External Validity: Is There Still a Problem?

Alexandre Marcellesi*†

I first propose to distinguish between two kinds of external validity inferences, predictive and explanatory. I then argue that we have a satisfactory answer to the question of the conditions under which predictive external validity inferences are good. If this claim is correct, then it has two immediate consequences: First, some external validity inferences are deductive, contrary to what is commonly assumed. Second, Steel's requirement that an account of external validity inference break what he calls the 'Extrapolator's Circle' is misplaced, at least when it comes to predictive external validity inferences.

1. Introduction. It is customary, at least since Campbell (1957), to distinguish between two properties of the conclusions drawn from studies purporting to estimate causal relationships: internal validity and external validity. This distinction is typically drawn as follows (see, e.g., Cook and Campbell 1979, 37): a causal conclusion drawn from a study is internally valid if and only if it is true of the population on which the study was conducted. And it is externally valid if and only if it is true not just of the study population but also of some other population(s) as well. Although there is fairly wide agreement on the strictures a study must satisfy in order for the causal conclusions one draws from it to be internally valid, matters tend to be muddier when it comes to external validity. There is no widespread agreement on the conditions under which one can infer from the truth of a causal claim in a study population to its truth in some other population.

In this article I propose, first, to distinguish two kinds of external validity inferences, predictive and explanatory. I will then argue that we have the correct answer—developed independently by Nancy Cartwright and Jeremy

*To contact the author, please write to: Department of Philosophy, University of California, San Diego, 9500 Gilman Drive 0119, La Jolla, CA 92093-0119; e-mail: amarcellesi@ucsd.edu.

†I wish to thank Nancy Cartwright for her comments on an earlier draft as well as audiences at the 2014 PSA conference in Chicago and at the UCSD Graduate Philosophy Colloquium.

Philosophy of Science, 82 (December 2015) pp. 1308–1317. 0031-8248/2015/8205-0048$10.00
Copyright 2015 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/684084 Published online by Cambridge University Press

Hardie on the one hand and Elias Bareinboim and Judea Pearl on the other—to the question of the conditions under which predictive external validity inferences are good inferences. If this claim is correct, then it has two immediate consequences: First, Daniel Steel's demand that any acceptable account of external validity inferences break what he calls the 'Extrapolator's Circle' is misplaced. Second, some external validity inferences are deductive, in contradiction with the widely accepted view according to which they are inductive.

I will say little about external validity inferences of the explanatory kind. My hope in this article is to show that, as far as predictive external validity inferences are concerned, the problem has been solved.

2. A Classification of External Validity Inferences. Issues of external validity (EV) typically arise in one of the following two contexts. They arise, first, in evidence-based policy (EBP), where EV inferences are often drawn from the results of field experiments (e.g., randomized controlled trials [RCTs]) aiming to estimate the effects of particular policies. Second, issues of EV arise when interpreting the results of laboratory experiments, especially in the cognitive and social sciences, from cognitive psychology to experimental economics. Let me give an example of what EV inferences look like in each of these two contexts, starting with the case of EBP.

Consider the following schematic case, which I will use for illustration throughout the article. Imagine that Mary is the superintendent of some school district d_1, a district that includes n high schools $S_1, \ldots, S_n$. Imagine also that Mary is dissatisfied with the scores achieved by her high school students on the SAT (Scholastic Assessment Test) and that she would like to take measures to increase these scores. Call Y_i the discrete variable representing the mean SAT score achieved by students from high school S_i. Because Mary is in charge of the entire district, and not of any specific school, the quantity she seeks to increase is the mean district-level SAT score $\mathrm{E}[Y_i]$. A colleague of Mary's who is the superintendent of some neighboring school district d_0 suggests to her that she implement a particular educational policy P, a policy that has been evidenced—say, by a well-conducted RCT—to increase the mean SAT score in d_0. Call X the binary variable representing whether policy P has been implemented ($X = 1$ when it is, 0 otherwise), and call m (where $m > 0$) the estimated size of the effect of $X = 1$ on $\mathrm{E}[Y_i]$ in d_0.

