is a direct or indirect cause of Y. Thus, P is both a direct cause of Y, as in $C \to Y$, and an indirect cause, as in $C \to N \to Y$. We refer to non-terminal nodes in directed paths as *mediators*. S is a mediator in the path $X \to S \to Y$.

In addition to laying out causal theories graphically, and with intuitive termi-

[4]By convention we confine the set of parents of Y to variables in $\mathbf{V}$. Hence we do not include U_Y in the set $\text{PA}(Y)$ even though U_Y is a direct cause of Y. One can think of such background variables as unobserved disturbances.

nology, causal diagrams have two additional properties. First, by Definition 3 a DAG of a set of variables $\mathbf{W}$ only qualifies as a causal diagram if it includes all common causes of the variables in $\mathbf{W}$ (see point 4 in the definition).[5] This ensures the diagram has some nice properties, including the ability to read conditional independencies directly.[6] For example, the diagram tells us that under the null of no effect (e.g deleting $C \rightarrow S$ in causal diagram 1b), and conditional on P and A, charter schools and test scores are distributed independent of each other. If C and Y remain associated despite controlling for these variables, then we read that as evidence that they are causally related under the assumptions laid out in causal diagram 1b.[7]

Second, the definition of causal diagrams relies on directed edges (e.g. arrows) in place of explicit functional relations to depict causal relations between variables in the graph. This is a feature not a bug. Detailed knowledge about specific functional forms is often completely unnecessary for causal identification. To wit, this diagrammatic representation of functional relations is in accordance with how most people store their causal knowledge. For example, most of us know that smoking causes lung cancer but few, if any of us, know the precise functional relation linking them together.

Figure 1 also shows that every causal model has a corresponding causal diagram (Figures 1a and 1b respectively). A causal model is defined as follows:

Definition 4 (Causal Model, adapted from Pearl (2009, 203))**.** A *causal model* $\mathbf{M}$ replaces the set of edges E in a causal structure $\mathcal{G}$ by a set of functions $\mathbf{F} = \{f_1, f_2, \ldots, f_N\}$, one for each element of $\mathbf{V}$, such that $\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$. In turn, each function f_i is a mapping from (the respective domains of) $U_i \cup \mathrm{PA}_i$ to V_i, where $U_i \subseteq \mathbf{U}$ and $\mathrm{PA}_i \subseteq \mathbf{V} \backslash V_i$ and the entire set $\mathbf{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. In other words, each f_i in

$$v_i = f_i(pa_i, u_i), \quad i = 1, 2, \ldots, N,$$

assigns a value to V_i that depends on (the values of) a select set of variables in $V \cup U$, and the entire set $\mathbf{F}$ has a unique solution $\mathbf{V}(\mathbf{u})$.

Like the causal diagram, a causal model is completely non-parametric. For example, casual model 1a specifies that being exposed to a charter schools is

[5] If two variables in $\mathbf{W}$ have a cause Z in common (e.g. $U_Y \leftarrow Z \rightarrow U_A$) but $Z \notin \mathbf{W}$, then DAG $\mathcal{G}$ is not a causal diagram. To make it one Z should be included in $\mathbf{W}$ and U_Y and U_A included in the set of endogenous variables $\mathbf{V}$.

[6] Formally a causal diagram meets the Causal Markov Condition, see Pearl (2009, 19, 30) for details.

[7] Causal diagrams are specially useful for determining the conditions under which a desired quantity of interest is identified. See Morgan and Winship (2007, §1.6) for a gentle introduction, and Martel García (2013) for a recent application to identification of causal effects in experiments subject to attrition.

a function $c = f_c(a, p, u_c)$. This function is compatible with any well defined mathematical expression in its arguments like $c = \alpha + \beta_1 a + \beta_2 p + u_c$, or $c = \alpha + \beta_1 a + \beta_2 p + \beta_3 a \times p + u_c$.

