Developmental Psychology
2010, Vol. 46, No. 6, 1454–1480

0012-1649/10/$12.00 DOI: 10.1037/a0020204

Causal Inference and Developmental Psychology

E. Michael Foster
University of North Carolina, Chapel Hill

Causal inference is of central importance to developmental psychology. Many key questions in the field revolve around improving the lives of children and their families. These include identifying risk factors that if manipulated in some way would foster child development. Such a task inherently involves causal inference: One wants to know whether the risk factor actually causes outcomes. Random assignment is not possible in many instances, and for that reason, psychologists must rely on observational studies. Such studies identify associations, and causal interpretation of such associations requires additional assumptions. Research in developmental psychology generally has relied on various forms of linear regression, but this methodology has limitations for causal inference. Fortunately, methodological developments in various fields are providing new tools for causal inference—tools that rely on more plausible assumptions. This article describes the limitations of regression for causal inference and describes how new tools might offer better causal inference. This discussion highlights the importance of properly identifying covariates to include (and exclude) from the analysis. This discussion considers the directed acyclic graph for use in accomplishing this task. With the proper covariates having been chosen, many of the available methods rely on the assumption of "ignorability." The article discusses the meaning of ignorability and considers alternatives to this assumption, such as instrumental variables estimation. Finally, the article considers the use of the tools discussed in the context of a specific research question, the effect of family structure on child development.

Keywords: causal inference, methodology, single-parent families

Causality is central to developmental psychology. Many developmentalists are motivated not only to understand but to improve the lives of children and their families. Psychologists want not only to identify developmental risks but also to understand mechanisms by which development can be fostered. This aim is noble but represents an enormous scientific challenge. In particular, achieving this goal requires knowledge of cause-and-effect relationships. This issue is especially challenging because many developmental questions involve putative risks, exposures, conditions, or characteristics that vary across children and families but to which individuals cannot be randomly assigned, such as growing up in a single-parent family.

In those situations, developmentalists correctly understand that causal inference—inferring causal relationships—is difficult. They recognize the fundamental problem of causal inference—that associations alone do not reveal causal relationships. One cannot distinguish the consequences of a behavior or program from those that shape the choice to engage or participate in the first place. Smokers, for example, definitely suffer from more illnesses than nonsmokers, but the causal question is whether smoking itself is responsible (Rosenberger & Lachin, 2002). Would eliminating smoking actually improve the population's health?

The last 30 years have produced superior methods for moving from association to causation (i.e., causal inference). These tools have been developed outside of developmental psychology, in fields as diverse as epidemiology, economics, and statistics. This article describes and illustrates these methods. However, a full and technically detailed treatment of the various methodological developments is beyond the scope of a single article, even a long one. Book-length treatments of these issues are available and cited below. Those treatments, however, are tailored to neither the interests nor the background of developmental psychologists.

The purpose and the organization of this article reflect the aims and current state of developmental psychology and are guided by four premises. First, causal inference is essential to accomplishing the goals of developmental psychologists. If psychologists hope to improve the well-being of children and families, then their research must tackle causality. This article claims that causal inference—the study and measurement of cause-and-effect relationships outside of random assignment—should be the goal of developmental research in most circumstances. If developmental psychologists want to avoid causality, they need to do so explicitly and explain the limitations of their work (Rutter, 2007).

Second, in many analyses, psychologists unfortunately are attempting causal inference but doing so badly—that is, based on many implicit and, in some cases, implausible assumptions. These assumptions are both statistical (such as linearity) and conceptual (such as no unobserved confounding).[1] Given that causal inference

This article was published Online First August 2, 2010.

E. Michael Foster, University of North Carolina, Chapel Hill.

Correspondence concerning this article should be addressed to E. Michael Foster, Gillings School of Global Public Health, University of North Carolina, Rosenau Hall, Campus Box 7445, Chapel Hill, NC 27599-7445. E-mail: emfoster@unc.edu

[1] Pearl (2000) drew a strong distinction between statistical and causal assumptions. However, in fixed-effects or instrumental variables estimation, the distinction is not so clear: The assumptions that shape estimation also have causal content.

is impossible without assumptions (some untestable)[2] (Holland, 1986; Pearl, 2000), the third premise is that these assumptions should be identified explicitly and checked empirically and conceptually.

