*plim *argmax

A great deal of empirical work in microeconometrics is conducted on the basis of its importance for real-world decisions and public policy. In the vast majority of cases this requires using empirical findings from populations that are different to those in which a potential policy would be implemented.[1] If microeconometric analysis identified stable parameters of a fully-specified model of the relevant relationships, and sufficient data was available on the population of interest, then extrapolation would be merely mechanical.[2] That scenario is rarely, if ever, realised in modern economics: economists have insufficient knowledge to correctly specify full structural models *ex ante* and the data available is often insufficient to identify and estimate all the parameters.

The structural literature on program evaluation has, in recent decades, attempted to address these shortcomings; most notably by developing reduced-form representations of structural parameters, and methods for bounding estimated treatment effects when certain identifying assumptions are relaxed. In the interim, however, there has been a dramatic shift in the predominant methods employed in applied microeconometrics: implementation and analysis of randomized experiments has become the preferred basis for identifying average causal effects.[3] The rationale is that data based on experimental (or otherwise plausibly exogenous) interventions offer the prospect of achieving identification of causal effects without requiring the knowledge or assumptions needed to adequately specify a structural model. The influence of this view has, in some respects, reoriented the research process: rather than beginning with a policy question of interest, the availability of actual or 'natural' experiments determines what policy questions are analysed. Scepticism of the — statistical or structural — assumptions required for identification of causal effects, or relationships, from observational data has been an explicit basis for preferring the experimental approach. However, unlike the case of a fully specified structural model, extrapolation does not follow mechanically from successful identification using experimental methods. While the appeal of the experimental approach is that it offers the prospect of identification without committing to assumptions about relationships between key variables, the absence of such assumptions further limits the basis for extrapolation.

Despite that, the rise of this 'design-based' literature has not been accompanied by corresponding development of rigorous methods for extrapolating experimental results to populations or questions of interest.[4] Yet, empirical analysis that identifies a causal relationship (achieves 'internal validity') without being generalisable outside the experimental sample ('external validity'), arguably has no *formal* basis for informing policy decisions.[5]

1 Proposal overview

Given the above, the review would examine the extant literature on external validity as it pertains to applied microeconometrics, with a particular emphasis on the use of estimated treatments effects to inform public policy decisions.[6] With a large number of empirical papers of this kind now being published annually, the review should therefore be of interest to many economists — as well as practitioners in the area of 'evidence-based policy' and researchers from a variety of other disciplines.

Necessarily, a review of this kind has to locate the issue of external validity within recent debates on the merits of design-based versus structural methods. A number of contributions have shown, or argued, that certain methods employed in the design-based literature are effectively nested by (are equivalent to specific formulations of) structural methods. Understandably, the proposed approach for finding common ground that has emerged from these contributions — including in an earlier review for this journal (Heckman, 2010) — is one that emphasises the role of a sufficiently generic structural model: referred to, following Roy (1951), as the *generalised Roy model.* Related to this, much of the existing criticism of design-based methods and their limitations has been premised on the concern of structural econometricians with the implications of subjects' behavior for internal validity. Failures of internal validity lead to corresponding failures of external validity: if design-based methods estimate a treatment effect different from the one that researchers intend, this will not provide the information needed to determine the policy effect of interest in a new population.

The relative merits of structural versus design-based methods in achieving external validity will be a focus of the review. However, adopting the structural approach from the outset would arguably pre-empt some important questions of interest – such as appropriate conceptualisations of external validity, and how extrapolation challenges manifest under different assumptions. The proposed structure is, instead, as follows.

The review will contain five main sections. The first is a novel synthesis of the disparate literatures on external validity of experimental results. This provides the reader with a broad background to recent developments and, in doing so, enables a better understanding of why, and how, the problem of extrapolation has been (relatively) neglected in economics and elsewhere. Furthermore, it serves as a basis for subsequent definition and use of the term 'external validity'. The concept originates in influential contributions

[6] External validity concerns, generically formulated, are also relevant to macroeconomics. However, the persistent differences in technical methods employed, and predominant concerns, of the microeconomic and macroeconomic literatures mean that attempting a single review encompassing both is too ambitious.

(Campbell and Stanley (1966), Cook and Campbell (1979))to the literature on experimental methods in social science, but is now used more broadly and often without being clearly specified. It follows from the microeconometric definition of external validity that factors compromising internal validity also compromise external validity. For the sake of completeness, this section therefore briefly notes conceptual and practical challenges to implementing successful experiments. Some consideration is also given to the contrast between the microeconometric literature, in which external validity is defined so as to require internal validity, and the macroeconometric literature in which this remains an open debate (albeit framed in different terms).

The second section discusses the problem of external validity within the standard, counterfactual framework of the design-based literature. The primary purpose of this section is to provide an exposition of the obstacles to external validity that exist even when the standard, statistically-formulated identification assumptions are satisfied. Under some formulations, this is the problem of moving from a *sample average treatment effect* to a *population average treatment effect.* As indicated above, the aim is to avoid pre-empting discussion of the tensions between this literature and the structural approach – which is primarily concerned with violations of these identifying assumptions implied by individuals' optimising behavior. Those tensions are addressed later in the review. The section also notes three problems — social interaction among subjects, general equilibrium effects and non-linear treatment effects — that are left for consideration in the fourth section of the review.

