I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15151

Treatment Effect Heterogeneity

Jeffrey Smith

MARCH 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15151

Treatment Effect Heterogeneity

Jeffrey Smith
University of Wisconsin-Madison and IZA

MARCH 2022

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Treatment Effect Heterogeneity*

Knowledge of treatment effect heterogeneity or "essential heterogeneity" plays an important role in our understanding of how programs work and in the design of systems to allocate them among the eligible. This paper provides a relatively non-technical survey of the current state of the treatment effect heterogeneity enterprise within economics from both substantive and applied econometric perspectives. It also suggests directions for research on treatment effect heterogeneity going forward.

JEL Classification: C10, C41
Keywords: treatment effects, essential heterogeneity, program evaluation

Corresponding author:
Jeffrey Smith
Department of Economics
University of Wisconsin
Sewell Social Science Building
1180 Observatory Drive
Madison, WI 53706
USA
E-mail: econjeff@ssc.wisc.edu

* This paper draws on my presentation entitled "The Implications of (Treatment Effect) Heterogeneity for Internal and External Validity" at the conference "Rigorous Impact Evaluation in Europe: A Conference in Honor of Alberto Martini." I thank the conference organizers for including me. In addition to the conference participants, Dan Black, Jacob Klerman, and Lois Miller provided helpful feedback and Ian Lundberg provided a useful pointer to the literature.

Introduction

Program evaluation has a long history both inside and outside of economics.[1] In the last 25 years or so, an important subtheme of the literature addresses heterogeneity in causal effects, both methodologically as a factor in the interpretation of estimates (especially estimates based on instrumental variables methods) and substantively in the design of programs and in the design of systems to more effectively target programs at those most likely to benefit from them. This paper considers the present state of the treatment effect heterogeneity enterprise, offers some critiques of current practice, and outlines various avenues for improvement.

The paper proceeds as follows: I begin in the conventional manner with notation and terminology. Following that, I discuss sources of treatment effect heterogeneity and why treatment effect heterogeneity matters. Then I consider how to obtain evidence of treatment effect heterogeneity, which leads naturally into some remarks on how to choose which subgroups (and other moderators) to look at and how to interpret the resulting estimates. In a novel twist, the final section provides concluding remarks.

Notation and terminology

Consider the standard potential outcomes framework, variously anticipated by or attributed to Mill (1843), Frost (1920), Neyman (1923), Fisher (1935), Roy (1951), Quandt (1972) and Rubin (1974). In this framework, each combination of a unit (a person, a firm, a county or whatever) and a treatment choice (which may be the null treatment or control condition) has associated

[1] Within economics, relatively early contributions include the work around the National Supported Work Demonstration summarized in Hollister, Kemper and Maynard (1984), Ferber and Hirsch (1981) on social experiments, as well as the non-experimental "old testament" of Heckman and Robb (1985). Outside economics, see e.g. Cook and Campbell (1979).

with it some potential outcome, which indicates the outcome the unit would experience given that treatment choice. The "potential" refers to the fact that the researcher only ever observes the outcome associated with at most one treatment choice in a given time period for each unit; the outcomes associated with other treatment choices thus necessarily remain unrealized potentials.

For simplicity, I focus here on binary treatments, with the treatment indicator $D = 1$ for treated units and $D = 0$ for untreated or control units. Adopting standard potential outcomes notation, Y_{1i} denotes the treated outcome for unit "i" and Y_{0i} denotes the untreated outcome for unit "i". The framework handles both discrete outcomes, such as employment, and continuous outcomes, such as earnings. The treatment effect for unit "i" consists of the difference between the treated and untreated potential outcomes, or $\Delta_i = Y_{1i} - Y_{0i}$. Until about 20 years ago, most of the literature assumed, usually implicitly, the strong restriction of a common treatment effect, so that $\Delta_i = \Delta$ for all "i". More recently, researchers have sought to relax this assumption. For example, a very important strand of applied econometric work initiated by Imbens and Angrist (1994) attempts to come to terms with the implications of treatment effect heterogeneity in the context of instrumental variables estimators.

I (mostly) assume no spillovers across units, so that the treated and untreated outcomes for a given unit do not depend on the treatment choice of any of the other units. The statistics literature applies the (awkward) label "Stable Unit Treatment Value Assumption" (SUTVA); more recent work in econometrics and statistics talks about "interference" between units.

Precisely because each unit reveals only one potential outcome in the data, the literature that concerns itself with treatment effect heterogeneity focuses almost exclusively on various averages of treatment effects. The Average Treatment Effect (ATE), defined in terms of the notation as $E(\Delta_i) = E(Y_{1i} - Y_{0i})$, provides the expected effect of treatment on units in some

population. The Average Treatment Effect on the Treated (ATET or TOT), defined as $E(\Delta_i|D_i = 1) = E(Y_{1i} - Y_{0i}|D_i = 1)$, gives the expected effect of treatment on those who choose (or get chosen) to take it, within some population.

As a running example, consider the treatment effect of a marriage counseling program for couples thinking about marriage, call it Marriage is Really Fun (MIRF), similar to the actually existing Building Strong Families program evaluated in Wood, Moore, Clarkwest, and Killewald (2014). For the example, let the population consist of never-married heterosexual couples ages 18-25 who respond to an email solicitation and let the experimental treatment consist of the offer of MIRF counseling. Further assume that not all individuals assigned to the treatment group actually receive the counseling and that some control group individuals receive MIRF counseling (or similar counseling from other programs) on their own from other sources. Thus, the experiment has "imperfect compliance" in both directions, because some members of the treatment group choose to become "no-shows" while some members of the control group "substitute" into the same (or very similar) treatment from other sources.[2]

The experimental data provide an unbiased estimate of the ATE for the offer of MIRF counseling within the study population, a parameter the literature (sometimes) calls the mean impact of the "intention to treat" or ITT. In contrast, the experimental data do not provide an obvious estimator of the ATE of actual receipt of MIRF counseling. In the absence of control group substitution, the Bloom (1984) estimator, which divides the experimental mean difference estimate by the fraction receiving counseling in the treatment group, provides the ATET within the study population for receipt of counseling. This estimator equates to using random assignment to the treatment group as an instrument for receipt of counseling. In the presence of

[2] See e.g. Heckman, Hohmann, Smith and Khoo (2000) for more on imperfect compliance and responses thereto.

control group substitution, using random assignment to the treatment group as an instrument for receiving counseling identifies a Local Average Treatment Effect (LATE) within the study population.[3] In particular, the LATE corresponds to the average impact of MIRF counseling on those who receive it if assigned to the treatment group but do not receive it if assigned to the control group.[4] With two-sided non-compliance, the experimental data do not provide a simple estimator for the ATET of counseling; see Black, Joo, LaLonde, Smith and Taylor (2020) for a discussion of generalizing the LATE to the ATET.

