Instrumental variables: From structural equation models to design-based causal inference*

Christopher L. Carter† and Thad Dunning‡

This draft: July 1, 2019

Instrumental-variables (IV) analysis bridges structural equation modeling and design-based methods for causal inference. In its original formulation, instrumental variables were designed to help overcome the endogeneity of price and quantity to derive supply and demand curves; by finding a third variable that was correlated with the supply but not the demand of a good (or vice versa), scholars sought to map, in a system of structural equations, how supply and demand respond to changes in prices. Later, social scientists would use insights from IV to improve analyses of experiments, in particular, to estimate causal effects for those units who comply with their assignment to a particular treatment status. Thus, the use of instrumental variables also spans observational and experimental research.

In both kinds of applications, instrumental variables appeals to an assumption of random or as-if random assignment of units to causal conditions. In structural equation modeling, randomization implies statistical independence of the causal variable(s) and the error term in a regression model—that is, exogeneity. In the observational world, such assignment occurs naturally, or is otherwise out of the control of the researcher—often thereby raising concerns about whether assignment is really as good as random. When plausible, however, this assumption allows researchers to obviate concerns about confounding variables that complicate drawing causal inferences from observational data. IV thus promises to marry the realism and macro focus of observational research to the rigor of experimental methods.

Yet, in observational and experimental work alike, (as-if) random assignment alone does not guarantee valid causal inference under the IV framework. Other assumptions are also needed, and these often cannot be fully tested from the data. Furthermore, the assumptions invoked by structural equation models that are fit to observational data (e.g., supply and demand curves) carry different weight and meaning, as compared to those required to estimate complier average causal effects in an experiment.

We draw attention to these points of overlap and divergence between different usages of instrumental variables in this chapter. We begin with a discussion of early IV work in

*Prepared for inclusion in the *SAGE Handbook of Research Methods in Political Science & International Relations*, edited by Luigi Curini (Università degli Studi di Milano) & Robert J. Franzese, Jr. (University of Michigan).

†University of California, Berkeley, email: christopher.carter@berkeley.edu

‡University of California, Berkeley, email: thad.dunning@berkeley.edu

the structural equation modeling (SEM) framework, highlighting the key assumptions and potential places where they may break down. We then discuss applications of IV to design-based research under a potential outcomes model. We detail similarities and differences in the assumptions that these two types of applications entail. A key distinction involves the stipulation of a linear response schedule with constant effects across units in the SEM framework. Relaxing this assumption in the potential outcomes framework allows for clear definition of heterogeneous unit-level causal effects, which proves particularly important in experiments with non-compliance. Moreover, the potential outcomes framework disaggregates and clarifies other key assumptions often left implicit in the SEM framework. Yet, both approaches face important challenges in generalizing effects beyond the variation induced by a particular instrument. In a final section, we illustrate these points by comparing two different instrumental-variables strategies—one observational and the other experimental—for investigating the effect of price changes on demand for coffee.

1 IV in structural equation models

The use of instrumental variables originated in simultaneous equation models, in which researchers sought to estimate supply and demand curves from equilibrium values of price and quantity (Angrist and Krueger 2001; Stock and Trebbi 2003: 179). Because supply and demand curves map how quantity supplied and demanded responds to changes in prices, they can be considered "response schedules" or "structural equations," where the regression of quantity, Q, on price, P, carries a causal interpretation (Freedman 2009; Imbens 2014: 9).[1] A researcher may stipulate, for instance, that demand is determined according to

$$Q_t = \beta_0 + \beta_1 P_t + \beta_3 X_t + \gamma_t, \tag{1}$$

where Q_t is the quantity of a product demanded at time t, P_t is its price, X_t is a matrix of exogenous variables affecting demand, and γ_t is a random error (disturbance) term.

A difficulty for estimating equation (1), however, is that the quantity of the good supplied is also a function of P_t. Suppose the supply curve is given by

$$Q_t = \beta_4 + \beta_5 P_t + \beta_6 Z_t + \gamma_t, \tag{2}$$

where Z_t is a matrix of variables affecting supply.[2] Were the supply curve to remain fixed while the demand curve shifted, data on equilibrium levels of price and output could allow a researcher to trace out the demand equation. Yet, both curves may shift as a function of shared market conditions. In an early analysis of the impact of tariffs in markets for butter and flaxseed, the mathematician and economist Philip G. Wright noted this problem: "If both supply and demand conditions change, price-output data yield no information as to either curve. Unfortunately ... [this case] is the more common" (Wright 1928: 296). Indeed,

[1] We refer in this paper to "structural equation models" in this sense. One stream of research uses the term more specifically to refer to systems of equations linking unobservable "latent" constructs; see e.g. Bollen (1989).

[2] We use the typical language of supply and demand "curves" here, even though the response schedules in equations (1) and (2) are linear in P_t.

if $X_t = Z_t$ in equations (1) and (2)—that is, the same variables affect the quantity of the good demanded and supplied—then data on quantities and prices cannot uniquely identify the supply and demand curves.

Wright (1928) proposed an initial solution to this problem by using variables that affected supply without independently shaping demand (and vice-versa).[3] When such variables can be found, the columns of the matrix X_t in equation (1) are not identical to the columns of Z_t in equation (2). Using what came to be called "instrumental variables"—that is, variables in X_t that are excluded from Z_t, and vice versa—Wright determined the elasticity of the supply (and demand) functions of flaxseed. One instrument Wright used to estimate supply elasticity was the price of a flaxseed substitute, cottonseed. This example already suggests difficulties in finding viable instrumental variables, however: shocks to substitutes might affect not only the demand for, but also the supply of flaxseed, perhaps because producers anticipate shifts in the demand curve.

Estimating equations (1) and (2) raises related difficulties. Manipulation of the price of the good affects quantity in both equations: supply and demand are jointly determined within a system of structural equations. Moreover, unmeasured variables that affect the quantity of demand may also affect supply, resulting in endogeneity—that is, correlation between disturbances and an explanatory variable (Freedman 2006: 699; Freedman 2009; Imbens 2014: 9). In that case, the Ordinary Least Squares (OLS) estimate of β_1 in Equation 1 is biased by $(P'P)^{-1}P'E(\gamma|P)$, when $E(\gamma|P) \neq 0$ (Freedman 2009: 181). Yet, as long as cottonseed is correlated with flaxseed price but uncorrelated with the disturbance term from the demand equation, instrumental-variables analysis can provide a consistent estimator of demand elasticity (Angrist and Krueger 2001: 70).

Wright's work went largely unnoticed and played little role in the development of the IV method in econometrics.[4] In fact, there was no further work on instrumental variables until the 1940s, when Reirsøl's (1945) dissertation demonstrated that model parameters can be identified using the additional information provided by an "instrumental set of variables" (Angrist and Krueger 2001; Morgan 1990; Aldrich 1993). Building on the further work of Geary (1949) and Durbin (1954), Sargan (1958) demonstrated the consistency of the instrumental variables estimator. Wald (1940) had previously shown the consistency of an equivalent "grouping" estimator.[5]

While much of this early research sought to address measurement error in independent variables, the IV framework has gained its most prominent use in addressing the problem of omitted variable bias. A researcher interested in a causal effect of an explanatory variable X_i on an outcome variable Y_i may stipulate the response schedule,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{3}$$

If unmeasured variables not included in equation (3) are correlated with the explanatory

[3] While Philip G. Wright's name is on the piece, the key finding is in Appendix B, which is believed by some to have been written by his son, Sewall. The elder Wright also discusses his son's closely related methods of causal path analysis. However, Stock and Trebbi (2003), using stylometric analysis, conclude that Philip was the most likely author.

[4] Wright's innovation was only recognized in the 1970s, when Goldberger (1972) highlighted Wright's contribution to structural equation methods (Aldrich 1993: 270, fn. 34).