Insights from the earlier experimental methodology literature and more recent contributions in econometrics are reviewed, and synthesized, to show that the most basic obstacle to external validity can be represented by interaction effects in causal production functions. This formulation implies heterogenous treatment effects — in the sense of treatment effects that vary with individuals' (observed or unobserved) characteristics — and relevant aspects of that literature are located accordingly. Given inconsistent usage in the literature, an explicit distinction is made between 'treatment effect heterogeneity' as referring to covariate-dependent treatment effects, and the consequences of varying compliance or selection *in the presence of* covariate-dependence.[7]

[7] For example, Heckman, Urzua and Vytlacil (2006) refer to 'heterogenous treatment effects' as relating to heterogeneity across *unobservables*, while selection based on this unobserved heterogeneity is referred to as 'essential heterogeneity'. Both usages are po-

Given this, it is straightforward to show that formal requirements for external validity are analogous to (better-known) formal requirements for obtaining causal identification using observational data: where simple causal identification requires balance across treatment and control populations, simple causal extrapolation requires balance across the experimental sample and the population of policy interest. While some researchers may believe this point to be obvious, until very recently papers in the design-based literature have made no attempt to formally, or systematically, address this challenge. And there is presently no consistent guidance or consensus in the economics literature on how to formally conduct such extrapolation.

The final part of the second section reviews a series of relatively recent theoretical and empirical contributions, in econometrics and statistics, that seek to address this lacuna in the design-based literature. The theoretical literature examines different approaches to estimating heterogeneous treatment effects and performing covariate-based extrapolation. The empirical literature illustrates how different forms of interaction effects have led to a failure of external validity. In some cases, reflecting the absence of an agreed approach, the empirical papers also contain *ad hoc* attempts to formulate methods for extrapolation, or tests for failure of external validity. Together these contributions provide a starting point for a more systematic consideration of the requirements and methods for successful, formal extrapolation of estimated treatment effects when using data from ideal randomized experiments.

The third section of the review relaxes the assumption of perfect random assignment and compliance, considering the implications of selective participation in treatment when treatment effects are heterogeneous. Besides attempting to eliminate such concerns through exogenous manipulation (randomization), the design-based literature addresses this concern primarily through consideration of 'quasi-experiments' and the *local average treatment effect* (LATE) (Imbens and Angrist, 1994). The section begins, therefore, with a review of the implications of the LATE literature for external validity. This then provides a natural transition to consideration of the structural literature on program evaluation.

In contrast with the LATE framework, the structural approach allows a broader conception of the influence of individuals' optimising behavior on

tentially confusing for the broader audience targeted by the proposed review. Related matters are covered in the discussion of marginal treatment effects.

estimated treatment effects. The challenges arising from participant behavior are, broadly, a function of two issues: selection into, or out of, treatment; and, responses to receipt or non-receipt of treatment. As regards program evaluation, the two leading approaches — which are not mutually exclusive — that address these issues are better described as 'reduced form structural models'.

The first, developed primarily from work by Heckman and Vytlacil (2005), defines the concept of a *marginal treatment effect* (MTE) within the Roy model (referred to above). Along with other treatment effects, LATE is shown to be a weighted average of MTEs. The MTE is further used to define what the authors call the *policy relevant treatment effect* (PRTE), which they argue can be used to predict the effects of new policies in new populations. The second approach, introduced in its most basic form by Manski (1990), takes the alternative perspective of *partial identification*: considering identification of informative *bounds* on treatment effects when relaxing some of the identifying assumptions described in section 2. For both approaches, the focus of the review is on elucidating the additional contribution of these frameworks to our understanding of the problem of external validity, as well as methods for obtaining it.

The final part of the section considers the relationship between the two reduced-form structural approaches and the challenges to external validity posed by treatment-covariate interactions discussed in section 2. This is an issue that has received little direct attention: conditioning on covariates is left implicit in most contributions to the structural literature in order to focus on the specific problems arising from selection. One consequence, for example, is that the discussion of external validity in the MTE literature has tended to focus on the more challenging problem of predicting the effects of new policies — as opposed to predicting the same policy in new populations. Yet the fact that estimated propensity scores play a fundamental role in empirical analysis using the MTE framework indicates that these two representations of covariates — as modifiers of treatment effects and determinants of selection — need to be reconciled when considering external validity.

The fourth section considers issues relevant to external validity that go beyond the standard program evaluation framework: non-linear treatment effects, social interaction, general equilibrium effects and other forms of scale effects. By non-linear treatment effects we mean instances in which the elasticity of the causal effect of treatment is not constant.[8] Under social interaction, program participants' outcomes are affected – for a variety of possible reasons – by the treatment status of others; peer effects in education are one example. The leading example of the importance of general equilibrium effects for external validity is in predicting the likely society-wide effect of labor market interventions. Finally, if scale effects are defined as referring to any differences in treatment outcomes that result solely from differences in the number of treatment recipients, they overlap with the aforementioned challenges. However, scale effects so-defined may also reflect factors such as limits to implementer capacity, effects on participant beliefs and so forth. The section aims to provide a succinct characterisation of the failures of external validity that result from all these factors, along with a review of any solutions that have been proposed in the extant literature.

The fifth section returns to the motivation behind the concern with external validity: optimal policy decisions. This issue can be framed as a cost-benefit analysis (CBA) problem — which is the usual approach in the design-based literature — or as a more complicated, decision-theoretic problem. In practice, the former approach has largely focused on fairly simple comparisons of treatment costs and average treatment effects. The latter takes a much broader view in considering the implications of different decision-maker welfare functions, the importance of ambiguous or uncertain treatment effects, and of course the relevance of treatment heterogeneity. Both approaches are reviewed with an emphasis on their relevance to external validity. Specifically, the definition of external validity can be reformulated to address the question: would the optimal policy implied by the estimated treatment effect in the program sample be the same as the optimal policy in the population of policy interest?