Should Mary go ahead and implement policy P in d_1 on the basis of the evidence that P caused an increase in SAT scores in d_0? Only if the causal claim that implementing P leads to an increase of magnitude m in SAT scores is externally valid, that is, only if it is true not just in d_0 but also in d_1. The question of EV thus is central to EBP: any policy decision that is based on evidence gathered in locations (whether spatial or temporal) other than the one targeted by the proposed policy must, it seems, involve some EV

https://doi.org/10.1086/684084 Published online by Cambridge University Press

inference. And although the case described above is fictional, it is representative of the inferential problem typically faced by practitioners of EBP.

The second context in which issues of EV typically arise is in discussions of results from laboratory experiments. Consider the following classic example from experimental economics. It is often the case that, in order to conduct cost-benefit analyses of environmental policies, economists and policy makers must first assess the monetary value of various nonmarket goods (e.g., clean air). One way to do so is to assess how much members of the population targeted by the policy in question are willing to pay for the good, assuming that they do not already possess it. Another way is to assess how much they are willing to accept as a payment to part with the good in question, assuming they possess it. According to standard preference theory, 'willingness to pay' (WTP) and 'willingness to accept' (WTA) are two measures of the same quantity, namely, of the monetary value the members of some population attach to a nonmarket good.

There is ample experimental evidence, however, that WTP and WTA diverge for identical goods (see, e.g., Starmer 2000): people tend to value the same good differently depending on whether they are asked how much they are willing to pay to acquire it or how much they are willing to accept as a payment to part with it. And they tend to value goods more highly when they possess them than when they do not; that is, the WTA for a given good tends to be higher than its WTP. This is what has come to be known as the 'endowment effect'. One prominent explanation for the endowment effect is loss aversion (see, e.g., Kahneman, Knetsch, and Thaler 1990). According to models incorporating loss aversion, agents are more sensitive to losses than they are to gains, and this is why the amount they are willing to pay to acquire a good they do not possess is lower than what they are willing to accept as a payment to part with the same good. And there is experimental evidence to the effect that loss aversion is in fact a cause of the endowment effect (see, e.g., Bateman et al. 1997).

There are at least two ways in which issues of EV arise in this classical case from experimental economics. First, one might wonder whether loss aversion, given experimental evidence to the effect that it is widespread, can help explain past instances in which the endowment effect has been observed, either inside the lab or outside. Second, one might wonder whether, given that one is trying to determine the monetary value of a nonmarket good for a population the members of which are known to be loss averse (with respect to this good), one should expect to observe an endowment effect and so a discrepancy between WTA and WTP (which should in turn inform one's choice of measure). Those are two kinds of EV inferences one might draw from the results of laboratory experiments on loss aversion and the endowment effect.

https://doi.org/10.1086/684084 Published online by Cambridge University Press

One might be tempted to classify EV inferences depending on whether they are drawn from the results of field experiments, as in the context of EBP, or from the results of laboratory experiments, as in the context of the economic experiments described above. And, indeed, discussions of EV are often conducted either in one context (see, e.g., Cartwright and Hardie [2012] for EBP) or the other (see, e.g., Guala [2005] for experimental economics) but rarely across those two contexts (although see Steel 2008). A more useful classification, I think, is based not on the nature of the experiments from which EV inferences are drawn but on the aim with which they are drawn. And there are two broad aims with which such inferences are drawn: prediction and explanation. Let me formulate the distinction between these two kinds of EV inferences as follows:

> **Predictive EV inference**. An account of predictive EV inference will tell you under what conditions an inference from the claim that the effect of a cause C on E in a population P_1 is of size m to the claim that C will have an effect of size m on E in a distinct population P_2 if it occurs there is a good inference.

> **Explanatory EV inference**. An account of explanatory EV inference will tell you under what conditions an inference from the conjunction of (1) the claim that the effect of a cause C on E in a population P_1 is of size m and (2) the observation of E in a distinct population P_2 to the claim that C has occurred in P_2 and has had an effect of size m on E there is a good inference.