The fourth premise is that once introduced to the broader issues, developmental psychologists will recognize the central importance of causal inference and naturally embrace the methods available. To some extent, as discussed below, the methods of causal inference are creeping into developmental psychology. For example, recent articles use fixed-effects and instrumental variables estimation (Dearing, McCartney, & Taylor, 2006; Gennetian, Magnuson, & Morris, 2008). (Other examples are cited below.) What is missing is an overarching framework in which to embed comparisons of alternative approaches. This article provides that framework and starts the process by which developmentalists can adopt these methods more generally.

One feature of this discussion will strike readers already familiar with the broader issues and methods of causal inference: The treatment here emphasizes propensity scores but not because such methods are the final word on causal inference. Propensity-score-based methods assume "ignorability" (or no unobserved confounding), and many researchers (especially economists) highlight the role of unobservables in shaping decisions and outcomes. However, given that the vast majority of articles in developmental psychology currently rely on some form of linear regression, propensity scores represent a solid step toward plausible causal inference.

A central goal of this article is to promote broader thinking about causal inference and the assumptions on which it rests. Methods that relax ignorability follow naturally after regression and propensity score methods. One characteristic of the broader literature, however, is that methodologists in different fields differ substantially (and rather heatedly, at times) as to the role of unobserved confounding. (The ignorability assumption represents something of a dividing line between econometrics and other fields.) Though less detail is provided on these methods, the illustrative example below represents the natural progression one might take through the various forms of causal inference, methodologically and conceptually, including relaxing ignorability.

Before describing specific methods, the article considers two conceptual frameworks for thinking about causal questions. No doubt the reader can anticipate that a key issue will involve selecting covariates used for adjusting between-groups comparisons of the treated and untreated. The reflection these methods precipitate also can help researchers refine the research questions of interest and help one select and apply statistical methods to answer those questions. Before doing so, however, this article reviews the state of current practice.

First, however, a note about language is in order. The methodological literature generally refers to causality as involving an "event, condition, or characteristic" (Rothman & Greenland, 2005, p. S144). Epidemiologists often use the term *exposure*. Consistent with the literatures in economics, statistics, and biostatistics, this article refers to any of these as *treatment*. In this framework, a treatment might include a program or a characteristic (such as living in a female-headed family) or different levels of a continuous characteristic (such as hostile attributions).

Current Practice: Confusion

In the typical developmental psychology article, the issue of cause and effect is acknowledged, perhaps as a limitation of the article. In identifying this issue, the authors may recognize that "only laboratory experiments or randomized controlled trials allow any firm causal inference" (Rutter, 2007, p. 377). In many instances, these authors urge the reader to recognize that "our results represent only associations."

At this point, the articles tack in one of two directions. Both are dissatisfying and potentially misleading. In the first case, the authors hold causal inference as unattainable. These authors will state, for example, that spanking is associated only with aggressive outcomes. Of course, one is left to wonder about the usefulness of such information. As Rutter (2007) noted, this approach "sounds safer, but it is disingenuous because it is difficult to see why anyone would be interested in statistical associations or correlations if the findings were not in some way relevant to an understanding of causative mechanisms" (p. 377). For example, some research suggests that television viewing is associated with attention problems (Christakis, Zimmerman, DiGiuseppe, & McCarty, 2004). If this information does not reveal whether a parent should limit his or her child's television viewing, thinking of a use for the finding is difficult (Foster & Watkins, 2010).

A second group of authors embrace causality. These researchers often rely on the longitudinal nature of their data to make the leap from associations to causality. Unfortunately, they often apply tools that have limitations for performing causal inference, such as linear regression, and make implausible assumptions about the nature of the association of interest. The authors often leave these assumptions unstated or may be unaware of these assumptions themselves. For example, in standard regression, the model assumes that the covariates are related linearly to both the outcome and the putative cause. For example, the researcher may recognize that maternal depression may influence both a child's outcomes and television viewing and potentially confounds the effect of the latter on the former. What the researcher may not recognize is that the standard (linear) regression assumes that the effect of depression on television viewing is linear. Misspecification of this relationship may leave residual confounding.

Regrettably, some authors straddle the two groups. These authors often stray into causal interpretations of what are associations. Rutter's (2007) diagnosis of this group is spot on: These authors "are careful to use language that avoids any direct claim for causation, and yet, in the discussion section of their articles, they imply that the findings do indeed mean causation" (p. 377). To finesse these issues, these researchers use other terms to describe the relationships they identify, such as *predictive*. They do not claim that a behavior causes another but that one predicts the other.