The paper concludes with a set of preliminary implications for empirical work and an assessment of fruitful directions for future theoretical and methodological contributions. While the review clearly identifies contributions from the literature that can improve empirical work and policy advice, it is evident that the dominant approaches in the design-based literature to date have been insufficiently rigorous as regards external validity. And though the structural literature does more to address challenges to external validity that arise from behavior of participants, the degree to which its

[8] For example, a reduction in class size by five students may result in a 0.3 standard deviation improvement in average test scores when the initial class size is fifteen students, but only a 0.1 standard deviation improvement when the initial size is thirty students.

methods are likely to be widely applicable and empirically successful remains an open question. Furthermore, because it uses reduced form models, the structural literature faces similar challenges to the design-based literature as regards interaction (heterogenous) effects and covariate-based extrapolation. A preliminary conclusion is that, given these considerations, a greater degree of caution would seem to be appropriate in making claims about policy relevance of econometric program evaluations than is currently the case.

2 Relevance of the proposed topic to JEL readers

The recent 'revolution' in the use of experimental methods in empirical work has now spread to many areas of empirical work, but largely originated in development economics (Angrist and Pischke, 2010). In that field the primary motivation for many studies, and their funders, was testing interventions in order to inform policy decisions. This is also the typical motivation for program evaluation studies using structural methods. However, taking the challenge of policy advice seriously means, as noted by Manski (2013*a*), tackling the problem of external validity. The proposed review will therefore be of interest to the large number of economists engaged in program evaluation using microeconometric methods, whether design-based or structural, as well as those more directly involved in policy work.

The expectation is that the review, as outlined above, will provide the reader with the following:

- a comprehensive introduction to external validity as a concept, and synthesis of previous contributions in economics and other disciplines (where relevant)

- a detailed exposition of the challenge posed by external validity in the presence of an ideal experiment, what recent contributions have proposed to address this, as well as empirical and theoretical considerations in implementing these

- a review of the contribution of quasi-structural methods to characterising additional obstacles to external validity and developing possible solutions

- consideration of potentially more intractable challenges to external validity such as non-linear treatment effects, social interaction and general equilibrium effects

- an assessment of the implications of the above for the strength of claims that can be made using findings from program evaluations contingent on techniques currently available, theoretical knowledge and data limitations.

3 Main references to be included in the review

The literature on external validity in economics and related disciplines is now developing rapidly, but many contributions remain disconnected from each other. While explicit efforts to address the issue in the microeconometric literature on program evaluation are quite recent, there is a larger and older literature in experimental economics focusing on the so-called 'artificiality of laboratory experiments'. In the broader literature, important contributions span econometrics, statistics, machine learning and generic methodological work in quantitative social science. With a few exceptions, these contributions are better organized by the components of the external validity problem they address. I propose delineating contributions using the following subheadings, though in the actual review some contributions will naturally merit consideration under multiple headings.

Background: RCT debates The first component of the literature concerns important methodological contributions that have questioned the manner in which treatment effects from randomized control trials (RCTs) are employed to inform policy. Banerjee and Duflo (2009), Card, DellaVigna and Malmendier (2011) and List and Metcalfe (2014) provide overviews of the adoption of these methods in applied microeconometrics. Imbens and Wooldridge (2009) give an overview of the various econometric approaches to program evaluation. Arguably the two most widely-read critiques of the typical use of RCTs in economics have been by Deaton (2008, 2010) and Keane (2005, 2010*a*,*b*), though related issues had been previously debated by others including Heckman and Smith (1995) and Burtless (1995). Furthermore, the contentious issues are not limited to economics: Deaton draws on work in philosophy of science by Nancy Cartwright, such as Cartwright (1989, 2007) — a more explicit summary of which can be found in Cartwright (2010).

These critiques can be seen as a reaction to the case made most explicitly by Angrist and Pischke (2010), and further defended by Imbens (2010), that randomized program evaluation constitutes a 'credibility revolution' in economics. Among the other notable comments on such claims are: Leamer (2010), who identifies the issue of 'interactive confounders' taken-up below; Heckman and Urzua (2010) and Heckman and Vytlacil (2007*a*)'s critique based on an extensively developed, structural econometric theory of treatment effects; Manski (2011) who argues that policy analysis based on randomized evaluations is too often premised on 'incredible certitude'; and, Wolpin (2013) who expands on Keane's theme of inference without theory. The review by Imbens (2013) of Manski (2013*a*), and response by Manski (2013*b*), in this regard is also informative. Although further consideration of macroeconomics is excluded from this review, the discussion by Giacomini (2015) of the relative predictive success of structural versus statistical models in macroeconomics provides a useful contrast with the emphasis on causal identification in microeconometrics.

Synthesis of previous and parallel literatures on external validity The references above provide the immediate rationale for reviewing the literature on external validity, although this evidently has considerable merit in-and-of itself for reasons already outlined in the introduction to this proposal. The term 'external validity' was coined by Campbell and Stanley (1966) and Cook and Campbell (1979) in their work on experimental methods in the social sciences. Though they use few econometric or statistical formalisms, those authors' insights regarding obstacles to extrapolation, as well as possible solutions, remain relevant to current developments. In the philosophy of science literature, Cartwright (2011*a*,*b*) has been influential in emphasizing — with reference to economics — the importance of addressing the question of external validity and whether a trialled treatment 'will work for us'.

The question of external validity has arisen elsewhere in the economics literature in relation to the 'artificiality' of laboratory experiments in experimental economics and there are some useful insights that are relevant to the proposed review. Harrison and List (2004), List (2011) and Al-Ubaydli and List (2015) discuss the merits of laboratory versus field experiments in that context. Other notable contributions are Guala (2005), Binmore (1999) and Samuelson (2005).