[5] We show the equivalence of the IV estimator and Wald's grouping estimator below.

variable, such that ϵ_i and X_i are statistically dependent, OLS will yield biased and inconsistent estimates of β_1. However, a third variable, Z_i that is correlated with X_i but not ϵ_i offers a way to identify β_1. Specifically, consistent estimation of β_1 can be obtained from the "first stage" regression of X_i on Z_i and then a second-stage regression of Y_i on the fitted values of X_i, or $\widehat{X_i}$, from the first stage.[6] In matrix notation, this "two-stage least squares" (2SLS) estimation of β_1 can be written as

$$\begin{aligned} \widehat{\beta}_{1,\text{2SLS}} &= (\widehat{X}'\hat{X})^{-1}\widehat{X}'Y, \\ \text{where } \widehat{X} &= Z(Z'Z)^{-1}Z'X. \end{aligned} \tag{4}$$

Other derivations of the multivariate instrumental-variables least squares (IVLS) estimator can be rearranged to show their equivalence with equation (4) (Freedman 2009: 178-9). In the bivariate model in equation (3), equation (4) is equivalent to dividing the regression coefficient of the "reduced form" regression of Y_i on a variable Z_i by the regression coefficient obtained in the first-stage regression of X_i on Z_i:[7]

$$\widehat{\beta}_{1,\text{2SLS}} = \frac{\frac{\widehat{\text{Cov}}(Y_i,Z_i)}{\widehat{\text{Var}}(Z_i)}}{\frac{\widehat{\text{Cov}}(X_i,Z_i)}{\widehat{\text{Var}}(Z_i)}} = \frac{\widehat{\text{Cov}}(Y_i,Z_i)}{\widehat{\text{Cov}}(X_i,Z_i)}. \tag{5}$$

We return to equation (5) in the next section on IV analysis in design-based inference, where the "reduced form" regression of Y_i on Z_i is referred to as "intent-to-treat" analysis.

Instrumental-variables analysis requires several crucial assumptions for consistent estimation of β_1 by the method in (4) or (5). Some of these assumptions are mechanical, meaning that the calculation of the 2SLS estimator requires them to be true. To solve Equation (4), the number of units must be at least as large as the number of independent variables (i.e., $n > q \geq p$ where n is the number of observations, q is the number of columns in Z, and p is the number of columns in X); and $Z'X$ and $Z'Z$ must have full rank of p and q respectively.[8] Additionally, we require in practice a sufficiently strong covariance between X_i and Z_i: the so-called "weak instrument" problem exacerbates finite-sample bias in the IV estimator.[9] Indeed, in finite samples with instruments that are only weakly related to the endogenous regressors, the asymptotic unbiasedness of the 2SLS estimator in a hypothetical, infinitely large sample—i.e., its consistency—may be of limited practical utility (Staiger and Stock 1997). We can diagnose weak instruments by examining the relationship between X_i and Z_i; as a rule of thumb, F-statistics of less than 10 indicate a weak instrument (Staiger and Stock 1997). Each of these assumptions may be evaluated from the data.

However, other key assumptions of structural equation models are more difficult or impossible to test. Each assumption merits careful consideration in applications of the

[6] The fitted values $\widehat{X_i}$ are sometimes called "predicted" values of X_i, though "post-dicted" is usually more accurate. Importantly, we cannot simply use the values of $\widehat{X}$ to calculate the variance-covariance matrix of $\widehat{\beta_1}$, as this produces inconsistent estimation of σ^2 (Greene 2003: 79).

[7] A derivation can be found in the Appendix; see (A.1).

[8] A model in which $q = p$ is "just-identified" while the case with $q > p$ is "over-identified."

[9] There is, of course, also a mechanical reason for this, related to the previous paragraph and the rank condition; if $Cov(X_i, Z_i) = 0$, then the estimator of β_1 found in Equation (5) is undefined.

method. First and perhaps most fundamentally, valid IV analysis of structural equation models requires that the data were generated according to the posited response schedule—that is, a regression model such as equation (3):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Because the model stipulates the effect of hypothetical interventions to alter values of X_i, β_1 is said to carry a causal interpretation: it is the causal effect of X_i on Y_i. However, in observational studies—by definition—no researcher intervened in the system to manipulate the value of X_i (Freedman 2009). Whether the model captures what would happen if, say, a researcher varied X_i experimentally is usually a matter of conjecture. We return to this idea of invariance to manipulation in our discussion of design-based inference.

The stipulation of this model embeds several auxiliary postulates, with specific implications for IV estimation. First—as often noted in methodological discussion of IV analysis, and as sometimes discussed in applications as well—the assumption that the response schedule is correctly specified implies an "exclusion restriction." That is, the instrument is *excluded* from equation (3). Thus, Z_i does not have a "direct" effect on Y_i: it does not itself belong in the response schedule (Equation 3), and if it is related to Y_i, it is only through its effect on X_i. We refer to this assumption as the "exclusion restriction," although some scholars use this term to refer to the combination of this assumption and the independence of Z_i and ϵ_i; we treat the latter as a distinct assumption (Angrist and Pischke 2008: 117).[10] Additional collection of qualitative and quantitative data can help to rule out plausible alternative channels through which Z_i might have a direct effect on Y_i. Yet, for reasons we discuss further below, convincingly demonstrating that the instrument only affects the outcome through the endogenous regressor of interest raises considerable difficulties.

In addition, the structural model critically implies a set of linearity and constancy assumptions. Equation (3) stipulates that the response schedule is linear in the parameter β_1: thus, the effect is proportional to the value of X_i. In addition, for each unit i, the response Y_i only depends on value of the regressor X_i: the exposure to this treatment of other units $j \neq i$ is irrelevant. This is an analogue to the "non-interference" assumption, a component of the "stable-unit treatment value assumption" (SUTVA), in the context of design-based inference under the potential outcomes model. Thus, for each unit i, the treatment effect is constant, in the sense that it does not depend on the treatment assignment of other units, which might be compromised if by communication or learning from other subjects in a study pool. The response schedule also presumes a treatment effect that is constant *across* all units i: β_1 is the same for every unit in the study. We further discuss these assumptions of linear and constant effects across units (and contrast it to the assumption of idiosyncratic unit effects in the potential outcomes framework) in the next section.[11]

Finally, models such as equation (3) assume a different kind of constancy assumption: effects are constant (or homogenous) across components of X. This assumption that has received somewhat less attention, yet is critical for understanding the leverage that IV analysis

[10] Given a model like equation (3), the unconfoundedness of the instrument and the exclusion restriction are implied by $\epsilon_i \perp\!\!\!\perp Z_i$ (Imbens 2014).

[11] Heckman and Robb (1986); Imbens and Angrist (1994); Angrist et al. (1996); Rosenzweig and Wolpin (2000); Freedman (2006); and Heckman et al. (2006) all draw attention to this IV assumption.

may—or may not—provide. Imagine a researcher who is interested in the effect of income (X_i) on attitudes toward taxation (Y_i). Among participants in a lottery, lottery winnings (Z_i) can be used as an instrument for income. Income (X_i) is the sum of winnings from the lottery and income from other sources ("earned income"); call these X_{1i} and X_{2i}. The model in equation 3 assumes that the effect of these two components of X_i are the same. If this is not the case, then in calculating a 2SLS estimate of β_1 we are getting the effect of a particular type of income shock—specifically, windfall gains X_{1i} (Dunning 2008). We are not getting the effect of an increase to earned income X_{2i}. Perhaps, then, the model we should be considering is, in fact, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$. If $\beta_1 \neq \beta_2$ in this equation, then assuming a constant effect of β_1 in equation 3 is misleading. But we cannot estimate the model with X_{1i} and X_{2i} without another instrument for X_{2i}. Were we to find such an instrument, concerns about the assumption that β_2 is constant across components of X_{2i} may arise. Dunning (2008) gives additional examples where the possibility of heterogeneous partial effects in IV models raises concerns about model specification. One way to reduce concerns of heterogeneous partial effects may be to define concepts more precisely a priori and limit causal claims to those aspects of a general concept that are actually measured through the IV analysis.[12] The point is that the stipulation of the response schedule is a key consideration for IV analysis.