This classification of EV inferences cuts across the two contexts described above. To be sure, most EV inferences drawn from RCTs in the context of EBP are in fact predictive, but they need not be. And while EV inferences drawn from laboratory experiments tend to be explanatory, at least some of them are predictive, as the second EV inference considered above in the case of the effect of loss aversion on the endowment effect illustrates. And the case of the Federal Communications Commission auction for telecommunication licenses—which Guala (2005) discusses under the heading "Economic Engineering"—is a concrete case in which a predictive EV inference has been drawn from the results of laboratory experiments.

There are various ways one might refine or reformulate the distinction between types of EV inferences drawn above. One could first replace the terms 'predictive' and 'explanatory' by the terms 'prospective' and 'retrospective', since predictive claims typically are about future events while explanatory claims typically are about past events. One could also appeal to Mill's distinction between the "causes of effects" and the "effects of causes" (Mill 1843, bk. 3, chap. 10). Predictive EV inferences are inferences about

https://doi.org/10.1086/684084 Published online by Cambridge University Press

the effects of causes; that is, they are inferences the conclusions of which are claims regarding which effects will occur if certain causes occur. Explanatory EV inferences, as I have characterized them above, are by contrast inferences regarding the causes of effects insofar as their conclusions are claims regarding what the causes of observed effects are.

Now that predictive and explanatory EV inferences have been distinguished, let me turn to an argument to the effect that, as far as predictive EV inferences are concerned, the problem of EV has been solved.

3. The Problem of Predictive External Validity Has Been Solved. The accounts of predictive EV developed by Cartwright and Hardie (2012) on the one hand and Bareinboim and Pearl (2013) on the other share a common core, even though they are set in different frameworks. Consider again the educational policy example introduced above. Traditional accounts of predictive EV (e.g., Cook and Campbell 1979) will enjoin Mary to ensure that her school district is sufficiently 'similar', in the relevant ways, to district d_0 before inferring to the claim that implementing policy P in d_1 will cause an increase of size m in the district-level mean SAT score. The Cartwright-Hardie and Bareinboim-Pearl accounts represent important advances because they tell you exactly the ways in which school district d_1 must be 'similar' to d_0 in order for Mary's predictive EV inference to be good.

The insight that is common to both accounts is that causes, including educational policy P, do not produce their effects in a vacuum. The size of the effect of P on SAT scores in any particular school S_i, for instance, will depend on a variety of school-level factors: the size of the classes in S_i, the quality of the teachers implementing P in S_i, and so on. Cartwright and Hardie (2012, 25) call these factors on which the size of the effect of P on SAT scores depends "support factors," while Bareinboim and Pearl (2013, 108) call them "difference-generating factors." Let me simply call them 'interactive factors', although I should make it clear that the interaction may be merely statistical; that is, these factors need not literally interact with the cause of interest.[1]

The idea behind both the Cartwright-Hardie and Bareinboim-Pearl accounts of predictive EV is that the inference from the claim that P had an effect of size m on SAT scores in d_0 to the claim that P will have the same effect if implemented in d_1 is a good inference if and only if the school-level interactive factors for P are 'the same' in d_1 and d_0. Let me briefly explain in more detail how the Cartwright-Hardie account develops this idea and how it explicates the meaning of 'the same' in the previous sentence. Limitations

1. One could also have considered district-, class-, and even student-level interactive factors here. Let me, however, focus on school-level interactive factors for the sake of simplicity.

https://doi.org/10.1086/684084 Published online by Cambridge University Press

of space unfortunately prevent me from discussing the Bareinboim-Pearl account and comparing it to the Cartwright-Hardie account. Let me simply say that these two accounts converge on similar sets of conditions for predictive EV inferences to be good.