Arguably a more useful literature, from the perspective of using RCTs for policy, comes from empirical analysis in medicine. In that case the problem for practitioners is to determine the relevance of results from randomized trials for individual patients. Notable references are Rothman and Greenland (1998) and Rothwell (2005*a*,*b*, 2010, 2006). That literature has given rise to more formal work on extrapolation Cole and Stuart (2010) in epidemiology, which has expanded into the broader statistical literature discussed further below.

External validity with an ideal experiment In the context of an ideal experiment, external validity arises as a concern in program evaluation when treatment effects differ across variables that themselves differ across contexts or populations of interest. Consequently, Cook and Campbell (1979) argued that challenges to external validity are fundamentally about interaction between the variable of interest and other factors. The program evaluation literature to date has dealt with the issue of causal interaction indirectly, through consideration of 'heterogeneity' in treatment effects. One problem has been that because these analyses are 'atheoretical' (Keane, 2010*b*), the variables across which authors test for simple heterogeneity are often chosen in an *ad hoc* manner that in some cases amounts to specification searching. Crump et al. (2008) provide a detailed formal discussion and test for heterogeneity to account for such concerns. (Studies such as Heckman, Smith and Clements (1997) and Djebbari and Smith (2008) take the issue of heterogeneity in a different direction, covered further below in references from the structural literature). However, little attention has been paid to formalising the implications of identified heterogeneity for obtaining external validity. Similar issues arise in the statistics literature, where extensive discussions of interaction analysis (Egami and Imai, 2014; VanderWeele, 2015) include no reference to the implications for external validity.

In the context of concerns about external validity of results from RCTs, but in the absence of formal guidance on how to obtain it, some contributions attempt to examine whether external validity appears to hold empirically. The typical strategy is to compare treatment effects of similar, or identical, programs in different populations or with different implementing institutions. Notable examples are the studies by Allcott and Mullainathan (2012); Allcott (2015), Bold et al. (2013), Pritchett and Sandefur (2013, 2015), Fischer and Karlan (2015), Vivalt (2015) and Gechter (2015).

It is only in the last decade that formal methods for testing or obtaining external validity have begun to be developed. Following Cook and Campbell (1979): if treatment effects in experimental samples differ from those in policy populations due to interacting factors, then an intuitive solution is to estimate the likely effect in new populations by accounting for population differences in the interacting variables. This is, in fact, the underlying theme of constructive proposals in the current literature, but typically without reference to the underlying role of interaction. The first attempt to address the extrapolation problem systematically in this fashion was by Hotz, Imbens and Mortimer (2005). More recently a number of authors have examined the prospects of utilizing propensity scores for this process in analogous fashion to the use of propensity scores in the literature on matching estimators of causal effects: Cole and Stuart (2010), Stuart et al. (2011), Hartman et al. (2015), Tipton (2013, 2014) and O'Muircheartaigh and Hedges (2014). Muller (2015) notes that explicitly recognizing the role of causal interaction reveals the direct similarities between requirements for external validity from ideal experiments and requirements for internal validity using observational data.

A related approach, anticipated by Cook and Campbell (1979) and discussed also by Allcott and Mullainathan (2012) and Tipton (2013), is to consider the difference in average treatment effects across populations as a *sampling* problem. That suggests two possible, potentially complementary, solutions. The first is selecting experimental samples that are maximally similar to the populations of policy interest. This has not found much favor because it is often impractical, future policy interest may not be clear and many authors aspire to generalizing their experimental results beyond a single policy population. A second approach, considered by Solon, Haider and Wooldridge (2015), is to represent the extrapolation problem as a sample re-weighting problem. Leaving aside the specific re-weighting methods used, the resultant process is analogous to the one based on the propensity score approach.

Besides the above contributions to the statistics and econometric literature based on counterfactuals, which is now familiar to many economists, the problem of extrapolation has also been addressed using the less well-known approaches of algorithmic extrapolation ('machine learning') and causal graphs. Bareinboim and Pearl (2011, 2013, 2014) and Pearl (2015) are the main contributions in the causal graphs literature to date. Athey and Imbens (2015) and Kleinberg et al. (2015) provide discussions of machine learning approaches to estimating heterogeneous treatment effects and extrapolating results for policy purposes. It is appropriate to consider these contributions in this section as they concern methods that are more statistical, than structural, in nature.

Compliance, selection and partial identification As noted above, much of the econometric literature on 'treatment effect heterogeneity' in fact concerns problems that arise from compliance or selection *in the presence of* heterogeneity of effects across individuals. The widely-used approach to estimating and interpreting LATEs – following Imbens and Angrist (1994) and given a textbook treatment in Angrist and Pischke (2009) – does this within the design-based framework, focusing on selective compliance in response to an instrumental variable. Similar to developments in the statistical literature cited above, Angrist and Fernandez-Vál (2013) examine the problem of extrapolating estimated LATEs utilising the propensity score. One recent example of an empirical contribution based on this approach is Bisbee et al. (2015). Another important tool within this quasi-experimental literature is the regression discontinuity design: DiNardo and Lee (2010) and Bertanha and Imbens (2014) consider the problem of external validity for such methods.

The general criticism that treatment effects obtained by practitioners favouring the design-based framework may have limited relevance for policy purposes has particularly focused on the LATE. In important early contributions, Heckman (1996) emphasized the importance of viewing randomized experiments as instruments and Vytlacil (2002) showed the equivalence between the assumptions of the LATE model and those of a Roy-type latent index model. Putting these issues in the context of the broader microeconometric literature, Heckman, Schmierer and Urzua (2010) show the equivalence between the correlated random coefficient model and the generalized Roy model.