If the many modeling assumptions hold—but X_i is endogenous, or statistically dependent on ϵ_i—and there exists an instrumental variable such that

$$Z_i \perp\!\!\!\perp \epsilon_i, \tag{6}$$

where $\perp\!\!\!\perp$ is read as "is independent of," then the instrumental-variables estimator in equation (5) consistently estimates β_1. This too is a matter of model specification: like the exclusion restriction, statistical independence of the instrument and the disturbance term implies that Z_i does not belong in the response schedule. If it did, then the response schedule in Equation (3) would be incorrectly specified. Given the model, however, random assignment of values of the instrument may imply $Z_i \perp\!\!\!\perp \epsilon_i$. The assumption could also hold in a natural experiment where treatment is merely "as-if" randomly assigned. Yet, the burden is then on the researcher to demonstrate why Z_i might be plausibly uncorrelated in expectation with pre-treatment causes of X_i and Y_i. Sovey and Green (2011) and Dunning (2012), among others, discuss tests that can be used to assess the validity of this assumption.

Relative to the design-based approach discussed next, the SEM framework adds additional complications for assessing the assumption in (6), however. Researchers often use multivariate IVLS regression—thus, the matrix form of the estimator in equation (4). In practice, however, they tend to focus on a single endogenous regressor and on whether a single instrumental variable is as good as randomly assigned; and they include putatively "exogenous" columns of the matrix X in the matrix of independent variables, Z. Little attention is typically paid to assessing the assumption that those other columns of Z are exogenous—that is, as good as randomly assigned—as required for valid estimation by the 2SLS estimator for multiple regression. We return later to discussing these assumptions,

[12] To be sure, improving conceptual precision by moving down Sartori's (1970) "ladder of abstraction" may lessen the perceived impact of the research: a paper on the effects of windfall earnings on political attitudes may generate less interest than one that purports to estimate the effect of income more generally.

after introducing the use of instrumental-variables analysis in design-based analysis under the potential outcomes framework.

.

2 IV analysis in design-based causal inference

The rise of experimental social science has provided a new use for instrumental variables as a tool for estimating a complier average causal effect (CACE). When conducting an experiment, researchers randomly assign units to treatment or control conditions. Interest is often in estimating the average causal effect (ACE) for the study group of units—that is, the experimental population. Scholars often stipulate the Neyman potential outcomes model, also called the Neyman-Rubin-Holland model (Splawa-Neyman et al. 1990; Rubin 1974; Holland 1986). According to this model, each unit has a potential outcome under treatment, $Y_i(1)$—i.e., the outcome that would materialize if it were assigned to the treatment group—and another potential outcome $Y_i(0)$ that would materialize if it were assigned to control. The two potential outcomes cannot be simultaneously observed for the same unit, because a unit assigned to the treatment group cannot be assigned to control; this is the "fundamental problem of causal inference" (Holland 1986). Nor can a researcher observe the average of the potential outcomess under treatment for the experimental population, without losing access to the average of the potential outcomes under control. In an experiment, however, units are assigned at random to treatment and control groups. It is as if the treatment group is a random sample from the experimental population; and the control group is another random sample from the same population. The mean of the treatment sample can therefore be used to estimate the average potential outcome under treatment, for all units in the study group; and the mean of the control sample similarly estimates the average potential outcome under control. The difference of the means is an unbiased estimator for the average causal effect.

This mode of inference is sometimes called "design-based," because the only stochastic element in the model is the random assignment to treatment and control groups—which is controlled by the researcher as a matter of research design (Cox 2009).[13] Scholars have also used the term more broadly to denote strategies for controlling for confounding variables that depend centrally on research design—rather than on regression adjustment, as in standard SEM frameworks (Freedman 2009; Dunning 2012). "Design-based" approaches are thus sometimes contrasted with "model-based" research, even though models for causal and statistical inference play a central role in both. The key difference, as we discuss in the next section, concerns the nature of the assumptions that must be made.

In design-based inference in experiments, the CACE—and instrumental-variables analysis—enters the picture when some units, despite having been assigned to the treatment condition, do not actually receive the treatment. Differential take-up of treatment generates a problem of non-compliance with treatment assignment. Imagine a case where a government offers a temporary employment program to unemployed citizens; many citizens apply, far more than the program can fund. The government decides to use a lottery to decide which applicants may participate. However, not all of those selected ultimately participate. Some have already

[13] This usage of "design-based" in statistics differs from a related but distinct use of the term in educational research.

located other employment; others may have already migrated elsewhere in search of employment; and still others may simply lose interest in participating. Similarly, some of those who were not offered enrollment may ultimately participate, say, if there is is non-take-up by those originally selected to participate.

With non-compliance, the difference-of-means estimator is "intent-to-treat" (ITT) analysis: it measures the effect of assignment to the program.[14] The effect of treatment assignment on outcomes such as future employment, or political support for the incumbent, may be of substantial policy as well as scholarly interest. Estimating it could tell us, for example, the likely marginal returns of offering the program to additional participants. Still, the estimator does not readily measure the effect of treatment receipt, i.e., actual participation in the TEP. The assigned-to-treatment and assigned-to-control groups include non-compliers; this may "dilute" the effect of treatment assignment. How to estimate the effect of program participation is not immediately obvious, however. We cannot naively compare those who received treatment to those who did not: those are self-selected groups, and participators may differ from non-participators in ways other than exposure to treatment. Put differently, these self-selected groups contain distinct mixes of compliers and non-compliers, and that asymmetry may confound valid inference about the effect of treatment receipt.[15]

Instrumental-variables analysis can assist in the estimation of an average causal effect among compliers—the CACE. To do so, we extend the potential outcomes model to allow for non-compliance.[16] Thus, we imagine that there are three types of subjects in the study pool: compliers, always-takers, and never-takers. Under the model, these types are fixed at the level of the subject; type is not affected by the assignment to levels of treatment. Compliers are those units who would receive the treatment if assigned to the treatment group—but otherwise receive the control. Always-takers receive the treatment, and never-takers receive the control, regardless of their assignment. A fourth type, defiers—who receive the treatment if assigned to the control group but receive the control if assigned to treatment—are ruled out; this assumption is required for identification of the CACE (Freedman 2006).[17] The trick is then to separate the responses of compliers, always-takers, and never-takers—in order to isolate the effect of treatment assignment among compliers. At the unit level, we often cannot directly observe who is a complier and who is not, as these definitions involve counterfactuals—that is, potential outcomes (Imbens 2014). For example, among those assigned to the control group who actually receive the control protocol, we do not observe whether they would have taken the treatment, had they been assigned to the treatment group.

However, we can estimate the group-level distribution of compliance types—and the average responses by type. Imagine first that there are no always-takers: this is a situation

[14] As we discuss below, the ITT analysis is equal to the reduced-form estimate discussed above.

[15] Relatedly, while manipulation checks can provide a useful measure of whether subjects understood or experienced the treatment in the way the researcher expected, treatment effects should not be calculated conditional on having passed a manipulation check, as the check is necessarily post-treatment (Aronow et al. 2015; Montgomery et al. 2018).

[16] See, inter alia, Angrist et al. (1996); Freedman (2006); Gerber and Green (2012); Dunning (2012); Imbens (2014).

[17] Angrist et al. (1996) call the no-defiers assumption "monotonicity": being assigned to treatment should never make it *less* likely that a unit actually receives treatment (see also Imbens 2014: 17).

of "single crossover" or "one-way non-compliance" (Gerber and Green 2012). In this case, we can tell which type is which among units assigned to the treatment group: the never-takers cross over to receive the control protocol, while the compliers receive treatment.[18] Thus, we observe the average responses of the group of compliers in the assigned-to-treatment group. In the assigned-to-control group, however, the compliers and never-takers look the same: they both follow the control-group protocol.