Causal models, like causal diagrams, are completely deterministic: Probability comes into the picture through our ignorance of background conditions $\mathbf{U}$, which we summarize using a probability distribution $\mathrm{P}(\mathbf{u})$. In turn, $\mathrm{P}(\mathbf{u})$ induces a probability distributions $\mathrm{P}(\mathbf{v})$ over all endogenous variables in $\mathbf{V}$.[8]

Definition 5 (Probabilistic Causal Model, Pearl (2009, 205))**.** A probabilistic causal model Γ is a pair $\langle \mathbf{M}, \mathrm{P}(\mathbf{u}) \rangle$, where $\mathbf{M}$ is a causal model and $\mathrm{P}(\mathbf{u})$ is a probability function defined over the domain of $\mathbf{U}$.

Finally, social scientists often talk about generalizability in terms of causal mechanisms. But what are causal mechanisms, and what is the difference between a model and a causal mechanism, if any? The present framework allows us to define such mechanisms precisely:

Definition 6 (Causal Mechanism)**.** A *causal mechanism* is any $\mathbf{F}' \subseteq \mathbf{F}$, where $\mathbf{F}$ is the set of functions in causal model $\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$.

For example, in Figure 1a function f_y is a causal mechanism generating y, and so too is the set $\mathbf{F}$ of all mechanisms in model $\mathbf{M}$. The difference between these two mechanisms is that f_y takes some endogenous variables as inputs, whereas mechanism $\mathbf{F}$ takes only background variables as inputs. For instance, in causal model 1a we say that ability (A) causes test scores (Y) via mechanism $\mathbf{F}_{A,Y} = \{f_c, f_s, f_y\}$.

After this brief introduction to causal diagrams we turn to the formal definitions needed to understand interventions, heterogeneity, and generalization.

3 Intervention, causal heterogeneity, and generalization

In this section we investigate effect heterogeneity ignoring causal identification issues. We start by laying out the notion of an intervention, then we examine the nature and causes of causal heterogeneity. Throughout we assume a perfectly randomized controlled experiment in one setting, and then consider why and how the exact same intervention may have different results in different settings.

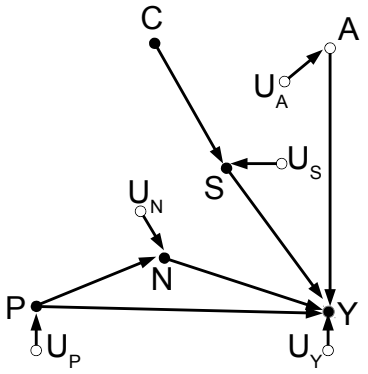
[8] This is exactly the same as characterizing disturbance terms in a regression context using some distribution, like $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Figure 2: Causal diagram and model representing an intervention whereby a researcher is completely able to set variable C to any desired value, like c'. Graphically this kind of control is represented by deleting all arrows pointing into C, which captures the idea that nothing causes C other than the researcher's intervention. In terms of the causal model the intervention 'wipes out' f_c by forcing the value of c'.

(a) Causal Model

$$
\begin{aligned}
y &= f_y(a, s, n, p, u_y);\\
a &= f_a(u_a);\\
s &= f_s(c, u_s);\\
n &= f_n(p, u_n);\\
p &= f_p(u_p);\\
\mathbf{c} &= \mathbf{c}';\\
u_y &\perp u_n \perp u_a \perp u_p \perp u_s.
\end{aligned}
$$

(b) Causal Diagram



3.1 Intervention

To continue with the charter schools example suppose causal diagram 1b in Figure 1 is a faithful representation of all that we know at time t about the effect of charter schools (C, versus public schools) on test scores (Y), and of possible confounders of this effect like parental education P and student ability A. Testing and estimating the effect of C on Y with observational data is complicated by the fact that student ability is both a confounder and unobserved, so we cannot control for it. Consequently we decide to carry out a randomized controlled trial on a convenience sample $\mathcal{S}$ from the population of interest $\mathcal{P}$. In particular, an equal number of students in $\mathcal{S}$ are randomly assigned to public schools and the rest to charter schools.