The key contribution in this literature is Heckman and Vytlacil (2005), which follows Imbens and Angrist (1994) but also builds on earlier work by Heckman and Robb (1985). The authors define the MTE – originally described by Björklund and Moffitt (1987) – using the Roy model, and show that other treatment effects defined in the literature can be expressed as functions of the MTE. It is critical to note (Heckman, 2010) that the MTE differs from LATE only when – in addition to heterogeneity of effects across participants – the treatment effect is correlated with treatment receipt. It is in this sense that the MTE literature is structural, and extends the preceding econometric literature on selection into the program evaluation literature through consideration of instrumental variable methods. As already noted, however, the approach is not that of a full structural econometric analysis; the discussions by Nevo and Whinston (2010) and Todd and Wolpin (2010) argue the merits of that approach, the limitations of which have already been briefly mentioned above and in detail by Heckman (2000, 2008) and Keane (2010*b*).

There are a number of overlapping contributions to this strand of the structural literature. Extensive surveys of the approach are provided by Heckman and Vytlacil (2007*a*,*b*), as well as Heckman and Urzua (2010). One useful aspect of Heckman and Vytlacil (2007*b*) is their distinction between external validity and prediction of new policies, along with consideration of the associated exogeneity assumptions required. Some notable individual papers are: Heckman and Vytlacil (2001), which introduces policy-relevant treatment effects; and Heckman, Urzua and Vytlacil (2006), which proposes and illustrates tests for the presence of selection effects ('essential heterogeneity'). Among the empirical contributions to this literature, besides some of the papers already cited, are Moffitt (2008) and Carneiro, Heckman and Vytlacil (2011).

However, the critical point for the purposes of the review is that the vast majority of these papers are not explicitly concerned about external validity: the focus is on what instrumental variables actually estimate in the presence of selection into treatment. Consideration of external validity is largely implicit in arguments that the authors' reduced form structural models imply the identification of stable ('structural') parameters in many cases of interest. It is that which then warrants – in the parlance of authors in the philosophy literature such as Cartwright (1989) – claims about the relevance of these estimates for other populations and policies. A useful recent exception is Kowalski (2015), who explicitly examines the additional explanatory power provided by using MTE-based methods for explaining disparate findings of similar interventions across different populations.

An alternative approach to failure of the assumptions underlying the design-based approach is to calculate bounds on estimated treatment effects. An early summary of this approach is provided by Manski (1990), with a detailed review in Manski (2003) and more recently by Tamer (2010). As above, the issue of interest is how this alternative approach to identification impacts on the external validity of empirical results. That question is addressed to some extent by Manski (2008, 2011, 2013*a*), but with an emphasis on the decision problem faced by policy makers – see references below.

Non-linearities, social interaction and general equilibrium effects Among the critics of the limitations of RCTs, Keane (2010*a*) notes the lack of consensus on the magnitude of causal effects of various policy interventions and the associated problem of nonlinearity in these effects. This arises in particular when researchers wish to predict the effect of an intervention in which treatment 'intensity' is different to that of past experiments. Heckman and Vytlacil (2007*b*) note that to the extent that this arises from functional form of causal relationships, it poses a challenge to both the design-based and reduced form structural literatures that may only be resolvable by expanding the support of the data on treatment interventions.

The literature on social interactions and econometric identification is significant in both size and complexity (Durlauf and Ioannides, 2010). It encompasses large sub-literatures on topics such as conspicuous consumption, peer effects in education and neighbourhood effects (Durlauf (2004); Kling, Liebman and Katz (2007)) in social outcomes. Manski (1993, 2000*a*) outlined the primary obstacles to identification of such effects. Brock and Durlauf (2001) review the literature on identification, to which Blume et al. (2015) makes a recent contribution in relation to linear interaction models. Garfinkel, Manski and Michalopolous (1992) is one of the key references in this literature as regards implications for program evaluation. Such effects may be relevant to external validity in a variety of ways, including: undermining internal validity (through influence between treatment and control groups); introducing complex forms of scale-dependence; and, changing the scope of the policymaker's decision problem through spillover effects. One influential, and still contested, empirical attempt to address the effects of one type of interaction on identification is the paper on an intervention to treat intestinal worms by Miguel and Kremer (2004).[9] In recent contributions: Baird et al. (2014) examine how experiments can be designed to account for possible interactions; and, Angelucci and Di Maro (2016) provide a survey for researchers of experimental and non-experimental methods for achieving identification of the treatment effects of interest.

A related problem for program evaluation is that of general equilibrium effects. Besides Garfinkel, Manski and Michalopolous (1992), Heckman, Lochner and Taber (1999*a*, *b*) considered this problem in the context of taxation-financed education policies. Such effects are of particular concern for predicting the impact of programs involving complex causal relationships, such as education (Moffitt, 2006) and labor market (Smith (2000), Heckman, Lalonde and Smith (1999)) interventions, when taken to a large scale. Heckman (2001) and Heckman and Abbring (2007) consider the problem from the perspective of social welfare maximization, which raises the issue addressed by the fifth section of the proposed review.

[9] Though note that there is also variable usage of terms here – Angelucci and Di Maro (2016) define the issues considered by Miguel and Kremer (2004) as 'externalities'.