Nonetheless, due to random assignment, we can estimate the proportion of each type in the study group. Indeed, the proportion of each type in the assigned-to-treatment group is an unbiased estimator for the corresponding proportions in the experimental population, since the treatment group is a random sample from the whole set of units in the experiment. In particular, the fraction of compliers in the treatment group—which we can observe in the case of single crossover from treatment to control—estimates the fraction of compliers in the experimental population. Moreover, the responses of never-takers in the treatment and control groups should be the same, in expectation: by assumption, treatment assignment has no effect on the response of never-takers, since they actually receive the control condition whether they are assigned to the treatment or the control group. Since we observe the overall response in the assigned-to-control group, and we impute the response of never-takers from the assigned-to-treatment group, we can therefore estimate the responses of the compliers in the control group. The assumption that treatment assignment does not affect the response of never-takers is akin to the "exclusion restriction" in the SEM framework, as we discuss further in the next section, though the potential outcomes framework helpfully clarifies the important distinction between the exclusion restriction and as-if random assignment.

Together, random assignment and the exclusion restriction therefore allow us to estimate the responses of the never-takers in the assigned-to-control group—and thus the compliers. An estimate of the CACE is just the average difference between the assigned-to-treatment and the assigned-to-control groups—i.e., what ITT analysis gives us—divided by the estimated proportion of compliers in the study group. The model can also be extended to the case of two-sided non-compliance (double crossover).[19] In this case, we estimate the proportion of compliers by subtracting the proportion of the assigned-to-control group that actually receives the treatment from the proportion of the assigned-to-treatment group that receives treatment. Thus, when treatment assignment is a binary variable (e.g., $X_i = 1$ when assigned to treatment, $X_i = 0$ when assigned to control), we can use the "Wald estimator,"

$$\hat{\beta}_{1,Wald} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{X}_1 - \overline{X}_0}, \tag{7}$$

where $\overline{Y}_1$ is the sample average in the assigned-to-treatment group, $\overline{Y}_0$ is the sample average in the assigned-to-control group, $\overline{X}_1$ is the proportion who receive treatment in the assigned-to-treatment group, and $\overline{X}_0$ is the proportion who receive treatment in the assigned-to-control group. The difference of means in the numerator of (7) is thus "intention to treat" analysis: it estimates the average causal effect of treatment assignment.[20] Note that $\overline{X}_1$

[18] This also assumes we can observe who actually takes the treatment, e.g., who follows the protocol in a drug trial (which is distinct from the even harder problem of observing counterfactual compliance types).
[19] See Freedman (2006); Dunning (2012); Gerber and Green (2012).
[20] Numerically, the value is equivalent to the reduced-form regression of Y on Z.

includes both compliers and always-takers, while $\overline{X}_0$ includes only always-takers. Effectively, the denominator subtracts off the proportion of always-takers in the control group from the joint proportion of compliers and always-takers in the treatment group. Because we expect the estimated proportion of compliers to be the same across the groups assigned to treatment and control (due to random assignment), the denominator of the Wald estimator estimates the proportion of compliers in the full study group.[21]

Why is equation (7) an instrumental-variables estimator? Numerically, it is equivalent to a 2SLS procedure in which we regress Y on the fitted values of X, which were obtained from a first-stage regression of X on Z. Indeed, with one treatment and one control group, and as we show in the appendix,

$$\hat{\beta}_{1,Wald} = \hat{\beta}_{1,\text{2SLS}}, \tag{8}$$

where $\hat{\beta}_{1,\text{2SLS}}$ is given by equation (5). Conceptually, treatment assignment serves as an instrumental variable for treatment receipt in a similar sense to that developed in the previous section: it is correlated with an endogenous variable (treatment receipt), is randomly assigned, and by assumption does not directly influence outcomes, other than through its influence on treatment receipt. Indeed, the proof that $\hat{\beta}_{1,Wald}$ is a consistent estimator for the CACE depends on both the randomization of the instrument and the stipulation that treatment assignment does not affect the responses of always-takers and never-takers—a kind of exclusion restriction.[22]

The beauty of the Wald estimator lies in its simplicity. If we have two potential treatment assignments, we can calculate an estimate for the complier average causal effect knowing only the first-stage difference in means (the numerator) and the estimated proportion of compliers in the study group (the denominator). This simplicity also rests on a model that seems reliably to capture core elements of the data-generating process: in any experiment, for example, the physical properties of random assignment of units to treatment and control groups seem to justify the metaphor of drawing potential outcomes at random from an urn.

Nonetheless, as in all causal and statistical inference—and certainly as also in the SEM framework—design-based analysis under the potential outcomes model involves maintained hypotheses. A key assumption is the response schedule itself. The Neyman model assumes each unit has potential outcomes—in its simplest formulation, a potential outcome under control and a potential outcome under treatment. While potential outcomes are free to vary across units, they are considered fixed, deterministic properties of each unit; and the treatment assignment of one unit does not affect the response of another. When extended to account for non-compliance, moreover, the model assumes that units are always-takers, never-takers, or compliers—but not defiers. And assignment to treatment only affects outcomes for compliers; the response of always-takers and never-takers is invariant to treatment assignment. As in the SEM framework, such modeling assumptions merit careful consideration in applications of the design-based approach.

[21] This is equivalent algebraically to the first-stage regression of X on Z.

[22] Note that Equation (7) suffers from ratio-estimator bias: the denominator is a random variable. However, by Slutsky's theorem, the estimator is consistent (asymptotically unbiased); see Freedman (2006) or Dunning (2012).

SEM vs. design-based IV: A comparison of assumptions

How, then, does the use of instrumental variables in the SEM framework compare to design-based approaches? The discussion so far highlights points of convergence—but also important areas of divergence. We detail similarities and contrasts in the core assumptions of the models in Table 1. Each row or set of rows in the table includes an assumption in the SEM framework (first column) and a corresponding or contrasting analogous assumption in the design-based approach (second column). Fundamental distinctions in the approaches involve the assumed response schedule; the population for which key estimands are defined; stipulations on stochastic process; and the manner of formulating validity conditions on instruments. In boldface, we indicate those assumptions that can be assessed, if at least partially, with data; we later explain the coding decisions.

First, for SEMs, the response schedule is a linear equation such as equation (3). Thus, the effect of X_i on Y_i is given by the constant of proportionality β. Linear structural equation models involve assumptions akin to potential outcomes, because the response schedule traces out counterfactual responses at different values of X_i (Freedman 2009). Yet, the levels of X_i (or Z_i) are not typically directly manipulated, and linearity implies that the response surface varies smoothly as a function of X_i. By contrast, models in the Neyman tradition stipulate unit-level potential responses to two or several categorical treatment conditions. In experiments, assignment to these conditions is directly manipulated by a researcher.

Second, the response schedule under SEM also implies an assumption of a constant effect across units. By contrast, the design-based approach explicitly allows for treatment effects to vary across compliance types. This assumed unit-level heterogeneity of effects is useful because it readily illuminates key assumptions—for example, the idea of monotonicity, discussed momentarily—which are otherwise buried in the stipulation of common effects across units in the SEM framework (Imbens 2014: 346). It also allows easy characterization of varying effects for specific sub-groups, including "local average treatment effects" (LATE) such as the CACE. This complier average causal effect is less readily characterized in an SEM model in which effects are presumed constant across units (Sovey and Green 2011).

One apparent point of convergence among the SEM and design-based approaches is that both appear to stipulate constant effects across components of a treatment. Yet, the issues this raises, as we discuss later, appear less troublesome for the design-based approach than in structural equation modeling: that effects are constant across "components" of treatment assignment in an experiment seems weaker and more plausible, compared to, say, the assumption that different types of income have the same effect on attitudes (Dunning 2008).