This situation creates a swamp of ambiguity in which confusion thrives. In many instances prediction is of interest because one wants to predict what would happen were conditions changed. This

[2] Heckman (Heckman & Vytlacil, 2007a) parsed the problem in three steps: defining the set of counterfactuals, identifying parameters from hypothetical population data, and identifying parameters from real data. The key is that all agree that producing causal estimates relies on assumptions that are inherently untestable.

"counterfactual" lies at the heart of causal inference. Rather than circumvent issues of causation, prediction only raises them.

Why Causal Inference?

Perhaps developmental psychologists can just avoid the issue entirely by studying associations only. As indicated above, causal thinking and, as a result, causal inference are unavoidable. Indeed they are essential to accomplishing the goals of developmental psychology. One can support this claim in three ways. First, as noted above, a major goal of psychology is to improve the lives of humanity. Much of developmental science is devoted to understanding processes that might lead to interventions to foster positive development. Second, causal analysis is unavoidable because causal thinking is unavoidable. Sloman (2005) argued that causality is one of the fundamentally invariant relationships that humans use to make sense of the world: Causal relations are one of the "the constant, regular, systematic relations that hold between the objects, events, and symbols that concern cognition" (p. 17). Causal relationships exert a gravitational pull on one's thinking.

Finally, even if researchers can distinguish associations from causal relationships, lay readers, journalists, policymakers, and other researchers generally cannot. For example, knowing that children in single-mother households have worse outcomes than other children is a useful association. However, that association is routinely interpreted as causal—that were the mother to marry, the child's outcomes would be improved. Such a conclusion depends on a causal relationship, and the support for such a causal relationship is fairly weak (Foster & Kalil, 2007). Laypersons, researchers, and policymakers find it difficult to distinguish these two notions. As a result, single-parent mothers can be stigmatized, and the belief that their decisions about marriage and fertility have caused their troubles leads to government inaction. Bad causal inference can indeed do real harm.

Furthermore, if a researcher resists the urge to jump from association to causality, other researchers seem willing to do so on his or her behalf. Interventionists, for example, move from an association to the notion that changing the behavior will change the outcome: Manipulation of this sort presumes causal relationships. When one manipulates one variable in order to influence another, the (causal) impact on the effect will often be smaller than the association observed in data. What has happened is that part of that relationship was explained by other, third factors (confounders or omitted variables) that were unaffected by the manipulation. The result is interventions that are much less powerful than the original observational studies would suggest (compare Conduct Problems Prevention Research Group, 1992, with Conduct Problems Prevention Research Group, 2007). One explanation is that the foundational studies involved associations only.

Causal Inference as the Goal of Developmental Psychology

Few issues have held the interest of philosophers, social scientists, and statisticians as consistently over centuries (Pearl, 2000). The lesson of this quest is not that causal relationships can never be established outside of random assignment, but that they cannot be inferred from associations alone—that some additional assumptions are required.

The goal of this research should be to make causal inference as plausible as possible. Doing so involves applying the best methods available among a growing set of tools. As part of the proper use of those tools, the researcher should identify the key assumptions on which they rest and their plausibility in any particular application. The researcher should check the consistency of those assumptions as much as possible using the available data. In many, if not most, instances, key assumptions will remain untestable. The plausibility of those assumptions needs to be assessed in light of substantive knowledge, such as how a parent, child, teacher, or other actor decides to engage in a treatment or not.

Of course, what constitutes credible or plausible (and how it relates to the strength of assumptions) is not without debate. Manski (2007), for example, argued that plausible is synonymous with simplicity; he identified the "the law of decreasing credibility" and argued that stronger assumptions are less credible. Many would dispute this claim, including reviewers of this article. What seems clear is that all assumptions are best made explicit. Stronger assumptions can (and perhaps should) generate more information. As Manski has demonstrated, one can identify a maximum and minimum for the estimated effect of a treatment under very weak assumptions or features of the data (e.g., if the outcome measure has maximum and minimum values). No doubt, however, many would find the lack of a single point estimate rather dissatisfying. On the other hand, complex econometric selection models can generate estimates of the correlation between one's outcomes when treated (or exposed) and untreated. This information is not available even in a randomized experiment.