A decision-theoretic approach to external validity An alternative perspective on external validity is the policy maker's question as to whether the effect of an analysed intervention is likely to be stable enough that the *optimal policy decision* remains the same in another population. The analysis of the first four sections of the review consider external validity only from the perspective of the stability of the effect across contexts, or the information it can be provide on new interventions. However, given that program evaluation is often motivated by a desire to inform policy it is important to give adequate attention to external validity as it relates to decision making. A related issue, which is increasingly arising in the epidemiological literature, is the question of optimal treatment choice by practitioners for individual patients - see MacLeod (2016).

Garfinkel and Manski (1992), Heckman and Smith (1998), Heckman (2001) and Berger, Black and Smith (2001) are important early contributions in relation to social welfare maximisation and individual allocation.[10] Manski (2000*b*, 2004, 2008, 2013*a*) and Dehejia (2005) situate program evaluation as a decision theoretic issue, and consider the associated implications from the literatures on choice under uncertainty and ambiguity. Linking to the preceding literature on partial structural models, Eisenhauer, Heckman and Vytlacil (2015) provide assumptions under which it is possible to identify and estimate heterogeneous marginal cost and benefits of treatment.

[10] Brock, Durlauf and West (2003, 2007) – and comment by Leeper and Sargent (2003) – are analogous contributions in macroeconomics.

References

Allcott, Hunt. 2015. “Site selection bias in program evaluation.” *Quarterly Journal of Economics*, 30(3): 1117–1165.

Allcott, Hunt, and Sendil Mullainathan. 2012. “External validity and partner selection bias.” *NBER Working Paper*, 18373.

Al-Ubaydli, Omar, and John A List. 2015. “Do Natural Field Experiments Afford Researchers More or Less Control than Laboratory Experiments?” *American Economic Review (Papers & Proceedings)*, 105(5): 462–466.

Angelucci, Manuela, and Vincenzo Di Maro. 2016. “Programme evaluation and spillover effects.” *Journal of Development Effectiveness*, 8(1): 22–43.

Angrist, Joshua, and Ivan Fernandez-Vál. 2013. “ExtrapoLATE-ing: external validity and overidentification in the LATE framework.” In *Advances in Economics and Econometrics: Theory and Applications, Econometric Society Monographs, Tenth World Congress (Vol.III).* , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel. Cambridge University Press.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics.* Princeton:Princeton University Press.

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

Athey, Susan, and Guido Imbens. 2015. “Machine Learning Methods for Estimating Heterogeneous Causal Effects.” *arXiv preprint arXiv:1504.01132.*

Baird, Sarah, Aislinn Bohren, Craig McIntosh, and Berk *Ozler.*2014.*“DesigningExperimentstoMeasureSpilloverEffects."WorldBankPolicyResear*

Banerjee, Abhijeet V, and Esther Duflo. 2009. “The experimental approach to development economics.” *Annual Review of Economics*, 1: 151–178.

Bareinboim, Elias, and Judea Pearl. 2011. “External Validity and Transportability: A Formal Approach.” *JSM Proceedings: Section on Statistics in Epidemiology*, 157–171.

Bareinboim, Elias, and Judea Pearl. 2013. "A general algorithm for deciding transportability of experimental results." *Journal of Causal Inference*, 1(1): 107–134.

Bareinboim, Elias, and Judea Pearl. 2014. "Transportability from Multiple Environments with Limited Experiments: Completeness Results." *Advances of Neural Information Processing*, 27: 280–288.

Berger, MarkC., Dan Black, and JeffreyA. Smith. 2001. "Evaluating profiling as a means of allocating government services." In *Econometric Evaluation of Labour Market Policies*. Vol. 13 of *ZEW Economic Studies*, , ed. Michael Lechner and Friedhelm Pfeiffer, 59–84. Physica-Verlag HD.

Bertanha, Marinho, and Guido W Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." *NBER working paper*, 20773.

Binmore, Kenneth. 1999. "Why experiment in economics?" *Economic Journal*, 109: 16–24.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. 2015. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect." National Bureau of Economic Research Working Paper 21663.

Björklund, A, and R Moffitt. 1987. "The estimation of wage gains and welfare gains in self-selection." *Review of Economics and Statistics*, 69(1): 42.

Blume, Lawrence E, William A Brock, Steven N Durlauf, and Rajshri Jayaraman. 2015. "Linear social interactions models." *Journal of Political Economy*, 123(2): 444–496.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ngángá, and Justin Sandefur. 2013. "Scaling Up What Works: experimental Evidence on External Validity in Kenyan Education." *Center for Global Development Working Paper*, , (321).

Brock, William A., and Steven N. Durlauf. 2001. "Chapter 54 - Interactions-Based Models." In . Vol. 5 of *Handbook of Econometrics*, , ed. James J. Heckman and Edward Leamer, 3297 – 3380. Elsevier.

Brock, William A., Steven N. Durlauf, and Kenneth D. West. 2003. "Policy Evaluation in Uncertain Economic Environments." *Brookings Papers on Economic Activity*, 1: 235–302.

Brock, William A., Steven N. Durlauf, and Kenneth D. West. 2007. "Model uncertainty and policy evaluation: Some theory and empirics." *Journal of Econometrics*, 136(2): 629 – 664.

Burtless, Gary. 1995. "The case for randomized field trials in economic and policy research." *Journal of Economic Perspectives*, 9(2): 63–84.

Campbell, Donald T, and Julian C Stanley. 1966. *Experimental and Quasi-experimental Designs for Research.* Chicago:Rand McNally College Publishing.

Card, David, Stefan DellaVigna, and Ulrike Malmendier. 2011. "The Role of Theory in Field Experiments." *Journal of Economic Perspectives*, 25(3): 39–62.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review*, 101: 2754–2781.