Next, the two approaches further imply different assumptions about the population for which key estimands are defined. In the design-based approach, the target of inference is clear: it is the average of potential outcomes under treatment and control (and their difference), for the set of units in the study group—-also known as the "experimental population." Since this study group is typically (though not always) a convenience sample, there need be no broader population to which formal statistical inferences are drawn: the ACE is defined for the experimental population (and effects for sub-groups, such as the CACE,

Table 1: SEM vs. design-based IV: A comparison of the assumptions

Structural Equation Modeling		**Design-based IV**	
Linear response schedule	Linearity in parameters	Neyman potential outcomes	Unit-level potential responses to categorical treatment conditions
	Constant effect across units		Varying effect across units
	Constant effect across treatment components		Constant effect across treatment components
	Infinite (or undefined) population		**Finite experimental population**
	Random disturbance term (i.i.d)		Random sampling of potential outcomes (not i.i.d)
	Y_i depends on X_i, not $X_{j \neq i}$		**Non-interference/SUTVA**
	Z_i does not belong in response schedule (i.e., $Z_i \perp\!\!\!\perp \epsilon_i$ and exclusion restriction)		**(As-if) random treatment assignment**
			Exclusion restriction
	Rank assumptions		No defiers (monotonicity)
	Strength of Instrument		**Strength of Instrument**

Note: Bolded assumptions indicate those that can be potentially or partially tested from the data.

are similarly defined in reference to compliers in the study group).[23] Thus, in statistical treatments, the design-based approach is sometimes known as "finite population" analysis. We code the presence of a finite experimental population as testable in Table 1, but indeed this is directly observable. This clarity on the target population is not always present in the SEM approach. To be sure, an equation such as (3) will be fit to data for a particular group of units; but the equation aspires to a level of generality that does not appear restricted to a particular set of data. This impression is heightened by assumptions on stochastic process. In the SEM framework, "Nature" draws random disturbance terms, ϵ_i in equation (3); in a classical regression model, these are independent and identically distributed (i.i.d). But how "Nature" draws error terms at random and with replacement, and from what broader population, is not clearly articulated. In contrast, the design-based approach assumes that potential outcomes are fixed in the particular study group at hand. Randomness enters only in the metaphor of sampling potential outcomes from an urn—i.e., in sampling from this experimental population. Thus, random assignment to treatment or control groups determines which potential outcomes are observed. Moreover, these draws from the urn are not generally i.i.d.: they are made without replacement, and the treatment and control samples are statistically dependent.

Both approaches require that the outcome for a given individual depends only on whether that individual received treatment—and not on the assignment of other units. Thus, under an SEM such as equation (3), Y_i depends only on X_i and not on any other unit's value of the endogenous regressor, X_j. The design-based framework makes an analogous stipulation: a unit's potential outcomes are fixed and do not depend on the treatment receipt of any other unit. This is "non-interference," or a component of what Rubin (1978) calls the Stable Unit Treatment Value Assumption (SUTVA). In the design-based framework, for example, a common concern is that units that were assigned to receive the treatment may talk with or otherwise affect units that were assigned to control. The stipulation of non-interference can be seen as an identifying restriction: if potential outcomes depend only on a given unit's treatment receipt, but also on the treatment receipt of other units, the number of parameters (potential outcomes) in the model multiplies quickly, and this increases the difficulty of identifying key causal parameters of interest. However, unlike in the SEM tradition, manipulation of an experimental design can provide the means to test the existence of such spillovers between treatment and control groups. For example, a researcher might assign clusters of households to the treatment and control group; but then further assign individuals at random to treatment and control within the treatment households. Comparison of the responses of control individuals in the treatment and control households allows assessment of the presence of spillovers; see, e.g., Nickerson (2008). For this reason, in Table 1, we code the non-interference assumption as potentially testable in the design-based approach.

Next, the two approaches differ in their approach to key validity assumptions on the instrument. The SEM framework assumes that the instrument Z_i does not belong in the response schedule given by equation (3). This, in turn, implies both Z_i independent of ϵ_i—

[23] Some work does nonetheless distinguish, not always with clarity, between a sample average treatment effect (SATE) and a population average treatment effect (PATE), where the study group is itself viewed as a sample from a broader population.

as secured by randomization of the instrument—and what we call the exclusion restriction. Yet SEM does not clearly separate these two assumptions (Imbens 2014: 346), while the design-based IV approach treats each assumption as distinct. An instrument needs to be, on the one hand, (as-if) randomly assigned, which allows for a causal interpretation of the first-stage regression of Y_i on Z_i (i.e., the ITT) (Angrist and Pischke 2008: 152-153). The second assumption requires that the instrument only affect the outcome through the endogenous regressor. This exclusion restriction implies that potential outcomes for a given level of X_i do not changed based on the value of Z_i.[24]. Angrist and Pischke (2008: 153) use the example of the Vietnam draft lottery, a "natural" experiment, to illustrate why these two validity assumptions should be treated as distinct. To serve in Vietnam, young men were randomly assigned a number based on their birthday; lower numbers were selected first to serve. The random assignment of draft order fulfills the first validity assumption (i.e., statistical independence of Z_i on ϵ_i). Yet, being assigned a low draft number might affect the outcome (i.e., future earnings) not only through the endogenous regressor of interest (i.e., higher probability of military service), but also through other channels (e.g., enrolling in a university in hopes of getting a deferment). Compared to the stipulation that $Z_i \perp\!\!\!\perp \epsilon_i$ in SEMs, the assumption of as-if random assignment in the design-based approach can be directly if only partially tested. In addition to a priori knowledge or theory about the randomization process, this assumption can be assessed using balance and placebo tests, which answer the question of whether the data are consistent with randomization to treatment conditions. By contrast, in neither approach (SEM or design-based) can the exclusion restriction be directly assessed.

Note, then, that none of the assumptions of the SEM framework discussed so far can be readily tested from data. Others, however, must be true in order to calculate the 2SLS estimator. We refer to these as "rank assumptions" in Table 1. For example, the number of units, n, must exceed the number of instruments, q, which must also exceed the number of endogenous covariates, p; also, the matrices $Z'X$ and $Z'Z$ must be full rank, p and q, respectively. In this first section, we referred to these assumptions as "mechanical": given particular matrices X and Z, they can be readily tested. We therefore put this item in boldface in Table 1. More deeply, however, the rank of the matrices also reflects substantive modeling decisions—such as the exclusion of covariates which might otherwise be included in X or Z but cannot because if they were, the number of independent variables would outstrip the number of observations. Thus, identification is accomplished through model specification. As Freedman (2009: 144) puts it, "Many statisticians frown on under-identified models: if a parameter is not identifiable, two or more values are indistinguishable, no matter how much data you have. On the other hand, most applied problems *are* under identified. Identification is achieved only by imposing somewhat arbitrary assumptions."

In the design-based approach, the no-defiers (monotonicity) assumption can similarly be seen as an identification restriction. Defiers are those who receive the opposite treatment from the one they were assigned: that is, they receive the control if assigned to treatment, and the treatment if assigned to control. If there exist defiers, the relationship between treatment assignment and treatment receipt is non-monotonic. The existence of both defiers and compliers also means there are more structural parameters than we can estimate from the

[24] Both of these assumptions, along with monotonicity and a strong instrument (discussed below), are necessary for valid estimation of the CACE (Angrist and Pischke 2008: 154)

data. While we can estimate the proportions of compliers; never-takers (i.e., non-compliers in the treatment group); and always-takers (i.e., non-compliers in the control group), we cannot estimate the proportion of defiers. Thus, if there are indeed defiers, the IV model will be under-identified (Freedman 2006: 706). In that case, the Wald estimator in equation 7 does not consistently estimate the complier average causal effect. The no-defier condition is not directly testable, so we do not bold it in Table 1. Nonetheless, the assumption is often viewed as one of the more plausible in design-based applications of IV (see, however, Freedman 2006: 700). Certainly, when defiers constitute a very small proportion of the study group, identification and estimation issues from violations of the monotonicity assumption should be limited (Angrist et al. 1996: 451). And certain designs allow for us to dismiss the monotonicity assumption entirely. In cases of one-sided non-compliance, where researchers (or governments, nature, etc.) prevent the control group from having access to the treatment, there are by construction neither always-takers nor, more importantly, defiers.