Figure 2 is the intervention equivalent of Figure 1 assuming c is under the complete conrol of the researcher. We call this an intervention $do(c)$ and what it does is replace the second equation in causal model 1a with $c = c'$, generating the new intervention causal model 2a in Figure 2. In effect experimental intervention deletes all arrows pointing into C, thereby eliminating any possibility of confounding. If the randomized controlled experiment is well implemented any endline association between c and y among the set of expermental subjects $\mathcal{S}$ cannot be due to some unobserved cause in common, or confounder, but to a causal effect of c on y.[9]

[9]There are other ways to represent experimental interventions in the context of a causal model (see Pearl (2009, §3)). One possibility is to replace $f_c(a, p, u_c)$ with $f'_c(a, p, z, u_z)$

Experimental outcomes are uncertain. The experimenter sets the value of C but Nature sets the value of all other background variables $\mathbf{U}$. In effect, an experiment is a probabilistic causal model (see Definition 5) $\Gamma = \langle \mathbf{M}, \mathrm{P}_{\mathcal{S}}(\mathbf{u}, c) \rangle$, where $\mathbf{M}$ is intervention causal model 2a in Figure 2, and $\mathrm{P}_{\mathcal{S}}(\mathbf{u}, c)$ is the joint distribution of background variables $\mathbf{u}$ and intervention variable c in sample $\mathcal{S}$. By randomization $\mathrm{P}_{\mathcal{S}}(\mathbf{u}, c) = \mathrm{P}_{\mathcal{S}}(\mathbf{u})\,\mathrm{P}_{\mathcal{S}}(c)$. Nature then solves this probabilistic model and yields the intervention distribution $\mathrm{P}_{\mathcal{S}}(y|do(c))$ defined over each randomized level of $c \in \{c', c''\}$.[10] This distribution describes the outcome data from the experiment, and it can be queried to calculate any quantity of interest $\mathcal{Q}(\Gamma)$. For example, for any two distinct values c' and c'' of c, the average treatment effect is defined as:

$$\mathcal{Q}(\Gamma) = E[y|do(c')] - E[y|do(c'')].$$

Suppose $\mathcal{Q}(\Gamma)$ is statistically and substantively significant. How can we use this information to predict $\mathcal{Q}(\Gamma)$ in a different sample $\mathcal{S}^*$ from the same population $\mathcal{P}$ (e.g. $\mathcal{S}^* \subseteq \mathcal{P}$)?[11] Moreover, what factors can give rise to systematic difference across samples? And what might be the causes of heterogeneity in the causal model?

3.2 Heterogeneity

We begin this section with some intuition, and then follow with some formal definitions needed for the analysis of heterogeneity.

3.2.1 Intuition

Structural causal models are completely non-parametric and potentially heterogeneous. To begin with, consider the sub-model $C \to S \leftarrow U_S$ of causal diagram 2b in Figure 2. Because C is under the direct control of the researchers, all variation in S across different samples will come from the background variable U_S. This can be problematic for two reasons. First, variables C and U_S may happen to interact in mechanism $f_s(c, u_s)$ (we will define interaction below), in which case $\mathcal{Q}(\Gamma)$ may be sensitive to changes in the distribution of the background variables $\mathrm{P}(\mathbf{u})$. If so we say U_y is a *moderator* of the effect of S on Y. Second, such changes in the distribution of background variables are likely to happen. The original experimental sample $\mathcal{S}$ is a convenience sample

and $z = z'$, which captures the notion that the researcher only has access to an imperfect instrument z for controlling C (imperfect because Nature still has some say in generating c).

[10] The researcher cannot solve the model because she never observes $\mathrm{P}(\mathbf{u})$.

[11] We focus on the external validity of quantities of interest as this is a less demanding task than predicting $\mathrm{P}(y^*|do(x))$. The latter requires knowledge of all the causes of Y

from population $\mathcal{P}$, and so not representative of all background conditions in the population. Consequently, $P_{\mathcal{S}^*}(c, u_s)$ in a new sample $\mathcal{S}^*$ is very likely to differ from $P_{\mathcal{S}}(c, u_s)$, even if $P_{\mathcal{S}}(c) \equiv P_{\mathcal{S}^*}(c)$.[12] In sum, if $f_s(c, u_s)$ involves some interaction, *and* if $P_{\mathcal{S}^*}(c, u_s) \neq P_{\mathcal{S}}(c, u_s)$, then changes in background conditions are likely to bring about changes in $\mathcal{Q}(\Gamma)$.