Cartwright and Hardie (2012, 27) follow Mackie in embracing the view that causes are INUS conditions.[2] An INUS condition for some event *e* is an *insufficient* but *necessary* part of an *unnecessary* but *sufficient* condition for the occurrence of *e*. Mackie's classic example is that of a house fire caused by a short circuit (Mackie 1965). The short circuit is not individually sufficient to produce the fire. Other factors, which I called 'interactive factors' above, are required: the presence of flammable materials, the presence of oxygen, and so on. These interactive factors, together with the short circuit, are jointly sufficient for the occurrence of the fire.[3] But note that they need not be necessary. There are other ways to produce house fires; that is, there are alternative sets of factors (e.g., sets that have lit cigarettes instead of short circuits) that are sufficient to produce the same fire.

Consider again our educational policy example. In what follows, I sometimes refer to policy P by using the binary variable representing whether it is implemented, namely, X, and to SAT scores by using the variable Y_i. Call $\mathbf{V}$ the vector of variables $\{V_1, \ldots, V_n\}$, each of which represents a school-level interactive factor for the effect of X on Y_i in d_0. If class size is an interactive factor for policy P, for instance, then $\mathbf{V}$ will contain a discrete variable V_i representing the average size of classes in school S_i. A full specification of the configuration of interactive factors for a particular school S_i in d_0 is simply a particular realization $\{v_1, \ldots, v_n\}$ of $\mathbf{V}$. Now define variable Z as ranging over realizations of $\mathbf{V}$, so that each value z of Z represents a possible configuration of interactive factors for the effect of X on Y_i in a particular school in district d_0. The probability distribution $P_{d_0}(Z)$ simply describes the way in which these interactive factors are distributed in schools in district d_0.

According to Cartwright and Hardie, where predictability can be expected, we can assume that the effect of X on Y_i in a school S_i is governed by a "causal principle" (2012, 23). Without making any commitments as to its precise functional form, one can write out this causal principle as follows:

$$Y_i = f_{S_i}(X, Z). \tag{1}$$

2. They do not, however, hold the view that INUS conditions always are causes, a view that faces well-known counterexamples.

3. It is assumed here that causation is deterministic. One can, however, extend discussion to the probabilistic case by redefining an INUS condition for *e* as an insufficient but necessary part of an unnecessary but sufficient condition for an increase in the probability of *e* occurring.

https://doi.org/10.1086/684084 Published online by Cambridge University Press

Let me assume, for the sake of simplicity, that (1) holds of all schools in d_0. Given this assumption, the effect of X on Y_i not at the school level but now at the district level for d_0 is simply the sum of the school-level effect of X on Y_i for each value of Z weighted by $P_{d_0}(Z)$, the probability distribution over values of Z in d_0.

Given this setup, one can formulate the following argument, call it argument A, to bridge the gap between the effect X had on $\mathrm{E}[Y_i]$ in d_0 and the effect it will have if implemented in d_1:

1. X had an effect of size m on $\mathrm{E}[Y_i]$ in d_0.
2. The causal principle governing the school-level effect of X on Y_i in d_0 also holds in d_1 (i.e., [1] holds of all schools in d_1).
3. The interactive factors summarized by **V** are distributed in the same way in d_1 as in d_0 (i.e., $P_{d_0}(Z) = P_{d_1}(Z)$).

$\therefore$ X will have an effect of size m on $\mathrm{E}[Y_i]$ if implemented in d_1.

What argument A says, to put it slightly differently, is that if the same educational policy P is implemented in both d_0 and d_1, then if the interactive factors for P are distributed in the same way in both districts and if they interact with P in the same way in both, P will produce an effect of size m on the district-level mean SAT score in d_1 just as it did in d_0.[4] This is true as a matter of definition: because the conjunction of the occurrence of P and its required interactive factors in d_0 is assumed to be sufficient for the production of an effect of size m on $\mathrm{E}[Y_i]$, the occurrence of a qualitatively identical set of conditions in d_1 cannot but produce the same result.

According to Cartwright and Hardie, then, a predictive EV inference is a good inference just in case it takes the form of a sound deductive argument of the kind of argument A.[5] Their account thus gives a clear and unambiguous answer to the question of the circumstances in which two populations are 'similar' enough that one can extrapolate a causal claim from one to the other.