Finally, both the SEM and potential outcomes approaches require a sufficiently strong relationship between the instrument(s) and endogenous regressor(s). Because weak instruments explain little of the systematic variation in X, the predicted values of X, or $\hat{X}$, approach X. The 2SLS estimator in equation (4) is thus biased in the same direction as the OLS estimator (Bound et al. 1995). In both approaches to IV, this assumption can be tested directly from the data by examining the strength of the relationship between X and Z; see discussion in our first section.

Overall, the discussion in this section suggests several conclusions. First, design-based approaches to IV tend to be more modest in terms of the underlying assumptions. The potential outcomes framework relaxes certain assumptions stipulated in the linear response schedule under SEM (e.g., linearity in parameters, constant effects across units). Moreover, the target of inference—the average causal effect for a particular study group, or the average effect for a sub-group of compliers—is readily characterized and estimated; the model does not presume to extrapolate those effects to units outside the experimental population. Next, while many IV assumptions under SEM remain implicit in the assumption of the response schedule, IV analysis under the potential outcomes framework does a clearer job of disaggregating the key assumptions. Finally—as indicated by the greater number of bolded items in the columns of Table 1—the assumptions of design-based analysis tend to be more directly testable, for example, by assessing balance on pre-treatment covariates across treatment and control groups or through modification of design. The next section illustrates these points through an empirical example.

An illustration: the demand for coffee

How do changes in prices affect demand for coffee? The question recalls those motivating Wright's original work on IV, in which a key issue is the identification of the demand curve for agricultural goods. Yet, one could also approach this question experimentally, by randomly assigning prices to coffee products and assessing how the demand changes in response. Here, we therefore describe two different approaches to answering this question: one in the SEM tradition, another in the design-based framework. The example further illustrates tradeoffs and limitations, as well as areas of convergence between the approaches.

Thus, as in Wright (1928), one option for studying this relationship would be to find an instrument that affects demand but not supply. A researcher might seek to use, say, rainfall as an instrument for coffee prices. Researchers have used rainfall (or deviations from average rainfall) as an instrument in a variety of settings, for example, estimating the effects of economic growth on a variety of dependent variables including civil war in Africa (Miguel et al. 2004) and land invasions in Brazil (Hidalgo et al. 2010). Scholars have also increasingly used rainfall to estimate the effects of turnout on support for particular parties in the United States (Hansford and Gomez 2010; Horiuchi and Kang 2018; Fujiwara et al. 2016), Germany (Arnold and Freier 2016), and Spain (Artés 2014).

Imagine, then, a researcher wants to use changes in rainfall patterns in Uganda to instrument for changes to world coffee prices. Ultimately, she wants to test whether increased prices reduce the demand for coffee. Because Ugandan rainfall should only affect coffee demand through its affect on coffee supply and thus prices, the researcher thinks it might be a valid instrument. If the researcher were to use rainfall as an instrument to estimate a model in the form of Equation (3), what assumptions must be met for a causal interpretation?

The key stipulations are found in the first column of Table 1. Each might raise concerns in this example. We mention only several. The demand schedule might be non-linear—i.e., a demand "curve"—rather than proportional to coffee prices. The elasticity of demand might vary across units, as a function of the availability of substitutes (say, tea), violating the assumption of constant effects across units. And the assumption of constant partial effects, according to which the treatment effect does not vary across the components of the endogenous regressor, might especially suggest issues. In this case, prices can change in response to a variety of events; for example, changes induced by variation in weather patterns may have very different effects than price changes induced by a merger between two large coffee producers.[25] Thus, in the rainfall example, we may not be identifying the effect of price changes generally but rather price changes induced by a particular impetus—rainfall. Additionally, the assumption that demand in unit i does not depend on the exposure to coffee prices in unit j may also be suspect: in an interdependent world economy, spillover is perhaps much more common than non-interference.

Next, concerns might focus on the validity assumptions on the instrument—specifically, that rainfall does not belong in the response schedule. These validity assumptions entail that rainfall is independent of the disturbance term in the response schedule linking price changes to demand and that rainfall only affects coffee demand through price changes. Researchers may be able to argue that changes to rainfall in Uganda are as good as randomly assigned, adding credibility to the assumption of $Z_i \perp\!\!\!\perp \epsilon_i$. However, the exclusion restriction assumption is considerably more difficult to test. Rainfall may change demand for coffee through channels other than supply and thus, price. Angrist and Krueger (2001) discuss the possibility that "sophisticated commercial buyers at the New York Coffee, Sugar and Coca Exchange, where coffee futures are traded, use weather data to adjust holdings in anticipation of price increases that may not materialize in fact" (79).[26]

[25] Similar critiques have arisen around the use of rainfall as an instrument for economic growth. Dunning (2008) suggests that rainfall may induce a very specific type of economic growth that is quite different from growth induced by, for example, technological change in agriculture or an increase in foreign aid.

[26] Research using rainfall as an instrument for economic growth and turnout has often been critiqued with respect to the exclusion restriction. In the case of Miguel et al. (2004), rainfall may lead to flooding on

This last point raises a final, empirical question regarding the strength of the instrument: is a change in rainfall patterns in Uganda enough to change world coffee prices? We could test this assumption using data from rainfall in Uganda and world coffee prices. Do higher levels of rainfall decrease world coffee prices? An F-statistic greater than 10 in the regression of world coffee prices on Ugandan rainfall may suggest the latter is a sufficiently strong instrument. Yet, this rule-of-thumb has weaknesses. The ideal range of rainfall for coffee production is 45-70 inches per year; less than thirty inches is considered too dry, while more than one-hundred inches is considered too wet for coffee to successfully grow (Shaw 1955: 278). Thus, we might expect the relationship between rainfall in Uganda and coffee prices to be U-shaped. The first-stage regression in equation (4) is the linear projection of X onto Z, however. While non-monotonicity in the relationship between rainfall and prices may not affect our interpretation of β (Imbens 2014: 346)—given the modeling assumption of a constant effect—estimating a linear first stage for data that is non-linear may lead us to dismiss an instrument as weak even if it, in fact, has a strong relationship to the endogenous regressor.

The example of rainfall-induced variation in coffee prices thus illustrates several challenges of IV analysis under the SEM framework. Particularly troublesome is the assumption of correct specification of the linear response schedule, which gives rise to a number of other assumptions that must be carefully addressed but often cannot be directly tested—and which are not very clearly illuminated by the model. The design-based approach may provide a way of addressing some of these concerns a priori. Through robust experimental designs, researchers can attempt to reduce many of the issues that arise in the SEM framework. This approach has its own limitations and unverifiable assumptions. A feature and perhaps virtue of this approach, however, is that limited generalizability is baked into the estimand—rather than obfuscated by an apparently general structural equation.

Consider, then, the direct experimental manipulation of coffee prices. Drawing partially on the innovative study of Hainmueller et al. (2014), we imagine a case in which researchers would like to work with supermarkets to manipulate coffee prices and ultimately derive the price elasticity of demand for coffee.[27] The hypothetical researchers work with one hundred supermarkets. Managers from fifty of the supermarkets are told to raise their prices while managers from the other fifty are told to hold their prices constant. The researchers track coffee sales in the one hundred grocery stores for four weeks and then compare how sales changed across the treatment and control groups. How might the researchers analyze their experiment?

One option is intent-to-treat analysis: coffee sales in the fifty supermarkets that were told to raise prices may be compared with the fifty that were told to keep their prices con-

roads and bridges, making it difficult to transport soldiers and thus decreasing the likelihood of conflict (Sovey and Green 2011; Dunning 2012). Sarsons (2015) shows that the relationship between rainfall and conflict in India is strongest in areas downstream of dams, where agricultural income is less susceptible to rainfall shocks due to access to irrigation. As for research on turnout and party support, Horiuchi and Kang (2018) demonstrate that weather directly changes voter support for parties, with rainfall making voters more likely to support Republicans. In fact, most of the benefit obtained by Republicans in rainy elections can be attributed to voters changing their preferences, rather than differential levels of turnout.