Treatment effect heterogeneity lies implicit in the definitions of parameters such as the ATE, ATET and LATE. To make it explicit, consider the distinction drawn by Djebbari and Smith (2008), who partition the available heterogeneity into "systematic" and "idiosyncratic" categories. Systematic heterogeneity refers to heterogeneity captured by observed characteristics of the agents or the context of the program. For a marriage counseling program, this includes participant characteristics not affected by the treatment (or measured prior to its influence) such as race and ethnicity or years of schooling, as well as contextual characteristics such as the sex ratio in the relevant marriage market and also possibly program characteristics that vary across sites. Some bits of the literature call these moderators, and usefully distinguish them from mediators (also known as intermediate outcomes or markers). The heterogeneity in the treatment effect not captured by the available moderators Djebbari and Smith (2008) call idiosyncratic heterogeneity. Note that the partition into systematic and idiosyncratic heterogeneity depends on the set of available moderators; more moderators enlarges the systematic component and diminishes the idiosyncratic component of the overall treatment effect variation.

[3] The statistics literature calls this a "Complier Average Causal Effect" or CACE, where compliers are individuals who change their participation status in response to a change in a binary instrument.
[4] An assumption of "monotonicity" lurks in the background of this interpretation; that assumption holds that all individuals respond in the same direction (or not all) in response to a change in a binary instrument.

Sources of treatment effect heterogeneity

Thinking about how treatment effect heterogeneity might arise in particular evaluation contexts can aid in study design, in data collection (both qualitative and quantitative), and in interpretation. In this section I consider two broader sources of variation in individual treatment effects: heterogeneity in the underlying content of treatment and heterogeneous responses to homogeneous treatments. In terms of the notation, heterogeneity in treatment means variation in the content or meaning of D_i across units while heterogeneity in the treatment effect means variation in $\Delta_i = Y_{1i} - Y_{0i}$ across units. Within the former, I distinguish between different types of treatment and different amounts of the same treatment.

To start, consider the common case wherein a researcher combines multiple underlying services into a single binary treatment for evaluation purposes. For instance, evaluators often code up a binary indicator for participation in some active labor market program, yet such programs typically provide some participants with classroom training, some with subsidized on-the-job training, some with GED preparation, and still others with job search assistance. Those are very different services, all combined for the purposes of evaluation into a binary treatment indicator. Similarly, different staff members may administer a nominally homogeneous program quite differently, as with the caseworkers in the ethnographic study of the U.S. Temporary Assistance to Needy Families (TANF) program by Watkins-Hayes (2009) or the child protection investigators whose heterogeneous child removal proclivities provide identifying variation in Bald, Chyn, Hastings, and Machelett (2021).

Even within a somewhat narrower category such as classroom training, treatment heterogeneity exists. A participant receiving classroom training may learn about welding or

about Microsoft Office. As noted in McCall, Smith and Wunsch (2016) the literature lacks a systematic examination of what one might call the optimal coarseness of treatment definition. How finely should we define treatment for the purposes of a particular evaluation, in contexts where the underlying treatment itself is heterogeneous and may in a real sense differ for every single unit, as when caseworkers tailor their approach to each client? The researcher faces a tough trade-off between interpretability and statistical power or, put differently, between learning about the effects of the underlying heterogeneous treatments and the sample size available for studying each treatment.[5]

Weiss, Bloom and Brock (2014) emphasize what we might call the intensive margin of treatment heterogeneity. This type of treatment heterogeneity results from different participants consuming different amounts of a treatment as in e.g. Behrman, Cheng, and Todd (2004). It differs from the type of treatment heterogeneity just considered, which revolves around different types of treatment combined into the single treatment indicator D_i.

To see the intuition, suppose that the MIRF program comprises eight in-person sessions that combine lectures, group exercises, guided discussions and so on. In the context of an experiment that randomly offers the opportunity to enroll in MIRF to eligible couples, those couples choose the intensity of the treatment they receive by choosing how many of the eight sessions to attend. If all treatment group couples choose to attend all eight meetings, then we have no treatment heterogeneity because every couple receives the same "dose" of the treatment, in the jargon ported over from the literature on clinical trials in medicine. If all treatment group members choose to attend either no meetings or all eight meetings, then we have a very simple case of one-sided non-compliance with no heterogeneity in the treatment conditional on

[5] The epidemiology literature has thought a bit about this question; see, e.g., VanderWeele, Tyler, and Miguel Hernan (2013).

compliance (i.e. on receiving any of it). Finally, in the most general case, couples assigned to the treatment group may choose different numbers of meetings to attend, with some attending none, some one or two, and some all or almost all. In this case, estimating a true "dose-response function" (more clinical trials talk) that shows how the treatment effect varies with the number of meetings received becomes both feasible and interesting, subject to the important proviso that the experiment does not provide exogeneous variation in the dose of treatment received among treatment group members. As such, doing so will require non-experimental methods.

Even homogeneous treatments can have heterogeneous treatment effects. Budget set treatments represent the canonical example of homogeneous treatments. Examples include the Earned Income Tax Credit (EITC) studied in e.g. Lim and Michelmore (2018), the Connecticut Jobs First program studied in Bitler, Gelbach, and Hoynes (2006), and the Self-Sufficiency Project (SSP) in Canada, which received an experimental evaluation documented in Michalopoulos, Card, Gennetian, Harknett, and Robins (2000). All these treatments move around the schedule of implicit tax rates on earnings for the populations subject to them (some of the treatments do other things as well, but still homogeneously). More prosaically, they all subsidize additional earnings over some range relative to the pre-existing tax and transfer system. The same altered tax schedule applies to all, and without human intervention via caseworkers or other institutional gatekeepers.

Yet even here we could, and do, see systematic heterogeneity in the response to treatment. Consider the standard textbook model of labor-leisure choice in which individuals face some market wage at which they can choose (rather unrealistically) to work whatever number of hours they prefer in each time period. Against the value of the additional earnings they receive from working more, they balance the value of additional time at home. Assume that

the market wage and the value of home time vary among workers. At the margin, in this simple model, each worker who chooses non-zero hours assigns an equal value to the earnings associated with the marginal hour at their market wage and the home time associated with the marginal hour at their valuation of that home time. When wages and the value of home time vary across workers, so too will their hours choices.

Now consider the extensive margin response to an earnings subsidy, which captures any response in moving individuals in the target population from no employment to employment (and thus from zero earnings to positive earnings). Workers who choose zero hours value the last hour of home time more than the earnings associated with the first hour of work. For some, the two will be close to equal, for others, far apart. An earnings subsidy will move some workers for whom the two are roughly equal into work but for workers away from the margin, whether due to very high values of home time or very low values of work (or both), the subsidy does not change observed behavior, even though it moves the underlying latent index that compares the net utility of working with the net utility of not working. Thus, in this example, heterogeneity in the value of home time or heterogeneity in wages, or both, generates heterogeneous treatment effects in hours and employment from a homogeneous budget set treatment. One can easily construct similar examples in other substantive domains.