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

Cartwright, Nancy. 2007. *Hunting Causes and Using Them: approaches in philosophy and economics.* Cambridge:Cambridge University Press.

Cartwright, Nancy. 2010. "What are randomised controlled trials good for?" *Philosophical studies*, 147: 59–70.

Cartwright, Nancy. 2011*a*. "Evidence, External Validity, and Explanatory Relevance." In *Philosophy of Science Matters: The Philosophy of Peter Achinstein.* , ed. Gregory J Morgan, 15–28. New York:Oxford University Press.

Cartwright, Nancy. 2011*b*. "Predicting 'It Will Work for Us': (Way) Beyond Statistics." In *Causality in the sciences.* , ed. Phyllis Illari McKay, Federica Russo and Jon Williamson. Oxford (UK):Oxford University Press.

Cole, Stephen R., and Elizabeth A. Stuart. 2010. "Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial." *American Journal of Epidemiology*, 172(1): 107–115.

Cook, Thomas D, and Donald T Campbell. 1979. *Quasi-Experimentation: design and Analysis Issues for Field Settings.* Wadsworth.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. "Nonparametric tests for treatment effect heterogeneity." *Review of Economics and Statistics*, 90(3): 389–405.

Deaton, Angus. 2008. "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development." *Keynes Lecture, British Academy.*

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, 48(2): 424–455.

Dehejia, Rajeev H. 2005. "Program evaluation as a decision problem." *Journal of Econometrics*, 125: 141–173.

DiNardo, John, and David S. Lee. 2010. "Program Evaluation and Research Designs." *NBER working paper*, 16016.

Djebbari, Habiba, and Jeffrey Smith. 2008. "Heterogeneous impacts in PROGRESA." *Journal of Econometrics*, 145: 64–80.

Durlauf, Steven N. 2004. "Chapter 50 Neighborhood effects." In *Cities and Geography.* Vol. 4 of *Handbook of Regional and Urban Economics*, , ed. Kenneth J Arrow and Michael D Intriligator, 2173 – 2242. Elsevier.

Durlauf, Steven N, and Yannis M Ioannides. 2010. "Social Interactions." *Annual Review of Economics*, 2: 451–478.

Egami, Naoki, and Kosuke Imai. 2014. "A New Approach to Causal Interaction." unpublished working paper.

Eisenhauer, Philipp, James J. Heckman, and Edward Vytlacil. 2015. "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs." *Journal of Political Economy*, 123(2): 413–443.

Fischer, Greg, and Dean Karlan. 2015. "The Catch-22 of External Validity in the Context of Constraints to Firm Growth." *American Economic Review (Papers & Proceedings)*, 105(5): 295–299.

Garfinkel, Irwin, and Charles F Manski, ed. 1992. *Evaluating welfare and training programs.* Cambridge (MA):Harvard University Press.

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

Gechter, Michael. 2015. “Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India.” *unpublished working paper.*

Giacomini, Rafaella. 2015. “Economic theory and forecasting: lessons from the literature.” *The Econometrics Journal*, 18(2): 22–41.

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

Harrison, Glenn W, and John A List. 2004. “Field experiments.” *Journal of Economic Literature*, 42(4): 1009–1055.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. 2015. “From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3): 757–778.

Heckman, James, and Edward Vytlacil. 2007*a*. “Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation.” In *Handbook of Econometrics (Volume 6B).* Vol. 6B, , ed. James Heckman and Edward Leamer, Chapter 70, 4779–4874. Amsterdam:Elsevier.

Heckman, James, and Edward Vytlacil. 2007*b*. “Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments.” In *Handbook of Econometrics (Volume 6B).* , ed. James Heckman and Edward Leamer, Chapter 71, 4875–5143. Amsterdam:Elsevier.

Heckman, James, and Jaap H Abbring. 2007. “Econometric evaluation of social programs, Part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation.” In *Handbook of Econometrics (Volume 6B).* , ed. James Heckman and Edward Leamer, Chapter 72, 5145–5303. Amsterdam:Elsevier.

Heckman, James, and Richard Robb. 1985. “Using longitudinal data to estimate age, period and cohort effects in earnings equations.” In *Cohort analysis in social research: Beyond the identification problem.* , ed. William M Mason and Stephen E Fienberg, 137–149. New York:Springer Verlag.

Heckman, James J. 1996. “Randomization as an Instrumental Variable.” *Review of Economics and Statistics*, 78(2): 336–341.

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. 2001. “Accounting for heterogeneity, diversity and general equilibrium in evaluating social programmes.” *Economic Journal*, 111(2): 654–699.

Heckman, James J. 2008. “Econometric Causality.” *International Statistical Review*, 76: 1–27.

Heckman, James J. 2010. “Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy.” *Journal of Economic Literature*, 48(2): 356–98.

Heckman, James J, and Edward Vytlacil. 2001. “Policy-relevant treatment effects.” *American Economic Review (Papers Proceedings)*, 91(2): 107–111.

Heckman, James J, and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects and Econometric Policy Evaluation.” *Econometrica*, 73(3): 669–738.

Heckman, James J, and Jeffrey A Smith. 1995. “Assessing the Case for Social Experiments.” *Journal of Economic Perspectives*, 9(2): 85–110.

Heckman, James J, and Jeffrey Smith. 1998. “Evaluating the welfare state.” In *Econometrics and Economic Theory in the Twentieth Century.* , ed. S Strom, 241–381. Cambridge:Cambridge University Press.

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

Heckman, James J, Daniel Schmierer, and Sergio Urzua. 2010. “Testing the Correlated Random Coefficient Model.” *Journal of Econometrics*, 158(2): 177–203.