[27] Using a randomized control trial with 26 grocery stores in New England, Hainmueller et al. (2014) manipulate both the price and labeling of coffee to understand whether consumers are willing to pay a higher price for fair trade coffee.

stant. This approach takes advantage of the element of the design over which the researcher had most control—the initial randomization of units to treatment and control conditions. Moreover, it can be analyzed with a simple and transparent difference in means estimator that relies on relatively weak assumptions. In cases where researchers can engage in extensive monitoring to reduce non-compliance, as Hainmueller et al. (2014) do, this strategy provides a robust form of estimating the treatment effect of interest. However, when such monitoring is not present and/or non-compliance is high, researchers may wish to estimate a complier average causal effect.

What might non-compliance look like in our hypothetical price-manipulation experiment? Never-takers are defined as those stores that never raise prices; always-takers are those that always raise prices; and defiers are those that raise prices when told not to raise and that do not raise when told to raise. The CACE would thus constitute the treatment effect for stores that raised their coffee prices when instructed to and did not raise their prices otherwise. Under the potential outcomes framework, we could estimate a CACE using the ratio in equation (7). What assumptions are required?

A first assumption in valid estimation of the CACE is that the instrument is (as good as) randomly assigned. In the hypothetical case here, researchers controlled random assignment, which they can subsequently check using balance tests.[28] A second assumption is that treatment assignment affects the outcome only through treatment receipt. In this case, does telling a supermarket manager to raise coffee prices affect coffee sales other than through the actual increase in coffee prices? Perhaps, telling a store to raise coffee prices will result in the store not only increasing coffee prices, but also lowering tea prices. Or perhaps the price increase will lead managers to change the placement of merchandise, such that lower priced coffee is moved to occupy a more visible place in the store. The researchers may send monitors to check whether coffee prices are actually being changed, which serves as both a test of the instruments' strength and also a measure of compliance. However, it would be difficult if not impossible for the researchers to account for all of the changes made in response to the coffee price change announcement that might also affect coffee sales. And finally, we assume absence of defiers, or monotonicity. Is it plausible that there exist stores in the sample that raise prices when told not to and do not raise prices when told to? It seems unlikely in this case, although we might imagine some managers might look to defy an outside researcher who tells them how to control prices in their own store.

Both implicit and explicit in the discussion above are a number of tradeoffs regarding IV under the SEM and design-based approaches. In the SEM framework, many of the assumptions are invoked by the response schedule itself. Researchers must carefully justify ex-post the correct specification of the response schedule, including the assumptions that follow from it. The design-based approach under potential outcomes allows for a weakening of the constant unit-level effects assumption but clarifies a new assumption—monotonicity.[29] This approach also draws on the potential outcomes framework, which relaxes the assumption of the linear response schedule and disaggregates the assumptions otherwise implied by the model into distinct parts. Concerns about these assumptions are addressed through

[28] Unlike the analysis, which was performed using store-weeks as units, the balance tests were performed at the store level, giving a sample size of only 26, which may limit their power.

[29] However, as we noted, the assumption of monotonicity exists implicitly in the SEM framework. Allowing for defiers would imply more structural parameters than can be estimated from the data.

careful research design; natural and field experiments provide a way to find or develop robust instruments that are plausibly exogenous and that only affect the outcome through the endogenous regressor. The design-based approach thus has a number of desirable properties, which include both the clear statement of and the possibility to test key assumptions.

A key limitation to both approaches, however, involves generalizability of interpretation, which the coffee demand example illustrates well. Often, instrumental variables analysis involves an intervention that addresses only one component of the treatment of interest. For example, it is hard to know exactly what the price manipulation experiment tells us about the relationship between coffee price and demand. Artificial manipulation of a coffee price may truly isolate the general effect of interest, but more likely, it tells us about only one component of a "price" treatment—the actual changing of prices by a store. This change occurs at the end of the supply chain and tells us little about the result of price changes due to farming, tariffs, increased fuel prices, etc. The same can be said of using rainfall as an instrument for coffee price changes. However, in the SEM model, this claim is particularly muddled by the specification of the linear model, where the researcher is claiming to identify through X_i in equation (3) the effect of "prices." In reality, and as discussed above, price changes induced by rainfall may have a very different effect than price changes induced by other "interventions." From this perspective, an advantage of the design-based approach is then not just the clarity and relative testability of the key assumptions – but that the framework makes clear its limitations. Here, modesty is a virtue: the SEM approach is subject to the same kinds of weaknesses but because of the lack of specificity embedded in the model specifications, it tends to overstate its ability to deliver on its ambitions.[30]

Conclusion

Since Wright's initial work on supply and demand, social scientists have used instrumental variables to study the effects of a number of independent variables that would otherwise be difficult (if not impossible) for the researcher to randomly assign. Some instruments, like rainfall, rely on plausibly exogenous natural variation that affects an endogenous independent variable of interest. Ramsay (2011) studies as-if random variation in natural disasters to understand how a country's level of democracy responds to a change in oil prices. Other instruments rely on lotteries, where the instrument is—due to actual randomization—independent of pre-treatment causes of X and Y. Researchers have used the Vietnam draft lottery as an instrument for military service (Angrist 1990; Erikson and Stoker 2011) and lottery winnings as an instrument for income (Doherty et al. 2006). While these cases assure that the instrument, Z, is—in expectation—uncorrelated with the disturbance term from the regression, there remain important concerns about both violations of the exclusion restriction and the correct specification of the response schedule.

In observational work, challenges often arise to validating the identifying assumptions, which researchers have shown varying degrees of willingness to acknowledge and address (see e.g., Sovey and Green 2011: 194 and Staiger and Stock 1997: 597, fn. 2). While certain assumptions can be directly tested from the data (relevance, rank, identifiability of parameters) and others may be plausible by design (independence of instrument and

[30] For a related point in the context of fixed effects regressions, see Aronow and Samii (2016).

pre-treatment causes of X, Y), the remaining assumptions of structural equation modeling generally raise concerns that often cannot be fully allayed. The exclusion restriction and specification of the response schedule remain particularly troublesome.

Instrumental variables in the structural equation modeling framework offers a potential solution to a key problem: identifying the causal effect of X on Y given the stipulation of a particular structural equation in which the model is presumed but X is assumed to be endogenous. However, the SEM framework may also prove restrictive; it imposes an assumption of a linear response schedule that does not allow for estimation of heterogeneous treatment effects. It further builds monotonicity into the estimand rather than addressing it as an assumption (Imbens 2014: 346).

The potential outcomes framework overcomes some of these limitations by making more explicit the underlying key assumptions, which are often more plausible, less restrictive, and easier to test from the data. Linearity and constant effects assumptions are relaxed under this approach, and other assumptions, like the exclusion restriction and random assignment of units to values of the instrument, are directly stated, such that they might be separately addressed and evaluated. The monotonicity assumption, an added assumption under heterogeneous treatment effects, is generally viewed to be plausible, although plausibility should be judged based on the specific intervention.

Ultimately, despite the promise of instrumental variables, there remain key limitations. Both the SEM and design-based approaches suffer a common challenge of generating results that generalize beyond the particular intervention that gave rise to the instrument. Often instruments affect the independent variable of interest through a specific (and perhaps, narrow) channel. Generalizing from that particular component of the treatment to formulating broader claims should be done only after consideration of other factors that may have induced change in the independent variable. Robust IV analysis thus requires that researchers consider both the story that can be told from the data given the assumptions and the story that *cannot* be told given the inherent difficulties of generalization under the IV framework.