Second, consider the full mechanisms by which C is theorized to exert its causal influence on Y as described by causal diagram 2b in Figure 2. As in the previous example, heterogeneity can arise if U_S and C interact in mechanism f_s, and $\mathrm{P}_{\mathcal{S}}(\mathbf{u}, c) \neq \mathrm{P}_{\mathcal{S}^*}(\mathbf{u}, c)$ in new sample $\mathcal{S}^*$. Heterogeneity can also arise if variable S interacts with any other argument of f_y, including variables A, N, P, U_Y. These variables can all – singly or jointly – moderate the effect of S on Y. Importantly, variables A, N, P are all endogenous, that is determined, at least in part, by $\mathrm{P}_{\mathcal{S}}(\mathbf{u})$. This is another reason why background conditions matter. Conditioning on observable variables is a way to account for the influence of unobservable background conditions.

In addition to moderators, variable U_s can also act as a *modulator* of the effect of C on Y. Modulator because it can regulate the effect of C on Y through its moderator effect on mediator S (assuming it has such a moderator effect). The focus on the total effect of C on Y allows us to introduce meaningful new labels, like modulator, which goes to show how the conceptualization of heterogeneous effects arises naturally from the causal structure.

3.2.2 Formal definitions

Definition 7. (Causal Effect Structure) A *causal effect structure* for the effect of a set of variables $\mathbf{X}$ on a set of variables $\mathbf{Y}$ in causal model $\mathbf{M}$, is a set of variables $\mathbf{E}_{X,Y}$ such that it only includes $\mathbf{X}$, and all descendants of $\mathbf{X}$ along all directed paths from variables in $\mathbf{X}$ to variables in $\mathbf{Y}$

For example, the causal effect structure for the effect of C on Y according to causal model 2a is the set $\mathbf{E}_{C,Y} = \{C, S\}$. Conventionally such a set of causes and mediators is what researchers have in mind when they think of "mechanisms", but this is at odds with how we defined mechanisms in Definition 6. Besides, it is easy to see that knowing this "mechanism" is not enough to guarantee replication out of sample. In particular, the faithfulness of the replication may also depend on other causes of Y or S, not in $\mathbf{E}_{C,Y}$, that may interact with the causal effect structure, like variable U_S and all parents of Y other than S in causal diagram 2.

[12] Even if the original experiment had been carried out on the full population $\mathcal{P}$, $P_{\mathcal{S}^*}(c, u_s)$ will likely differ due to the randomized nature of c.

Definition 8. (Direct Causal Context) A *direct causal context* for the effect of one set of variables $\mathbf{X}$ on another set of variables $\mathbf{Y}$ in causal model $\mathbf{M}$ is a a set of variables $\mathbf{C}_{X,Y}$ such that:

1. it excludes the casual effect structure $\mathbf{E}_{X,Y}$;
2. it includes all remaining parents of Y; and
3. it includes all parents of all mediator variables in $\mathbf{E}_{X,Y}$.

For example, in causal diagram 2b the direct causal context for the effect of C on Y is $\mathbf{C}_{C,Y} = \{A, N, P, U_S, U_Y\}$. Conditioning on this set of variables *guarantees* replication in any other setting, without committing ourselves to any functional form assumptions about interactions. That is, these variables may, or may not, interact with other variables in the causal effect structure but so long as they are conditioned on, faithful replication is guaranteed. Obviously this conditioning strategy fails if some of these variables are unobserved and have moderator effects. The second instance where conditioning strategies fail is when we are asked to replicate in settings that fall outside the original range of observation.

Definition 9. (Probabilistic Direct Causal Context) A *probabilistic causal context* for the effect of one set of variables $\mathbf{X}$ on another set of variables $\mathbf{Y}$ in probabilistic causal model Γ is a distribution $\mathrm{P}(\mathbf{C}_{X,Y})$, defined over a direct causal context $\mathbf{C}_{X,Y}$.