4. Two Consequences

Steel's 'Extrapolator's Circle'? Steel (2008, 4) claims that any satisfactory account of EV inferences (or extrapolations, as he calls them) should break what he calls the 'Extrapolator's Circle' (EC). What would that mean

4. Assuming here that there are no competing sets of jointly sufficient conditions poised to produce an opposite contribution to $\mathrm{E}[Y_i]$ in either d_0 or d_1.

5. It should be clear that the existence of such an argument is sufficient but not necessary for the truth of the conclusion that X will have an effect of size m on $\mathrm{E}[Y_i]$ if implemented in d_1. My focus here, however, is on the conditions under which one can infer the truth of this conclusion on the basis of claims about the effect of X in d_0 and about similarities between d_0 and d_1, not on the conditions under which this conclusion is true.

https://doi.org/10.1086/684084 Published online by Cambridge University Press

in the case I have used above? To break the EC in this case would be to show that one can establish that d_0 and d_1 are sufficiently similar (in the relevant respects) for a predictive EV inference from the former to the latter to be good without having to know so much about d_1 that one could accurately predict the effect P will have on SAT scores if implemented there without appealing to information about the effects of P in d_0. The idea behind the requirement that an account of EV inference break the EC is that such an account should not imply that EV inferences are redundant or pointless.

Does the Cartwright-Hardie account of predictive EV inferences break the EC? It seems that, in order to establish that premises 2 and 3 in argument A above are true, one must know a lot about d_1, since one must know what the relevant causal principle is in d_1 and how the required interactive factors are distributed there. And if one possessed this knowledge, then would one not be in a position to accurately predict the effect of P on SAT scores in d_1 without having to appeal to information about d_0? If this is the case, then the Cartwright-Hardie account does not break the EC, and so it is not, by Steel's standards, acceptable as an account of predictive EV inferences.

Should one then reject the Cartwright-Hardie account? This seems too hasty a conclusion. I claimed above that this account gives the correct answer to the question of the circumstances under which predictive EV inferences are good. Assuming that this claim is true, what more could one want of an account of EV inferences? One might argue that such an account should, in addition, tell you how go about drawing such inferences. But one should be careful to distinguish between *analyses* (for lack of a better word) of the conditions under which EV inferences are good inferences and *methods* for drawing such inferences. Cartwright and Hardie (2012) provide both in their book-length treatment of EV inferences in EBP. For instance, they give methods for finding out about interactive factors. Breaking the EC might be a legitimate desideratum for a method for drawing EV inferences, but it is not for analyses of the conditions under which they are good inferences. Such analyses should be evaluated according to whether they provide clear and principled accounts of the conditions under which one can infer from the truth of a causal claim in one population to its truth in another population. And, by this standard, the Cartwright-Hardie account is successful.

Deductive EV Inferences. EV inferences, whether predictive or explanatory, are widely assumed to be inductive. Guala (2005, 196) expresses this received view when he claims, for instance, that "external validity inferences surely involve an inductive step of some sort." If the Cartwright-Hardie account of predictive EV inference is correct, however, then this assumption is false: some EV inferences are deductive.

One might, to be sure, be uncertain as to the truth-values of the premises required for good EV inferences according to the Cartwright-Hardie account. One might, for instance, be uncertain as to whether the causal principle

https://doi.org/10.1086/684084 Published online by Cambridge University Press

governing the effect of X on $\mathrm{E}[Y_i]$ is the same in d_0 and d_1. But an uncertainty regarding the truth-values of the premises of argument A is obviously not the same thing as an uncertainty regarding whether the conclusion of this argument is entailed by its premises. I hope to have shown that, given the way Cartwright and Hardie define the notions of causal principle and interactive factors, there is little room to dispute the claim that argument A is a valid deductive argument.

One might worry that, by interpreting predictive EV inferences as purported deductive arguments of the kind of argument A, the Cartwright-Hardie account places the bar too high. Since it seems that, in any concrete case, at least some of the required interactive factors are likely to be 'local' (i.e., be present in the study population but not in the target one), the Cartwright-Hardie account implies that predictive EV inferences will rarely (if ever) be good inferences. Is this skeptical upshot problematic for Cartwright and Hardie? It is, after all, what motivates Steel, Guala, and others to reject an earlier account of EV inference due to LaFollette and Shanks (1995).