Appendix

We derive the algebraic equivalence of the two-stage least-squares (equivalently, the instrumental-variables) estimator in Equation (5) and the Wald estimator of the Complier Average Causal Effect for a finite population in Equation (7), in the bivariate case with one treatment and one control group. Here, Y_i is the outcome variable; $X_i = 1$ if unit i receives treatment and otherwise $X_i = 0$; and $Z_i = 1$ if unit i is assigned to treatment and otherwise $Z_i = 0$. The number of units is given by N, and the number of units assigned to treatment is $m < N$. Without loss of generality, index the units assigned to treatment by $i = 1, \ldots, m$ and the units assigned to control by $i = m + 1, \ldots, N$. Thus, we have

$$
\begin{aligned}
\widehat{\beta}_{1,\text{2SLS}} &= \frac{\widehat{Cov}(Y_i, Z_i)}{\widehat{Cov}(X_i, Z_i)} \\
&= \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^{N}(X_i - \overline{X})(Z_i - \overline{Z})} \\
&= \frac{\sum_{i=1}^{N}(Y_i Z_i) - \overline{Y}\sum_{i=1}^{N} Z_i - \overline{Z}\sum_{i=1}^{N} Y_i + N\overline{Y}\,\overline{Z}}{\sum_{i=1}^{N}(X_i Z_i) - \overline{X}\sum_{i=1}^{N} Z_i + \overline{Z}\sum_{i=1}^{N} X_i + N\overline{X}\,\overline{Z}} \\
&= \frac{\sum_{i=1}^{N}(Y_i Z_i) - m\overline{Y}}{\sum_{i=1}^{N}(X_i Z_i) - m\overline{X}} \qquad (\text{because} \sum_{i=1}^{N} Z_i = m,\ \overline{Z} = \frac{m}{N}) \\
&= \frac{m(\overline{Y_1} - \overline{Y})}{m(\overline{X_1} - \overline{X})} \qquad (\text{because} \sum_{i=1}^{N} Y_i Z_i = m\overline{Y_1}) \\
&= \frac{m(\overline{Y_1} - \frac{m}{N}\overline{Y_1} - \frac{N-m}{N}\overline{Y_0})}{m(\overline{X_1} - \frac{m}{N}\overline{X_1} - \frac{N-m}{N}\overline{X_0})} \qquad (\overline{Y} = \frac{\sum_{i=1}^{m} Y_i + \sum_{i=m+1}^{N} Y_i}{N} = \frac{m}{N}\overline{Y}_1 + \frac{N-m}{N}\overline{Y}_0) \\
&= \frac{\frac{m(N-m)}{N}(\overline{Y_1} - \overline{Y_0})}{\frac{m(N-m)}{N}(\overline{X_1} - \overline{X_0})} \\
&= \frac{\overline{Y_1} - \overline{Y_0}}{\overline{X_1} - \overline{X_0}}
\end{aligned}
\tag{A.1}
$$

The first step uses the definition of the sample covariance; we divide through by n/n. Next, we multiply out terms, then use the definition of $Z = 1$ as the units assigned to treatment (and thus, the sum, $\sum Z_i$, is m, while the mean, $\overline{Z}$, is m/N) and cancel terms. In the following step, we use the fact that the product, $Y_i Z_i$, will be zero when $Z_i = 0$ and Y_i when $Z_i = 1$. The sum of this product will thus equal the mean outcome for the treated units, $\overline{Y_1}$, times the number of treated units, m. The next step uses the fact that the mean, $\overline{Y}$, is simply a weighted average of the mean outcome under treatment, $\overline{Y_1}$, and the mean outcome

under control, $\overline{Y_0}$. The final steps factor out common terms and reduce the equation to the Wald estimator.

References

Aldrich, J. (1993). Reiersøl, Geary and the idea of instrumental variables. *Economic & Social Review*.

Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, 80(3):313–336.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15(4):69–85.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arnold, F. and Freier, R. (2016). Only conservatives are voting in the rain: Evidence from German local and state elections. *Electoral Studies*, 41:216–221.

Aronow, P. and Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250267.

Aronow, P. M., Baron, J., and Pinson, L. (2015). A Note on Dropping Experimental Subjects Who Fail a Manipulation Check. SSRN Scholarly Paper ID 2683588, Social Science Research Network, Rochester, NY.

Artés, J. (2014). The rain in Spain: Turnout and partisan voting in Spanish elections. *European Journal of Political Economy*, 34:126–141.

Bollen, K. A. (1989). *Structural Equation Models with Latent Variables*. John Wiley & Sons.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430):443–450.

Cox, D. (2009). Randomization in the design of experiments. *International Statistical Review / Revue Internationale de Statistique, Vol. 77, No. 3 (December 2009), pp. 415-429*, 77:415–429.

Doherty, D., Gerber, A. S., and Green, D. P. (2006). Personal Income and Attitudes toward Redistribution: A Study of Lottery Winners. *Political Psychology*, 27(3):441–458.

Dunning, T. (2008). Model Specification in Instrumental-Variables Regression. *Political Analysis*, 16(3):290–302.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press. Google-Books-ID: ThxVBFZJp0UC.

Durbin, J. (1954). Errors in Variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3):23–32.

Erikson, R. S. and Stoker, L. (2011). Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes. *The American Political Science Review*, 105(2):221–237.

Freedman, D. A. (2006). Statistical Models for Causation: What Inferential Leverage Do They Provide? *Evaluation Review*, 30(6):691–713.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Fujiwara, T., Meng, K., and Vogl, T. (2016). Habit Formation in Voting: Evidence from Rainy Elections. *American Economic Journal: Applied Economics*, 8(4):160–188.

Geary, R. C. (1949). Determination of Linear Relations between Systematic Parts of Variables with Errors of Observation the Variances of Which Are Unknown. *Econometrica*, 17(1):30–58.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Goldberger, A. S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6):979–1001.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education. Google-Books-ID: njAcXDlR5U8C.

Hainmueller, J., Hiscox, M. J., and Sequeira, S. (2014). Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment. *The Review of Economics and Statistics*, 97(2):242–256.

Hansford, T. G. and Gomez, B. T. (2010). Estimating the Electoral Effects of Voter Turnout. *The American Political Science Review*, 104(2):268–288.

Heckman, J. J. and Robb, R. (1986). Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In Wainer, H., editor, *Drawing Inferences from Self-Selected Samples*, pages 63–107. Springer New York, New York, NY.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Hidalgo, F. D., Naidu, S., Nichter, S., and Richardson, N. (2010). Economic Determinants of Land Invasions. *The Review of Economics and Statistics*, 92(3):505–523.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Horiuchi, Y. and Kang, W. C. (2018). Why Should the Republicans Pray for Rain? Electoral Consequences of Rainfall Revisited. *American Politics Research*, 46(5):869–889.

Imbens, G. W. (2014). Instrumental Variables: An Econometricians Perspective. *Statistical Science*, 29(3):323–358.

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

Miguel, E., Satyanath, S., and Sergenti, E. (2004). Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy*, 112(4):725–753.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, 62(3):760–775.

Morgan, M. S. (1990). *The History of Econometric Ideas*. Cambridge University Press.

Nickerson, D. (2008). Is voting contagious? evidence from two field experiments. *American Political Science Review*, 102:49–57.

Ramsay, K. W. (2011). Revisiting the Resource Curse: Natural Disasters, the Price of Oil, and Democracy. *International Organization*, 65(3):507–529.

Reiersøl, O. (1945). *Confluence analysis by means of instrumental sets of variables.* Almquist & Wiksells, Uppsala. OCLC: 793451601.

Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural "Natural Experiments" in Economics. *Journal of Economic Literature*, 38(4):827–874.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58.

Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, 26(3):393–415.

Sarsons, H. (2015). Rainfall and conflict: A cautionary tale. *Journal of Development Economics*, 115:62–72.

Sartori, G. (1970). Concept Misformation in Comparative Politics. *The American Political Science Review*, 64(4):1033–1053.

Shaw, E. B. (1955). *World economic geography: with an emphasis on principles*. Wiley. Google-Books-ID: 4cE1AQAAIAAJ.

Sovey, A. J. and Green, D. P. (2011). Instrumental Variables Estimation in Political Science: A Readers Guide. *American Journal of Political Science*, 55(1):188–200.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

Stock, J. and Trebbi, F. (2003). Who Invented Instrumental Variable Regression? *Journal of Economic Perspectives*, 17:177–197.

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan Company.