Suppose the direct causal context $\mathbf{C}_{C,Y}$ in causal diagram 2b is fully observed in sample $\mathcal{S}$ as $\mathrm{P}(a, n, p, u_s, u_y)$.[13] Now suppose we draw another sample $\mathcal{S}^* \subseteq \mathcal{P}$, and we observe values $a^*, n^*, p^*, u_s^*, u_y^*$ s.t. $\mathrm{P}^*(a^*, n^*, p^*, u_s^*, u_y^*) > 0$ but $\mathrm{P}(a^*, n^*, p^*, u_s^*, u_y^*) = 0$. In this case the conditioning needs of the target environment go beyond the conditions available in the source environment (e.g. they are outside the support of $\mathrm{P}(a, n, p, u_s, u_y)$). Predicting quantities of interest for instances where $\mathrm{P}(a^*, n^*, p^*, u_s^*, u_y^*) = 0$ will require extrapolation or interpolation, namely making some functional form assumptions about all mechanisms that take elements of $\mathbf{C}_{C,Y}$ as inputs.[14]

Definition 10. (Interaction (adapted from VanderWeele (2009, 864))) For a given probabilistic causal model Γ, there is said to be an *interaction* between two or more parents of an effect Y, call them set X and set Z, if the quantity

[13] E.g. suspend belief and assume we can observe a, u_s, u_y.

[14] Discussing the relevant methods of extrapolation or interpolation is well beyond the scope of this study. The main criterion is that they give good predictions.

of interest computed from Y, $\mathcal{Q}(\Gamma)$, is such that:

$$\begin{aligned}&\mathcal{Q}\left[\mathrm{P}\left(Y|do(x'),do(z')\right),\mathrm{P}\left(Y|do(x''),do(z')\right)\right]\\&\neq\\&\mathcal{Q}\left[\mathrm{P}(Y|do(x'),do(z'')),\mathrm{P}(Y|do(x''),do(z''))\right],\end{aligned}$$

for some distinct (possibly vector valued) observations x' and x'' of X, and z' and z'' of Z.

Interaction is a functional form property of mechanisms. By the definition of a mathematical function we do not need to know the function itself, only its arguments and the values they take, in order to be able to accurately predict quantities of interest across settings using previous realizations. For example, if we are only interested in studying how the effect of a cause X on an effect Y varies across contexts, then we only need to know the arguments to the derivative of f_y^* with respect to X, where f_y^* is the reduced form mechanism for the effect of X on Y. This mechanism is at most a function of variables in causal context $\mathbf{C}_{X,Y}$ that interact with variables in causal effect structure $\mathbf{E}_{X,Y}$. That is, the variables needed to fully explain the variation of the effect out of sample is a set $\mathbf{H} \subseteq \{\mathbf{C}_{X,Y}, \mathbf{E}_{X,Y}\}$

Because interactions are a property of the set of mechanisms $\mathbf{F}$ in causal model $\mathbf{M}$, model transformations can be used to limit interactions.[15] But here we must think of transformations as functional form transformations of mechanisms $\mathbf{F}$ in model $\mathbf{M}$, and not as simple variable transformations.

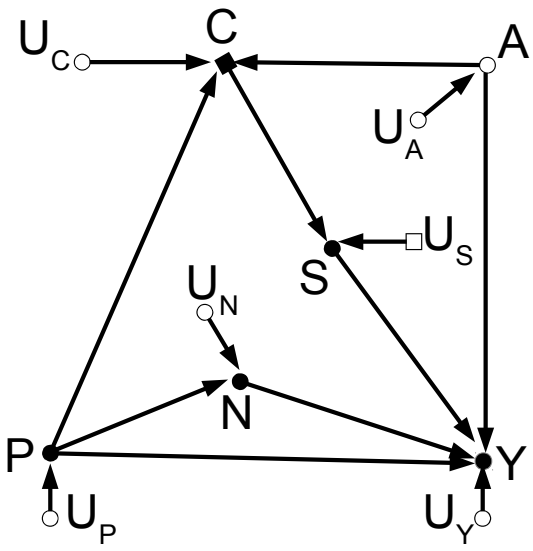
Definition 11. (Direct Causal Context Interaction) In considering the effect of a set of variables X on another set of variables Y in model $\mathbf{M}$, we say there is a *direct causal context interaction* of the effect of X on Y according to quantity of interest $\mathcal{Q}(\Gamma)$ whenever any subset $\mathbf{E}_I$ of causal effect structure $\mathbf{E}_{X,Y}$, interacts with any subset $\mathbf{C}_I$ of causal context $\mathbf{C}_{X,Y}$. We refer to the set of interacting sets as $\mathbf{I}_{X,Y} = \{\mathbf{E}_I \cup \mathbf{C}_I\}$. If there are no causal context interactions $\mathbf{I}_{X,Y} = \emptyset$.