The preceding paragraphs offer examples of how to think about the nature and definition of a treatment, and about the choice problems that lead potential participants to participate or not in the treatment or to participate in different ways and in different amounts. Thinking about both of these dimensions can help inform a prior regarding the nature and extent of treatment effect heterogeneity in particular contexts. An informed prior can then guide decisions about data collection. For example, in the budget set case described above, the simple model I laid out

suggests that the effect of the budget set treatment on employment and earnings outcomes should vary among individuals as a function of the value of time spent at home. An evaluation of such a program could collect data on variables that proxy for the value of home time, such as the presence of young children in the household, or of sick or disabled individuals needing informal care, and use these as theoretically-informed moderating variables in an analysis of subgroup effects. Collecting such data transforms idiosyncratic heterogeneity into systematic heterogeneity, which makes it practically useful.

Why treatment effect heterogeneity matters

There are many reasons to care about treatment effect heterogeneity. This section describes five of the most important. First, following Deaton (2010), Deaton and Cartwright (2018), and many others, treatment effect heterogeneity may provide information about the mechanisms by which an intervention produces impacts.

Second, programs that engender widely varying treatment effects among their participants may also increase inequality in outcomes in undesirable ways. Some programs may actually make some participants worse off. Learning that a program has a lot of idiosyncratic treatment effect heterogeneity, particularly when it suggests negative impacts for some participants, may provide a spur to program improvement via more attention to implementation fidelity, to the screening of applicants, or to participants' progress through the program.

Third, understanding the nature and extent of treatment effect heterogeneity associated with a particular program and, moreover, the extent to which potential participants and/or program gatekeepers have some sense of individual treatment effects, plays a critical role in a complete understanding of the program participation process. For voluntary programs,

individuals must want to participate; their desire presumably springs primarily from expectations of a treatment effect sufficiently large to outweigh the costs of participation. The literature offers very little information regarding the ex ante beliefs of potential program participants about their person-specific impacts, the extent to which those beliefs coincide with empirical reality, and the extent to which they drive participation choices. Cunha and Heckman (2007) consider this question in the context of choosing to attend college or not; McKenzie (2018) provides some evidence in the context of an experimental evaluation of a business plan competition in Nigeria.

We know a bit more about what participants think about their impacts ex post. Their ex post beliefs about program impacts may influence their future choices regarding similar opportunities, as well as what they tell others about their program experience, and whether or not they support or oppose funding for similar programs. Heckman and Smith (1998) and Philipson and Hedges (1998) consider program dropout as a crude participant evaluation measure. Smith, Whalley and Wilcox (2020), Byker and Smith (2021), and Calónico and Smith (2021) consider whether ex post participant evaluations obtained from surveys of program participants correlate with econometric estimates of program impacts. In general, they do not, though Brudevold-Newman, Honorati, Jakiela, and Ozier (2017) offer a bit of optimism based on improved survey measures. Both the ex ante and ex post beliefs of participants (and potential participants) regarding their individual-specific treatment effects remain grossly under-studied.

The literature also offers some tantalizing bits of evidence regarding what program staff know about unit-specific treatment effects. In the context of active labor market programs, caseworkers often play a gatekeeper role. Bell and Orr (2002) report the findings from a very interesting exercise, in which they asked intake workers in the AFDC Homemaker Home Health Aide demonstrations to predict, prior to random assignment, the potential as a homemaker-health

aide of each person randomly assigned. They find that these measures of potential do predict outcome levels but do not predict (very well) the impacts of the program.[6]

Depending on the program context, caseworkers may function as more than just passive gatekeepers; instead, they may actively determine who does and does not pass through the gate, or which of several gates they may enter. Lechner and Smith (2007) provide disappointing evidence on the performance of caseworkers within Swiss active labor market programs at assigning their unemployed participants to particular service categories. Indeed, they find that the caseworkers do no better in terms of average employment outcomes than random assignment of participants to services.

The silver lining within the cloud of gatekeeper ignorance of unit-specific treatment effects concerns external validity. To the extent that gatekeepers include or exclude potential participants from a program based on beliefs about unit-specific treatment effects uncorrelated with actual unit-specific treatment effects, the estimated average treatment effects on participants generalize immediately to the (de facto) randomly excluded potential participant population.

Fourth, understanding the nature of treatment effect heterogeneity, and its implications for the interpretation of evidence, plays a crucial role in extracting meaningful policy guidance from evidence. For instance, consider the Waite and Gallagher (2000) book entitled *The Case for Marriage*. The heart of the book consists of a set of chapters, each of which reviews, in a serious way but at a level accessible to non-academics, the literature on the causal effect of marriage on a particular class of outcomes. Chapter 4, for instance, looks at the health effects of marriage, while Chapter 7 reviews the effects on the wages of the spouses. In general, the studies that undergird the literature reviews in these chapters estimate ATETs; that is, they estimate the

[6] The question wording regarding "potential" admits of multiple interpretations. I interpret it as asking about predicted impacts, but others, such as McKenzie (2018), do not.

effect of marriage on the married (or, in some cases, they read as if estimating a common effect, but the ATET represents a more reasonable interpretation). In contrast, the policy chapter at the end of the book reads as though these studies all estimate a common effect of marriage. As most of the literature finds positive mean effects of marriage on the married, interpreting these estimates as common effects, or as ATEs, implies we should spend with enthusiasm on marriage promotion policies.

Alternatively, if individuals have some (correct) sense of their likely treatment effect of marriage, and act on that basis, then the treatment effect of marriage on the unmarried, even the marginal unmarried, may differ substantially from the average treatment effect of marriage on the married. Thinking about the results in the book this way suggests a new and valuable line of research on the extent to which individuals know their treatment effect from marriage, and it completely changes the policy implications that one would draw from the substantive literatures so ably reviewed in the book's earlier chapters.

Fifth and finally, systematic treatment effect heterogeneity underlies statistical treatment rules that attempt to direct treatment toward those who will benefit the most from it. These rules take estimates $\widehat{\Delta(X)}$ of $\Delta(X)$, the systematic component of treatment effect heterogeneity, and use them to assign program offers, or program participation, or program components, to specific individuals. They have the potential to substantially improve program efficiency (in the economic sense of that term) and as a result inspire much of the practical interest in treatment effect heterogeneity.

The simplest (and by far the most common) form of statistical treatment rule informally adjusts program spending or targeting based on estimated differences in the mean effects of the program across one-dimensional subgroups. For instance, if an evaluation of our marriage

intervention MIRF found larger impacts (at the same cost) for high school completers than for college graduates, it would make sense to focus limited program resources on that group. Back in the 1990s, the experimental evaluation of the U.S. Job Training Partnership Act (JTPA, then the largest federal training program) found evidence of larger impacts for adults served by the program than for out-of-school youth. In response, Congress substantially reduced funding for the out-of-school youth portion of the program.[7]

More sophisticated statistical treatment rules involving multiple participant characteristics already exist in several policy domains. For example, the Worker Profiling and Reemployment Services (WPRS) system profiles new Unemployment Insurance (UI) claimants into mandatory re-employment services early in their UI claim based (in many states) on their predicted probability of benefit exhaustion. Of course, this represents profiling on a predicted value of Y_{1i} rather than on a predicted value of $(Y_{1i} - Y_{0i})$, but its advocates imagine a strong correlation between the two, something not obvious from studies such as Black, Smith, Berger and Noel (2003).