This article cannot resolve these issues even for a single substantive problem; its broader purpose is to establish plausible causal inference as the goal of empirical research in developmental psychology. At this point, much, if not most, of developmental psychology involves implausible causal inference. Such inference could be improved even without dramatically changing the complexity of the analysis. As the field develops, the debate over whether complicated econometric models are required will become more salient.

The article now turns to the first of two frameworks that are useful for conducting causal inference.

Two Frameworks for Causal Inference

Two conceptual tools are especially helpful in moving from associations to causal relationships. The first involves the directed acyclic graph (DAG). This tool assists researchers in identifying the implications of a set of associations for understanding causality and the set of assumptions under which those associations imply causality. Moving from association to causality requires ruling out potential confounders—variables associated with both treatment and the outcome. The DAG is particularly useful for helping the research to identify covariates and for perhaps understanding unanticipated consequences of incorporating these variables.

Tool 1: DAGs

Like other scientists, computer scientists have been interested in causality, in particular, in identifying the circumstances under which an association can be interpreted as causal. Because they are directional, causal relationships among sets of variables imply

different covariance matrices. When placed in the context of their relationship to other variables, a given pattern of covariances (associations) can rule in or out causal relationships working in different directions involving two variables (Spirtes, Glymour, & Scheines, 2000).

For that reason, computer scientists have developed a symbolic representation of dependencies among variables, the DAG (Greenland & Pearl, 2008; Greenland, Pearl, & Robins, 1999; Pearl, 2000). A DAG comprises variables and arrows linking them. The DAG should be grounded in one's conceptual understanding of the treatment or exposure of interest. In some instances, key variables may be unobserved, unmeasured, or otherwise unavailable, but one should still include these variables in the DAG.

The DAG is directed in the sense that the arrows represent causal relationships. The model assumes a certain correspondence between the arrows in the graph and the relationships between the variables (i.e., their joint probability function). In particular, if one cannot trace a path (or sequence of arrows) from one variable to another, then the variables are not associated. Those paths involve stepping from one variable to the next—no relationships skip over another variable. This is the causal Markov assumption: The absence of a path implies the absence of a relationship. However, not all paths are apparent, and some are created by the analyst, perhaps unintentionally. For that reason, the DAG does more than represent the obvious.

A key feature of the DAG is structural stability: An intervention on one component of the model does not alter the broader structure. Intervening on a variable may change how it relates to other variables statistically. If one assigns children randomly to preschool, then enrollment may no longer reflect parents' education. However, the downstream consequences of preschool enrollment are not altered. For example, the relationship between preschool and early literacy is left unchanged (Pearl, 2000).

The DAG also assumes a preference for simplicity and probabilistic stability. Simplicity means that models that represent data with fewer linkages are preferred to the more complex. Stability refers to the robustness of a set of relationships across a range of possible magnitudes. For example, one might link two statistically independent constructs using two paths representing different psychological mechanisms. This model, however, would represent the data (i.e., the lack of association) only if the two mechanisms worked in different directions and exactly offset each other. Such a pair of relationships would not be considered stable. Stability strengthens one's ability to infer that no association between variables means no causal relationship.

A DAG looks like a path diagram or a structural equations model. Key features distinguish DAGs. First, the DAG is not linear or even parametric. Any discussion of conditioning on one variable does not imply a particular statistical method, such as stratification or regression. The DAG helps the reader to distinguish the various steps in causal inference; the choice of covariates is a separate decision from how to select a specific statistical method.

Second, unlike a structural equations model, the DAG contains no bidirectional arrows implying simultaneity. This assumption is not quite as restrictive as it seems. If one wanted to model jointly determined outcomes, some additional assumptions are needed. One possibility involves timing of measurement (as discussed below). In that case, the same construct measured at different times appears as different variables, and the DAG would incorporate a lag structure. If two variables are simultaneously determined, the DAG could incorporate this possibility by treating the two as reflecting a common cause.[3]

Under these assumptions, the DAG can accurately represent the joint probability function describing a set of variables. (In this case, the DAG is said to be consistent with the probability function [Pearl, 2000]).

The essence of the DAG can be grasped by thinking about three variables, *X*, *Y*, and *Z*. These variables may be related in any of three ways.