Being completely non-parametric causal diagrams do not convey any functional form information. One possibility is to expand the notation to convey the location of interacting variables. For example, suppose exposure to charter schools interacts with background conditions U_S. We might label the variables $\mathbf{I}_{C,S} = \{C, U_S\}$ explicitly in the causal diagram using edges with square ($\square$) origins, as shown in Figure 3, where filled squares refer to observed variables (C), and unfilled ones to unobserved variables (U_S). Graphically, the process of generalization, or of explaining away heterogeneity, requires abstracting and

[15] At times this has consequences for prediction out of sample on the original scale (Kennedy 1983).

measuring from background U_S the observable variables generating the heterogeneity thus replacing the empty square with an empty circle in U_Y. We call such (semi-parametric) diagrams *interaction causal diagrams*.[16]

Figure 3: Interaction causal diagram. Direct causal context interactions are denoted with an empty square if unobserved (U_S), or a full square if observed (U_C), where we assume $\mathbf{I}_{C,S} = \{C, U_S\}$ (see Definition 11).



Definition 12. (Robustness) The effect of a set of variables $\mathbf{X}$ on a set of variables $\mathbf{Y}$ according to quantity of interest $\mathcal{Q}(\Gamma)$ is said to be *robust* if causal model $\mathbf{M}$ admits no causal context interaction for this effect ($\mathbf{I}_{X,Y} = \emptyset$).

Robustness is a strong but powerful property of some causal models. One that allows the researcher to completely ignore the causal context in predicting a given quantity of interest out of sample. The graphical equivalent is an interaction causal diagram without any square nodes.

3.3 Generalization

The process of *generalization* involves explaining away causal heterogeneity. Suppose we started with causal diagram 2b in Figure 2, and that repeated experimentation across samples from population $\mathcal{P}$ show significant variation in the effect of charter schools (C) on self-regard (S), and hence on test scores

[16] Pearl and Bareinboim (2011) use a similar notation – though not focused only on interactions –, which they call *selection* diagrams.

(Y). Suppose we observe much less variation in this effect within levels of the residential neighbourhood variable (N), than across levels of it. Could it be that N is a cause of S, that we should replace $f_s(c, u_s)$ with $f'_s(c, n, u_s)$, and that, conditional on N, U_S no longer modifies the effect of C? It might be that feelings of self-worth are relative to a students neighbourhood, as in feeling privileged to be in a charter school within a poverty ridden neighbourhood. We could carry out a two-way randomization of students to neighbourhoods and schools to test this hypothesis. We might find that the evidence is indeed consistent with mechanism $f'_s(c, n, u_s)$, and that, conditional on N, there is no evidence U_S interacts with C (or N).

That neighbourhood causes feelings of self-worth is one possibility. Another possibility is that N is an effect of U_S, or, more likely perhaps, that they share an unobserved cause in common (Z); in which case N serves as a *proxy* for their cause in common. More generally, the knowledge that N correlates with the quantity of interest might seem sufficient to condition and predict effects out of sample. We refer to this as the prediction or robustness approach to generalization. By contrast, generalization offers a theory driven analytical approach to validity (Martel García and Wantchekon 2010).