The criminal justice system and related criminology literature teem with statistical treatment rules. When applied to criminal sentencing, they bear the name "selective incapacitation," and date back to early work by the Rand Corporation.[8] In many cases, these rules also rely on predicted outcome levels, sometimes interpreted as proxies for predicted impacts. Bushway and Smith (2008) critique the literature on statistical treatment rules in criminology and attempt to introduce some of the knowledge gained from statistical treatment rules for labor market programs into the criminological conversation.

[7] See Bloom, Orr, Cave, Bell, and Doolittle (1993) for the 18-month impacts from the JTPA evaluation and Orr, Bloom, Bell, Doolittle, and Lin (1996) for the 30-month impacts.
[8] Harcourt (2006) provides an engaging history of these efforts, along with a critique of the whole idea.

Finally, the most technically sophisticated application I know of occurs in those health-related fields in which scholars use Sequential Multiple Assignment Randomized Trials (SMART) to develop compelling statistical treatment rules for "adaptive treatments." In this context, an adaptive treatment consists of a sequence of treatment recommendations wherein the later recommendations condition on patient responses to earlier treatments. Lei, Nahum-Shani, Lynch, Oslin, and Murphy (2012) lay out the methodological basics of SMART.

Evidence on the existence and extent of treatment effect heterogeneity

In a common effect world, it makes sense to talk about "the" treatment effect associated with a given intervention, as in "the" return to schooling or "the" effect of training. This paper argues for the general empirical irrelevance of the common effect model and, more concretely, calls for the systematic accumulation of evidence against that model in empirical applications.

Heckman, Smith and Clements (1997) provide and apply a basic non-parametric test of the null of the common effect model in their Appendix E.[9] The test presents a modest statistical challenge because the null hypothesis, that the variance of the treatment effects equals zero, lies on the boundary of the parameter space, because variances are by definition greater than or equal to zero. To see this issue in another way, consider an experimental evaluation (i.e. a randomized control trial) with equally sized treatment and control groups. In a common effect world, we have "rank preservation" meaning that the treatment preserves ranks in the outcome distribution. Thus, in a loose sense, the highest outcome in the treatment group outcome distribution represents the counterfactual for the highest outcome in the control group outcome distribution,

[9] Djebbari and Smith (2008) and Buhl-Wiggers, Kerwin, Muñoz-Morales, Smith and Thornton (2022) also apply the test. A small literature of related tests of the common effect null has emerged in econometrics in recent years. See Chung and Olivares (2021) for a recent example and further citations to this literature.

the medians represent each other's counterfactuals, the 12th percentiles represent each other's counterfactuals and so on. But in any given sample, treatment effects constructed by taking differences between treated and control outcomes at each rank will always have a non-zero variance just due to sampling variation, even if the common effect model holds in the population.

To get around this problem, Heckman, Smith and Clements (1997) propose comparing the variance constructed by taking (in their implementation) outcome differences at percentiles of the treatment and control distributions to the distribution of variances obtained using two random samples from the control group outcome distribution to construct the impact variance. The latter distribution of variances approximates the distribution under the null because both outcome distributions come from the control group so that all treatment effects equal zero (and thus the null of the common effect model holds) in the population. Heckman, Smith and Clements (1997) implement their test using data on adult females from the experimental evaluation of the U.S. Job Training Partnership Act, the largest federal employment and training program from 1982 to around 2000, and easily reject the common effect model null.

A simple version of this test for experimental data estimates quantile regressions of the outcome of interest on an indicator for assignment to the treatment group for many different quantiles, perhaps every percentile or, more modestly, every ventile. The common effect model implies equality of the quantile regression coefficients; Stata will do this joint test for you and correctly handle the statistical dependence among the estimates. It puzzles me that every experimental evaluation that considers "continuous" outcomes does not implement this test.[10]

[10] In contexts with continuous outcomes that include mass points in the control group outcome distribution, an even simpler test checks whether the mass point moves. For example, if 10 percent of the control group has zero earnings, then in a common effect world no observations in the treatment group should have zero earnings but 10 percent should have earnings equal to the common effect.

Magnitudes matter as well as (even more than?) the results of classical statistical tests. As described in Heckman, Smith and Clements (1997), the literature provides a quantitative lower bound on the variance of the impacts that builds on the Fréchet-Höffding bounds on the joint distribution of treated and untreated outcomes (i.e., of Y_{1i} and Y_{0i}) implicit in the marginal outcome distributions. Reporting this lower bound seems like something that should always happen in experimental evaluations. If the data indicate a substantively small lower bound on the impact variance, perhaps one worries a bit less about treatment effect heterogeneity; in contrast, a substantively large lower bound demands researcher attention to the issue of treatment effect heterogeneity.[11]

The lower bound on the impact variance corresponds to the treatment effects obtained under the assumption of rank preservation as described above. A crude (but likely sufficient in many contexts) way to construct an estimate of the lower bound simply obtains quantile treatment effects at every percentile of the outcome distribution and then takes their variance. Heckman, Smith and Clements (1997) perform a somewhat more computationally elaborate version of this exercise and obtain a substantively large value for the lower bound on the impact variance for the adult women in the JTPA experiment.[12]

One can combine the bounding approach with the estimation of subgroup effects (or context effects, etc.) in the following way: First, estimate the lower bound on the impact variance. Second, turn some idiosyncratic treatment effect heterogeneity into systematic heterogeneity by estimating some subgroup effects (I say more about how to choose the subgroups below). Subtract off the estimated subgroup effects from the outcomes and repeat the

[11] In practice, the upper bound usually turns out large enough that it provides little useful information.
[12] Though more computationally elaborate, their procedure still relies only on percentiles rather than some finer quantile and neglects to adjust for sampling variation in the estimated lower bound leading (most likely) to some modest overstatement.

bounding exercise. Large reductions in the lower bound on the impact variance in response to the removal of the available systematic treatment effect heterogeneity suggest perhaps less need to fuss about the remaining idiosyncratic heterogeneity. Section 6 of Djebbari and Smith (2008) applies this strategy to the data from the Progresa experiment while Bitler, Gelbach, and Hoynes (2017) and Ding, Feller, and Miratrix (2019) provide paper-length treatments that delve deeper into various methodological issues.

On the origin of subgroups

How do researchers choose which subgroups (or other moderators) to consider when attempting to locate systematic variation in the treatment effects of some intervention? Existing practice reveals five main pathways by which evaluations come to examine particular subgroups.

First, many evaluations examine basic demographic categories of male and female, or black, white and Hispanic due to equity concerns.[13] Second, (very) informal theory sometimes guides the ex post interpretation of differences in impacts between particular subgroups, if not always the ex ante decision to include them in the search for systematic treatment effect heterogeneity. For example, evaluations of educational interventions often consider subgroups defined by teacher experience with the idea that more experienced teachers may do a better job of implementing interventions. The site-level differences in impacts found in many evaluations of multi-site programs get casually attributed (because most such evaluations lack enough sites to do more) to differences in managerial quality and/or implementation fidelity among sites.