Heckman, James, Jeffrey Smith, and Nancy Clements. 1997. “Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts.” *Review of Economic Studies*, 64(4): 487–535.

Heckman, James J, Lance Lochner, and Christopher Taber. 1999*a*. "General equilibrium cost benefit analysis of education and tax policies." *NBER working paper*, 6881.

Heckman, James J, Lance Lochner, and Christopher Taber. 1999*b*. "Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy." *Fiscal Studies*, 20(1): 25–40.

Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. "Chapter 31 - The Economics and Econometrics of Active Labor Market Programs." In . Vol. 3, Part A of *Handbook of Labor Economics*, , ed. Orley C. Ashenfelter and David Card, 1865 – 2097. Elsevier.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *Review of Economics and Statistics*, 88(3): 389–432.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics*, 125: 241–270.

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

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

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent developments in the econometrics of programme evaluation." *Journal of Economic Literature*, 47(1): 5–86.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Keane, Michael. 2005. "Structural vs. Atheoretic Approaches to Econometrics." *Keynote Address at the Duke Conference on Structural Models in Labor, Aging and Health.*

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

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction policy problems." *American Economic Review (Papers & Proceedings)*, 105(5): 491–495.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Kowalski, Amanda. 2015. "Marginal Treatment Effects and the External Validity of the Oregon Health Insurance Experiment." *Unpublished working paper.*

Leamer, Edward. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives*, 24(2): 31–46.

Leeper, Eric M, and Thomas J Sargent. 2003. "Policy Evaluation in Uncertain Economic Environments: Comments and Discussion." *Brookings Papers on Economic Activity*, 1: 302–322.

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

List, John A., and Robert Metcalfe. 2014. "Field experiments in the developed world: an introduction." *Oxford Review of Economic Policy*, 30(4): 585–596.

MacLeod, W. Bentley. 2016. "Viewpoint: The Human Capital Approach to Inference." *NBER working paper*, 22123.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *American Economic Review*, 80(2).

Manski, Charles F. 1993. "Identification of endogenous social effects: the reflection problem." *Review of Economic Studies*, 60: 531–542.

Manski, Charles F. 2000*a*. "Economic analysis of social interactions." *Journal of Economic Perspectives*, 14(3): 115–136.

Manski, Charles F. 2000*b*. "Identification and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice." *Journal of Econometrics*, 95(2): 415–442.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions.* New York:Springer-Verlag.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica*, 72(4): 1221–1246.

Manski, Charles F. 2008. *Identification for Prediction and Decision.* Cambridge (MA):Harvard University Press.

Manski, Charles F. 2011. "Policy Analysis with Incredible Certitude." *Economic Journal*, 121(554): F261–F289.

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

Manski, Charles F. 2013*b*. "Response to the review of 'Public policy in an uncertain world'." *Economic Journal*, 123: F412–F415.

Miguel, Edward, and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica*, 72(1): 159–217.

Moffitt, Robert. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Annales d'Économie et de Statistique*, , (91/92): 239–261.

Moffitt, Robert A. 2006. "Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective." In *Scale-Up in Education: Ideas in Principle (Volume 1).* , ed. Barbara Schneider and Sarah-Kathryn McDonald, 173–186. Cambridge:Rowman and Littlefield.

Muller, Seán M. 2015. "Causal interaction and external validity: obstacles to the policy relevance of randomized evaluations." *World Bank Economic Review*, 29: S217–S225.

Nevo, Aviv, and Michael D. Whinston. 2010. "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference." *Journal of Economic Perspectives*, 24(2): 69–82.

O'Muircheartaigh, Colm, and Larry V Hedges. 2014. "Generalizing from unrepresentative experiments: a stratified propensity score approach." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2): 195–210.

Pearl, Judea. 2015. "Generalizing Experimental Findings." *Journal of Causal Inference*, forthcoming.

Pritchett, Lant, and Justin Sandefur. 2013. "Context matters for size." Center for Global Development Working Paper 336.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review (Papers & Proceedings)*, 105(5): 471–475.

Rothman, Kenneth J, and Sander Greenland. 1998. "Causation and causal inference." In *Modern Epidemiology.* , ed. Kenneth J Rothman and Sander Greenland, Chapter 2, 7–28. Lippincott, Williams and Wilkins.

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

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

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

Rothwell, Peter M. 2010. "Commentary: External validity of results of randomized trials: disentangling a complex concept." *International Journal of Epidemiology*, 39: 94–96.

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

Samuelson, Larry. 2005. "Economic theory and experimental economics." *Journal of Economic Literature*, 43(1): 65–107.

Smith, Jeffrey. 2000. "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies." *Swiss Journal of Economics and Statistics (SJES)*, 136(III): 247–268.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?." *Journal of Human Resources*, 50(2): 301 – 316.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2): 369–386.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annual Review of Economics*, 2: 167–195.

Tipton, Elizabeth. 2013. "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts." *Journal of Educational and Behavioral Statistics*, 38(3): 239–266.

Tipton, Elizabeth. 2014. "How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations." *Journal of Educational and Behavioral Statistics*, 39(6): 478–501.

Todd, Petra E., and Kenneth I. Wolpin. 2010. "Structural Estimation and Policy Evaluation in Developing Countries." *Annual Review of Economics*, 2(1): 21–50.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press.

Vivalt, Eva. 2015. "Heterogeneous Treatment Effects in Impact Evaluation." *American Economic Review (Papers & Proceedings)*, 105(5): 467–470.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1): 331–341.

Wolpin, Kenneth. 2013. *The Limits of Inference Without Theory.* MIT Press.