Generalization differs from pure prediction in two crucial aspects. First, it provides theoretically motivated explanations for the causes of heterogeneity. In effect, the process of generalization involves observation, theorizing, abstracting potential moderators from within the set of background variables, and including them explicit in the model as endogenous variables. The second difference between generalization and the predictive approach is that the former is, at least in principle, more robust than the former. Of course, both causal models and purely predictive ones can be proved wrong by the data. The question is how much more fallible are they. Intuitively, causal explanations are more direct and so more robust. In the previous example, if N is shown to cause S, then heterogeneity of the effect of C on S can still arise if there are more variables amongst the background conditions that interact with C. However, if N is only a proxy for some hidden cause Z in common with U_S, then heterogeneity in the conditional effect can arise at multiple points. For example, due to interactions between U_N and Z, or Z and U_S, in addition to between U_S and C. This is three times more opportunities for failure compared to the direct causal explanation.[17]

Finally, if for some reason we are only interested in predicting the effect of C on S out of sample, then we can ignore most other variables in causal diagram $\mathcal{G}$. That is, the relevant causal context is specific to the causal relation under study. In this instance, pruning $\mathcal{G}$ can help focus our attention on

[17]The robustness of the causal interaction approach stems from the Causal Markov Condition. At the same time establishing causality is more involved and expensive, so there is a tradeoff between robustness and convenience.

possible moderators within the background variables in U_S. Causal diagrams make explicit the relevant causal context to be considered for predicting out of sample.

4 Conclusion

Few scientists begin an experimental investigation by laying out their best guess about the structure of the causal effect, the causal context, and the likely sources of heterogeneity. With the advent of causal diagrams there is little excuse for this practice, as anyone can draw arrows, circles, and squares. In the interest of generalization we would encourage practitioners to lay out threats to external validity explicitly at the outset of the study design in an interaction causal diagram . This way they can plan in advance what sorts of measurements should be taken for generalization, highlight potential threats to generalization, and suggests what measurements might be taken to predict the effect of intervening in a different context.

In some instances theories or educated guesses might not be available but there might be plenty of data on covariates. In these situations it is natural to search the covariate space for evidence of interactions. This can generate new hypotheses to be tested out of sample, including testing whether these covariates are part of the causal context of effect structure, or only proxies for such variables. We would want to test this because, as already noted, causal knowledge is more robust than knowledge about correlations. Also, the approach we have taken thus far relies mainly on non-parametric stratification, though there is much to be said about using hierarchical models for summarizing the inference, especially when there are numerous strata, or they are thinly populated.

Generalization is key to science yet its meaning remains highly ambiguous. Most extant theories have defined generalization in an *ex post* fashion, emphasizing whether a particular inference holds out of sample. Such a robustness approach obviates the need for theory driven research, emphasizing instead replications across all imaginable contexts. Building on the analytical approach of Martel García and Wantchekon (2010), and the more recent structural approach of Pearl and Bareinboim (2011), this study argues for a theory driven approach. Specifically, interaction causal diagrams can be used to encode *ex ante* potential sources of heterogeneity on the basis of existing knowledge and theories; to guide the design of experiments, follow-up experiments and measurements that might be needed to further justify external validity claims; and to communicate simply, clearly, and transparently to the broadest audience possible what the researchers know about the sources of causal heterogeneity. Science is a communal endeavor that ought to begin with clear

definitions and accessible language.

References

Bareinboim, Elias and Judea Pearl. 2012. Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ed. J. Hoffmann. and B. Selman. pp. 698–704.

Kennedy, Peter. 1983. "Logarithmic Dependent Variables and Prediction Bias." *Oxford Bulletin of Economics and Statistics* 45(4):389–92.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Harvard University Press.

Martel García, Fernando. 2013. Definition and Diagnosis of Problematic Attrition in Randomized Controlled Experiments. Working Paper 2302735. Available at SSRN: http://ssrn.com/abstract=2302735.

Martel García, Fernando and Leonard Wantchekon. 2010. "Theory, External Validity, and Experimental Inference: Some Conjectures." *The ANNALS of the American Academy of Political and Social Science* 628(1):132–147.

Morgan, Stephen L. and Christopher Winship. 2007. *Couterfactuals and Causal Inference: Methods and principles of Social Research*. Cambridge Univ. Press.

Morgan, Stephen L. and Christopher Winship. 2012. Bringing Context and Variability Back in to Causal Analysis. In *The Oxford Handbook of Philosophy of Social Science*, ed. Harold Kincaid. Oxford Handbooks in Philosophy Oxford University Press chapter 14, pp. 319–354.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, Judea and Elias Bareinboim. 2011. Transportability across studies: A formal approach. Technical report UCLA.

Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. Houghton Mifflin Company.