Third, sometimes the literature has identified subgroup differences in impacts in earlier evaluations of similar programs, so researchers continue to look for them in evaluations of

[13] At least in economics, researchers usually avoid making the useful distinction between the biology ("sex") and the social identity ("gender") and fail to go beyond traditional binary categories.

current programs, even without any clear (formal or informal) story to provide motivation. In the literature on employment and training programs in the U.S., the early evaluations summarized in LaLonde (2003) tended to find larger ATETs for female participants than for male participants. This pattern has lingered on in the literature for a very long time (several decades) without, so far as I know, any serious attempt to provide a theoretical underpinning that would provide some ideas about mechanisms that researchers could empirically examine or that might address concerns about confounding related to some missing moderator variable correlated with sex and/or gender.[14]

Fourth, some evaluations consider more-or-less ad hoc collections of potential subgroups comprised of whatever variables happen to appear in the evaluation dataset, in most cases presumably for reasons unrelated to the search for systematic variation in treatment effects. For instance, the 18-month impact report from the National Job Training Partnership Act Study, Bloom, Orr, Cave, Bell, and Doolittle (1993), offers the reader many pages of subgroup impact estimates (e.g., Exhibit 4.15 on pages 116 and 117 for adult women) that draw on nearly all of the variables collected at the time of random assignment via the study's Background Information Form (BIF). The only motivation offered, and that in a footnote, is the rather vague one of "relevance to policy discussions."[15]

Fifth, and finally, a few studies actually theorize ex ante about likely sources of systematic variation and then go looking for the variation predicted by their theories. Pitt, Rosenzweig, and Hassan (2012) develop a model of a "brawn-based economy" that explains differences between males and females in the extent of human capital investments and in the

[14] Though one could read the Lechner and Wiehler (2011) paper that considers fertility as an outcome in the evaluation of active labor market programs as providing such a theory.
[15] The footnote reassures the reader that "... we did not select them on the basis of the size or significance of the program effects presented here," surely a mortal sin relative to the venal sin of lack of theoretical motivation.

returns to those investments. Dillon and Smith (2020) examine how the impact of college quality varies with student ability, in light of theories in Rothschild and White (1995) and elsewhere that see them as complements in the production of educational and labor market outcomes. Though not exactly a study of subgroups in the narrow sense, Bitler, Gelbach, and Hoynes (2006) also illustrates this idea. They apply the standard static model of labor-leisure choice in economics to the context of Connecticut Jobs First welfare reform program and show that it predicts a particular pattern of quantile treatment effects. Upon estimating the quantile treatment effects using data from the experimental evaluation they find support for the model predictions in the data.[16] In my view, the literature would benefit from more of this fifth approach relative to the other four, at least when conceived of as serious, empirically grounded theory, not just wild speculation.

The sage in the machine

Recently fashionable machine learning methods take a very different approach to locating meaningful subgroup variation: they outsource the process to sophisticated statistical model selection algorithms. Recent papers that apply these methods to track down systematic heterogeneity include Buhl-Wiggers, Kerwin, Muñoz-Morales, Smith and Thornton (2022), Chernozhukov, Demirer, Duflo, and Fernández-Val (2018), Davis and Heller (2017), Knaus, Lechner and Strittmatter (2022), and Wager and Athey (2018).

In thinking about the application of machine learning methods to the problem of systematic treatment effect heterogeneity, I like to distinguish between two very different questions: a statistical one and an economic (or substantive) one. The statistical question asks:

[16] See also the fine paper by Kline and Tartari (2016).

given a set of variables that embody potential subgroups (or potential moderators more generally), how should the researcher locate substantively meaningful subgroup effects in the data? Machine learning methods address the statistical question by using well-defined algorithms that, in various ways, systematically search through the space of possible subgroup effects, subject to researcher guidance in the form of various tuning parameters, and report back on what they find. They have the potential to locate systematic impact variation not easily found via traditional approaches, such as second- or third-order interactions among moderators.

Of course, algorithmic model selection methods date back well into the last century. Stepwise regression, an antique model selection tool generally mocked by economists as atheoretic back in its heyday, provides one example. Linhart and Zucchini (1986) summarize the statistical model selection literature as of the early 1980s. Even newer methods date back farther than one might image when reading the recent economics literature; for instance, Section 5 of Heckman, Ichimura, Smith and Todd (1998) provides an analysis using Classification and Regression Trees (CART).[17] Relative to older statistical model selection techniques, machine learning methods exploit reductions in the cost of computing that allow much more thorough searches in model space. They also benefit from conceptual and computational developments over the past couple of decades and a wealth of practical experience built up via application of these methods in different empirical contexts in both industry and academia.

Like older formalized model selection algorithms, machine learning methods have the advantage of replicability. Two researchers applying the same method to the same data with the same set of potential moderators should learn the same lessons about subgroup effects. When

[17] Heckman, Ichimura, Smith and Todd (1998) seek to solve closely-related problem of choosing a functional form for their conditioning variables when estimating a treatment effect under the conditional independence assumption. See e.g. Farrell (2015) or Knaus (2021) for recent discussions of the application of machine learning tools to this problem.

combined with a pre-analysis plan specifying the moderators under consideration and the tuning parameter choices for the algorithm, they also tie the researcher's hands in ways that (at least partially) avoid issues of fishing for statistically significant results. An additional layer of protection from false positives comes from interpreting subgroup effects found via application of machine learning methods as "exploratory" rather than "confirmatory" (to borrow some useful jargon from outside of economics), i.e. as preliminary and not actionable in policy terms until they appear at least once more in a separate study. Finally, some machine learning methods, such as regression trees and the random forests that build upon them, greatly lower the cost to the researcher of exploring subtle interactions among moderators, something rarely done in the pre-machine-learning literature.

While machine learning tools promise real improvements in the ability of researchers to answer the statistical question I posed above, they provide no help whatsoever in addressing the second, substantive question: what potential moderators should the researcher consider? No machine learning algorithm can find subgroup effects when the data handed to the algorithm lack variables capturing membership in the relevant subgroups.[18] In addition to the important role of behavioral theory in suggesting candidate moderators noted earlier, institutional knowledge and the existing empirical literature have roles to play as well.

The literature on teacher effects (a.k.a. teacher value-added) provides a compelling example of the potential gains at the research margin to better theorizing and measuring of potential moderators. A large literature estimates teacher value-added, defined as the expected increase (or not) in scores on some standardized test associated with particular teachers.

[18] A related issue concerns the support of potential moderators in the data. For example, machine learning methods cannot estimate a subgroup effect for youth in data containing only adults. Hotz, Imbens and Mortimer (2005) emphasize this issue.