The skeptical upshot of the Cartwright-Hardie account would be problematic if (i) the predictive EV inferences we actually draw often are good inferences or (ii) if it implied that Mary, the superintendent of d_1, cannot learn anything valuable from the effect of P on SAT scores on d_0. The evidence for i, however, is hardly overwhelming. And ii is plainly false.

Let me start with i. Imagine that Mary implements P in d_1 on the basis of its effect on SAT scores in d_0 and that, as it happens, P produces the same result in d_1 as in d_0. Should we conclude that the inference drawn by Mary was a good predictive EV inference? No, because for all we know, Mary might have used naive induction ('It worked there, therefore it will work here') and simply gotten lucky. In the vast majority of cases, it is impossible to determine whether policy decisions that led to desirable results were in fact based on good predictive EV inferences. This is because, in the vast majority of cases, the process by which evidence is used to arrive at a policy decision is not public. And cases of bad predictive EV inferences leading to policy decisions with undesirable results are not hard to find (see, e.g., Cartwright and Hardie 2012, sec. II.B.4). An argument against the Cartwright-Hardie account thus can hardly rely on the premise that most of the predictive EV inferences we draw are in fact good inferences.

Consider now ii. To support a prediction regarding the effect of P on SAT scores in d_1, Mary will need to figure out whether the required causal principle and interactive factors are present there. And looking at the effect of P in d_0 can be useful for that purpose, even if argument A is unsound because the relevant causal principles differ or because some of the required support factors differ between d_0 and d_1. For instance, if teacher quality is a support factor for P in d_0, then it would be wise for Mary to check whether it might also play this role in d_1. In other words, what happened with P in d_0

https://doi.org/10.1086/684084 Published online by Cambridge University Press

does not become entirely irrelevant to what will happen with P if implemented in d_1 simply because the predictive EV inference from one to the other would not be a good inference. This means, however, that we should probably stop thinking of the problem of using knowledge garnered from experimental or observational studies to inform policy decisions as being identical to the problem of drawing predictive EV inferences.

5. Conclusion. I have argued above that, as far as predictive EV inferences are concerned, the problem of external validity has been solved. The accounts independently developed by Cartwright and Hardie on the one hand and Bareinboim and Pearl on the other converge on the correct answer to the question of the circumstances under which one can infer from the truth of a causal claim in some population to its truth in another population.

If this claim is true, then two consequences immediately follow. First, Steel's demand that any acceptable account of EV inferences break the Extrapolator's Circle is misplaced. Although the demand might be legitimate for methods for drawing EV inferences, it is not for analyses of the circumstances under which such inferences are good. Second, not all EV inferences are inductive, contrary to what Guala and others have claimed. According to the Cartwright-Hardie account, predictive EV inferences are good just in case they are sound deductive arguments of the kind of argument A.

REFERENCES

Bareinboim, Elias, and Judea Pearl. 2013. "A General Algorithm for Deciding Transportability of Experimental Results." *Journal of Causal Inference* 1:107–34.

Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer, and Robert Sugden. 1997. "A Test of the Theory of Reference-Dependent Preferences." *Quarterly Journal of Economics* 112:479–505.

Campbell, Donald. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54:297–312.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Cook, Thomas, and Donald Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Kahneman, Daniel, Jack Knetsch, and Richard Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98:1325–48.

LaFollette, Hugh, and Niall Shanks. 1995. "Two Models of Models in Biomedical Research." *Philosophical Quarterly* 45:141–60.

Mackie, J. L. 1965. "Causes and Conditions." *American Philosophical Quarterly* 2:245–64.

Mill, John Stuart. 1843. *A System of Logic*. London: Parker.

Starmer, Chris. 2000. "Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk." *Journal of Economic Literature* 38:332–82.

Steel, Daniel. 2008. *Across the Boundaries*. Oxford: Oxford University Press.

https://doi.org/10.1086/684084 Published online by Cambridge University Press