Hanushek and Rivkin (2012) ably review the substance and methods of this literature while Chetty, Friedman, and Rockoff (2014), Horváth (2015) and Buhl-Wiggers, Kerwin, Smith and Thornton (2022) provide recent empirical examples. Often, researchers consider how their estimated teacher effects covary with particular observed characteristics of teachers, such as their years of teaching experience, whether or not they have a Master's degree, their college major, how selective a college they attended, pre-college test scores, the usual demographics and so on. Perhaps surprisingly, other than years of experience the variables considered to date either do not matter enough for statistical detection or, if detectable, account for little of the observed variation.[19] Yet Jacob and Lefgren (2008) find that the subjective performance evaluations of school principals correlate with estimates of teacher value-added. Clearly principals observe features of teachers that researchers currently do not; research designed to measure those features would add a lot of value to the literature. While no one has, to my knowledge, attempted a full frontal assault on this question making use of the entire arsenal of machine learning methods, I read the literature as shouting out the need for better theory and better measurement, rather than better statistical model selection schemes.

Summing up, machine learning methods have a distinct role to play in the search for systematic variation in treatment effects; in the language of economics, they represent a complement to, rather than a substitute for, empirically and theoretically informed thinking about the likely nature of such heterogeneity.

[19] The lack of an effect for teacher MA degrees seems less surprising in light of an institutional context in which teacher union contracts provide automatic raises upon MA attainment while providers of MA programs compete for students by making the programs more enjoyable. See Wiswall (2013) for a somewhat contrarian view of the literature on teacher experience effects.

Speaking truth about power

Evaluators often power their evaluations to detect overall impacts of a particular magnitude. For instance, educational evaluations often aim to detect impacts of 0.2 standard deviations on a test score outcome. Sadly, sample sizes that allow the evaluator to detect meaningful overall impacts will not provide sufficient statistical power to detect much larger impacts for subgroups (or differential impacts between subgroups), even in the "best" case with only two subgroups of roughly equal size. Put differently, in most program evaluation contexts the available data impose real limits on the evaluator's ability to learn about effect heterogeneity. Of course, subgroup impact estimates still contain substantively valuable information even if they do not differ statistically from zero or from one another at conventional significance levels.

In some cases, evaluation researchers face an unhappy tradeoff between using an identification strategy with (usually) higher internal validity, such as a randomized control trial or a discontinuity design, but a smaller sample size, or using a (usually) less internally valid strategy based on selection on observed variables (i.e., conditional independence) applied to much larger numbers of observations drawn from administrative data. A focus on the detection of systematic heterogeneity makes the latter approach more attractive at the margin. See Klerman, Saunders, Dastrup, Epstein, Walton, Adam, and Barnow (2019) for a fine example of an evaluation that thinks a lot about statistical power and subgroup analyses and Klerman (2017) for a related discussion in the context of helping local program operators select a program variant for their site.

Interpreting subgroup effects

This section considers three issues related to the interpretation of estimated subgroup effects: These issues arise regardless of whether machine learning or more traditional methods produced the estimated effects and generalize to other types of moderators.

The first issue concerns confounding, not in the sense of a failure of conditional independence but in the sense that a given subgroup variable may only represent a proxy for the true causal source of the treatment effect heterogeneity. For example, in the Pitt, Rosenzweig and Hassan (2012) paper mentioned above, the data reveal subgroup differences in impacts by sex, but sex does not actually cause the difference in average treatment effects; instead, differences in physical strength that correlate with sex, rather than sex itself, drive the heterogeneity. When seeking to understand the mechanisms underlying average treatment effects, such differences matter, as they do when thinking about the external validity of subgroup effects estimated in particular populations. Both Hotz, Imbens and Mortimer (2005) and Muller (2015) highlight the confounding issue, which becomes even more important in contexts wherein the researcher dumps a bucket of potential moderators lacking either theoretical motivation or prior empirical foundation (or both) into a machine learning algorithm.

The second issue concerns heterogeneous treatment effects within subgroups. The fact that, say, men and women have different average treatment effects from some intervention does not imply that individual men and individual women do not have treatment effects that differ from the average treatment effects within the group to which they belong. Indeed, substantial overlap in the distributions of individual treatment effects in different groups can easily coincide with substantively meaningful differences in the group-level average treatment effects.

Surprisingly, many discussions of subgroup effects in evaluation research presume, and so see no need to justify, common effects within subgroups.

A simple example serves to illustrate one practical implication of the neglect of within-subgroup effect heterogeneity: Consider a program that serves both men and women, with a cost of participation equal to five for everyone. Half of the men in the relevant population have a treatment effect of 10, and half have a treatment effect of four; in contrast, among women, half have a treatment effect of 12 and half have a treatment effect of one. Assume that everyone in the population knows their treatment effect and the program cost, and that the population consists entirely of good economizers who participate when their treatment effect exceeds the cost and do not participate otherwise. In this world, half of men participate, namely those with an impact of 10, and half of women participate, namely those with an impact of 12. Doing an experimental evaluation that randomly assigns would-be participants yields subgroup impacts (exclusive of program costs) of 12 for women and 10 for men. The standard common effect interpretation of these subgroup estimates says that the program works better for women, and so it should serve more of them, perhaps via a targeted subsidy. In fact, at the margin, the program works better for men than for women, because the marginal untreated man has a treatment effect of four while the marginal untreated women has a treatment effect of one, so the usual interpretation, and the usual policy prescription that follows from it, yield the wrong answer. In essence, this example repeats the point about the difference in the treatment effect of marriage on average and at the margin, but in the context of subgroups.[20]

The third and final issue concerns whether or not to give a structural interpretation, in the sense that economists use that term, to differences in average treatment effects across subgroups

[20] Note that one could test the common effect model *within* subgroups using the methods already discussed.

or other moderators in particular substantive contexts. One could, and someone should, write an entire paper on this topic but I will confine myself to linking the economics notion of "structural" to standard assumptions in the treatment effects literature, and then making a couple of conceptual points that illustrate important interpretational dilemmas.

First, recall that economists use structural to mean "policy invariant"; like some relationships in the natural sciences (e.g., force = mass times acceleration) structural parameters or other objects such as production functions represent conceptual objects or technological relationships that do not depend on a particular policy environment. Compare this notion to the standard (probably too standard) SUTVA assumption in the treatment effects literature introduced earlier in the paper. Recall that SUTVA says that the treated and untreated outcomes for particular units are invariant to which other units, and how many other units, get treated. Formally, it states that (Y_{1i}, Y_{0i}) do not vary with $\{D_j\}, j = 1, \ldots, J$ where J is the total number of units in the relevant population. Put differently, SUTVA rules out spillovers or equilibrium effects.

Structure and SUTVA embody two distinct but closely related ideas. To see the distinction, consider the case where a different policy environment implies different counterfactual outcomes for potential participants in a particular program. As noted in Sandner, Cornelissen, Jungmann, and Herrmann (2018), home visiting programs have large impacts in the United States but smaller (or zero) impacts in Europe, perhaps because the untreated outcomes differ substantially in Europe due to the many other related programs on offer. In this example, the treatment effects fail the structural criterion but not necessarily SUTVA as these programs likely have relatively modest (if any) spillover effects. To see the commonality, consider the case of an active labor market program that improves the job search skills and/or increases the job

search effort of participants. Now consider a policy change that doubles the number of program participants. As described in Lise, Seitz and Smith (2004) and Crépon, Duflo, Gurgand, Rathelot, and Zamora (2013), we would expect programs that affect search skills or search effort to have spillover effects. If some workers search smarter and harder because of a program, other workers may find themselves displaced from jobs they otherwise would have taken, an effect on the untreated outcome that should increase with the number of treated individuals. In this case, both a structural interpretation of the treatment effects and SUTVA fail, because the policy change affects the untreated outcomes of everyone (and likely the treated ones too).

Now think about subgroup effects in light of the preceding discussion. Consider as an example the evaluation of some voluntary health intervention. Suppose that the intervention focuses on some non-contagious health condition so as to minimize issues with spillovers. Suppose further that an experimental evaluation estimates an ATET of 0.2 standard deviations on some health index for male participants and of 0.1 standard deviations for female participants. The following year, the intervention experiences a policy change that alters the participation patterns in the program while decreasing overall participation rates among both men and women. A new experimental evaluation estimates ATETs of 0.3 for men and 0.2 for women. Clearly, the ATETs for men and women do not merit a structural interpretation. At the same time, the experimental evidence does not rule out structural *individual* treatment effects nor a structural ATE in the population; under this interpretation, all of the difference in ATETs across years results from changes in the process of selection into the program that correlate with individual treatment effects. In this case, the reduced size of the program and increase in average impacts for both subgroups suggests a policy that induced stronger selection on program impacts.

The bottom line for this section: the potential outcomes framework has a substantive depth that many program evaluations fail to exploit, at the cost of learning less than they could about the nature of the treatment effects they study and about the policy relevance of their findings.

Summary and conclusions

Treatment effect heterogeneity matters both conceptually, for how we do applied econometrics and statistics, and substantively, for how we think about the findings of evaluations and for how we design programs and mechanisms that target programs. As my brief review shows, the literature has made tremendous progress on both fronts in recent years. At the same time, much remains to be done. The insights from the methodological literature (really literatures, inside and outside of economics) on treatment effect heterogeneity have penetrated deeply in some substantive domains while others remain largely untouched. The practical, and natural, generalization to the case of heterogeneous coefficients on continuous causal variables (i.e., the "correlated random coefficient model") remains largely aspirational. Too many policy discussions read as if programs have the same effect on everyone, everywhere and always.

I have suggested important ways forward via a program of deliberate testing of the null of treatment effect heterogeneity, which I predict the data will reject in nearly every context with sufficient statistical power, and routine estimation of the lower bound on the impact variance. I also argued for a deliberate program of theory and measurement to improve the extent to which we capture meaningful (and actionable) systematic variation in treatment effects in particular contexts. As argued by Weiss, Bloom, and Brock (2014), we need to have theories about where subgroup variation comes from and then we need to operationalize those theories in terms of

things that we measure and interact with the treatment indicator. The same applies to other moderators. Such a program complements rapid developments in machine learning methods. In sum, there are huge opportunities for research here, if people are looking for useful (and publishable) things to do.

References

Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett. 2021. "The Causal Impact of Removing Children from Abusive and Neglectful Homes." *Journal of Political Economy*. Forthcoming.

Behrman, Jere, Yingmei Cheng, and Petra Todd. 2004. "Evaluating Preschool Programs when Length of Exposure to the Program Varies: A Nonparametric Approach." *Review of Economics and Statistics* 86(1): 108-132.

Bell, Stephen and Larry Orr. 2002. "Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker-Home Health Aide Demonstrations." *Labour Economics* 9: 279-301.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review* 96(4): 988–1012.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2017. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *Review of Economics and Statistics* 99(4): 683–697.

Black, Dan, Joonwhi Joo, Robert LaLonde, Jeffrey Smith, and Evan Taylor. 2020. "Simple Tests for Selection: Learning More from Instrumental Variables." HCEO Working Paper No. 2020-048.

Black, Dan, Jeffrey Smith, Mark Berger, and Brett Noel. 2003. "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System." *American Economic Review* 93(4): 1313-1327.

Bloom, Howard. 1984. "Accounting for No-shows in Experimental Evaluation Designs." *Evaluation Review* 82(2): 225-246.

Bloom, Howard, Larry Orr, George Cave, Stephen Bell, and Fred Doolittle. 1993. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. Bethesda: Abt Associates.

Brudevold-Newman, Andrew, Maddalena Honorati, Pamela Jakiela, and Owen Ozier. 2017. "A Firm of One's Own: Experimental Evidence on Credit Constraints and Occupational Choice." World Bank Policy Research Working Paper No. 7977.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Muñoz-Morales, Jeffrey Smith, and Rebecca Thornton. 2022. "Some Children Left Behind: Variation in the Effects of an Educational Intervention." *Journal of Econometrics*, forthcoming.

Buhl-Wiggers, Julie, Jason Kerwin, Jeffrey Smith, and Rebecca Thornton. 2022. "Learning More about Teachers: Value-Added and Treatment Effects on Value-Added in Northern Uganda." Unpublished manuscript, University of Wisconsin-Madison.

Bushway, Shawn and Jeffrey Smith. 2007. "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us." *Journal of Quantitative Criminology* 23(4): 377-387.

Byker, Tanya and Jeffrey Smith. 2021. "Chapter 6: Evidence from Connecticut Jobs First." In Jeffrey Smith, Alexander Whalley, and Nathaniel Wilcox (eds.), *Are Participants Good Evaluators?*" Kalamazoo: W.E. Upjohn Institute for Employment Research. 145-196.

Calónico, Sebastian and Jeffrey Smith. 2021. "Chapter 5: Evidence from the National Supported Work Demonstration." In Jeffrey Smith, Alexander Whalley, and Nathaniel Wilcox (eds.), *Are Participants Good Evaluators?*" Kalamazoo: W.E. Upjohn Institute for Employment Research. 101-144.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." NBER Working Paper No. 24678.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593-2632.

Chung, EunYi and Mauricio Olivares. 2021. "Permutation Test for Heterogeneous Treatment Effects with a Nuisance Parameter." *Journal of Econometrics* 225(2): 148-174.

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

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531-580.

Cunha, Flavio, and James Heckman. 2007. "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling: A Survey of Recent Developments." *Labour Economics* 14(6): 870-893.

Davis, Jonathan, and Sara Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." *American Economic Review* 107(5): 546-550.

Deaton, Angus. 2010. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives* 24(3): 3-16.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2-21.

Dillon, Eleanor and Jeffrey Smith. 2020. "The Consequences of Academic Match between Students and Colleges." *Journal of Human Resources* 55(3): 767-808.

Ding, Peng, Avi Feller, and Luke Miratrix. 2019. "Decomposing Treatment Effect Variation." *Journal of the American Statistical Association* 114(525): 304-317.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Impacts in PROGRESA." *Journal of Econometrics* 145(1-2): 64-80.

Farrell, Max. 2015. "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations." *Journal of Econometrics* 189(1): 1–23.

Ferber, Robert, and Werner Hirsch. 1981. *Social Experimentation and Economic Policy*. Cambridge: Cambridge University Press.

Fisher, Ronald. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Frost, Robert. 1920. "The Road Not Taken." In: Robert Frost (ed.), *Mountain Interval*. New York: Henry Holt.

Hanushek, Eric, and Steven Rivkin. 2012. "The Distribution of Teacher Quality and Implications for Policy." *Annual Review of Economics* 4: 131-157.

Harcourt, Bernard. 2006. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: University of Chicago Press.

Heckman, James, Neil Hohmann, Jeffrey Smith, with the assistance of Michael Khoo. 2000. "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics* 115(2): 651-694.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd.1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017-1098.

Heckman, James and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press for Econometric Society Monograph Series. 156-246.

Heckman, James and Jeffrey Smith. 1998. “Evaluating the Welfare State” in Steiner Strom (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*. Cambridge University Press for Econometric Society Monograph Series, 241-318.

Heckman, James, Jeffrey Smith, with the assistance of 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.

Hollister, Robinson, Peter Kemper, and Rebecca Maynard. 1984. *The National Supported Work Demonstration*. Madison: University of Wisconsin Press.

Horváth, Hedwig, 2015. “Classroom Assignment Policies and Implications for Teacher Value-Added Estimation.” Unpublished manuscript, Institute of Education, University College London.

Hotz, V. Joseph, Guido Imbens, and Julie Mortimer. 2005. “Predicting the Efficacy of Future Training Programs Using Past.” *Journal of Econometrics* 125(1-2): 241-270.

Imbens, Guido, and Joshua Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects” *Econometrica* 62(2): 467-475.

Jacob, Brian A., and Lars Lefgren. 2008. “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education.” *Journal of Labor Economics* 26(1): 101–36.

Klerman, Jacob. 2017. “Editor in Chief’s Comment: External Validity in Systematic Reviews.” *Evaluation Review* 41(5): 391-402.

Klerman, Jacob, Correne Saunders, Emily Dastrup, Zachary Epstein, Douglas Walton, and Tara Adam, with Burt Barnow. 2019. *Evaluation of Impacts of the Reemployment and Eligibility Assessment (REA) Program: Final Report*. Prepared for the U.S. Department of Labor. Cambridge, MA: Abt Associates.

Kline, Patrick and Melissa Tartari. 2016. “Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach.” *American Economic Review* 106(4): 972-1014.

Knaus, Michael. 2021. “A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student's Skills.” *Journal of the Royal Statistical Society: Series A* 184(1): 282-300.

Knaus, Michael, Michael Lechner, and Anthony Strittmatter. 2022. “Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach.” *Journal of Human Resources* 57(2): 597-636.

LaLonde, Robert. 2003. “Employment and Training Programs.” In Robert Moffitt (ed.) *Means-*

Tested Transfer Programs in the United States. Chicago: University of Chicago Press. 517-585.

Lechner, Michael and Jeffrey Smith. 2007. "What is the Value Added by Case Workers?" *Labour Economics* 14(2): 135-151.

Lechner, Michael and Stephan Wiehler. 2011. "Kids or Courses? Gender Differences in the Effects of Active Labor Market Policies." *Journal of Population Economics* 24: 783–812.

Lei, Huitian, Inbal Nahum-Shani, Kevin Lynch, David Oslin, and Susan Murphy. 2012. "A `SMART' Design for Building Individualized Treatment Sequences." *Annual Review of Clinical Psychology* 8: 14-28.

Lim, Katherine and Katherine Michelmore. 2018. "The EITC and Self-Employment Among Married Mothers." *Labour Economics* 55: 98-115.

Linhart, H. and Walter Zucchini. 1986. *Model Selection*. New York, Wiley.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2004. "Equilibrium Policy Experiments and the Evaluation of Social Programs." NBER Working Paper No. 10283.

McCall, Brian, Jeffrey Smith, and Conny Wunsch. 2016. "Government-Sponsored Vocational Training" in Eric Hanushek, Stephen Machin, and Ludger Woessman (eds.), *Handbook of the Economics of Education, Volume 5*. Amsterdam: North-Holland. 479-652.

McKenzie, David. 2018. "Can Business Owners Form Accurate Counterfactuals? Eliciting Treatment and Control Beliefs About Their Outcomes in the Alternative Treatment Status." *Journal of Business & Economic Statistics* 36(4): 714-722.

Michalopoulos, Charles, David Card, Lisa Gennetian, Kristen Harknett, and Philip Robins. 2000. *The Self-Sufficiency Project at 36 Months: Effects of a Financial Work Incentive on Employment and Income*. Ottawa: Social Research and Demonstration Corporation.

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

Muller, Sean. 2015. "Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations." *World Bank Economic Review* 29: S217-S225.

Neyman, Jerzy. 1923. "Statistical Problems in Agricultural Experiments." *Journal of the Royal Statistical Society* 2:107-180.

Orr, Larry, Howard Bloom, Stephen Bell, Fred Doolittle, and Winston Lin. 1996. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute Press.

Philipson, Tomas, and Larry Hedges. 1998. "Subject Evaluation in Social Experiments." *Econometrica* 66(2): 381-408.

Pitt, Mark, Mark Rosenzweig, and Mohammad Hassan. 2012. "Human Capital Investment and the Gender Division of Labor in a Brawn-Based Economy." *American Economic Review* 102(7): 3531-3560.

Quandt, Richard. 1972. "Methods of Estimating Switching Regressions." *Journal of the American Statistical Association* 67: 306-310.

Rothschild, Michael and Lawrence White. 1995. "The Analytics of the Pricing of Higher Education and Other Services in Which the Customers are Inputs." *Journal of Political Economy* 103(3): 573-586.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3: 135-146.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66:688-701.

Sandner, Malte, Thomas Cornelissen, Tanja Jungmann, and Peggy Herrmann. 2018. "Evaluating the Effects of a Targeted Home Visiting Program on Maternal and Child Health Outcomes." *Journal of Health Economics* 58: 269-283.

Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox. 2020. "Are Program Participants Good Evaluators?" IZA Working Paper No. 13584.

VanderWeele, Tyler, and Miguel Hernan. 2013. "Causal Inference under Multiple Versions of Treatment." *Journal of Causal Inference* 1(1): 1-20.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523): 1228-1242.

Waite, Linda, and Maggie Gallagher. 2000. *The Case for Marriage: Why Married People are Happier, Healthier and Better Off Financially*. New York: Doubleday.

Watkins-Hayes, Celeste. 2009. *The New Welfare Bureaucrats: Entanglements of Race, Class, and Policy Reform*. Chicago: University of Chicago Press.

Weiss, Michael, Howard Bloom, and Thomas Brock. 2014. "A Conceptual Framework for Studying the Sources of Variation in Program Effects." *Journal of Policy Analysis and Management* 33(3): 778-808.

Wiswall, Matthew. 2013. "The Dynamics of Teacher Quality." *Journal of Public Economics* 100: 61–78.

Wood, Robert, Quinn Moore, Andrew Clarkwest, and Alexandra Killewald. 2014. "The Long-Term Effects of Building Strong Families: A Program for Unmarried Parents." *Journal of Marriage and Family* 76(2): 446-463.