***Z* is a common cause of *X* and *Y*.** The DAG in Figure 1 illustrates this possibility. In the figure *Z* is actually a vector, Z_1 and Z_2. One can see that Z_1 is a common cause of *X* and *Y*, and the effect of Z_1 on the latter is mediated by Z_2.

The figure reveals that treatment and the outcome will be related, because the former causes the latter and because of their link to the two *Z* variables. Even in the absence of a treatment effect, treatment and outcome will be associated. This is classic confounding. In causal inference, the path passing through the confounders is known as a "backdoor path."

Any veteran of multivariate regression knows the solution here: One would want to adjust for the effect of Z_1. It is important to note that regression is only one way of adjusting,[4] but the various methods all have the same effect on the DAG. As apparent in the bottom half of the figure, paths from and to the confounder are eliminated. (In Figure 1, one conditions on Z_2.) In effect, the backdoor path is blocked; conditioning (or holding constant) means that the confounder is no longer related to the treatment or the outcome. Having blocked all backdoor paths, *X* and *Y* are known as "*d*-separated." The essence of the DAG is that once all backdoor paths are blocked, one can infer that an association is a causal relationship given the assumptions on which the DAG rests.

***Z* is a common effect of *X* and *Y*.** The implications of conditioning depend on how *X*, *Y*, and *Z* are related. *Z* might be caused by both *X* and *Y*. Variables like *Z* are referred to as "colliders" because the arrows from the other variables converge at that variable (see Figure 2). The existence of a collider does not mean that *X* and *Y* are related in any way. However, conditioning on the collider does establish a relationship. This insight is critical.

[3] This is consistent with the idea (in ordinary regression) that omitted variable bias and simultaneity bias are essentially the same problem: Both generate a covariance between a regressor and the model's error term (Greene, 2008).

[4] One might use stratification, regression, matching, or a range of other methods. The essence is as follows. Suppose the variable *Z* takes on one of *k* possible values and that one is interested in the effect of *X* on the probability distribution of *Y* [$P(Y|X)$]. When one conditions, one calculates

$$P(Y|X) = \sum_{i=1}^{k} P(Y|X,z_i)P(z_i)$$

In other words, one calculates the relationship between *Y* and *X* within strata defined by the values of *Z* and then combines across values of *z*. If *Z* is continuous, the calculation involves an integral. The issues of how to write the probability distribution of *Y* and its relationship to *X* and *Z* (e.g., linear regression) pose a technical question rather than a conceptual one.

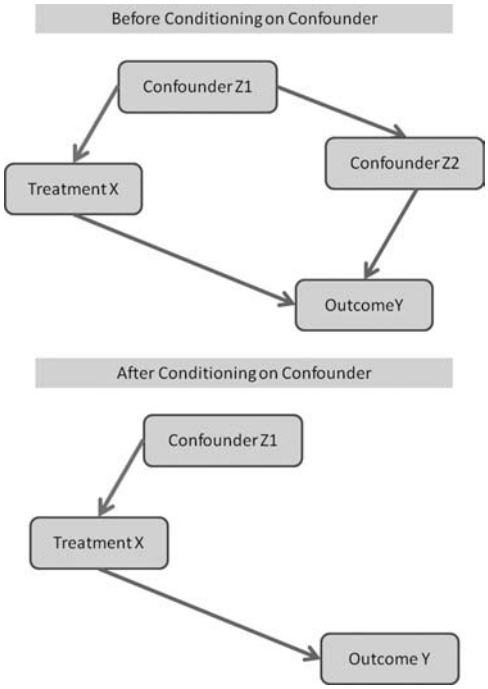

Figure 1. Conditioning on confounder.

For example, suppose that *S* is success in first grade, *P* is attending a high-quality preschool, and *A* has asthma. In the population, these two (*P* and *A*) are unrelated, yet both contribute to the likelihood of school success. Now condition on school success (i.e., limit the sample to the successful). What is the relationship between asthma and preschool attendance? Among the successful, those with asthma were more likely to have attended preschool. Conditioning on success creates dependence between the two variables. (This relationship is easiest to see when *P* and *A* are the only determinants of *S*. In that case, successful children who did not attend preschool could not have had asthma.)

Conditioning on a collider creates a spurious relationship between *X* and *Y*: It opens a backdoor path involving the determinants of the collider. Such a relationship can inflate or suppress (or hide) a true causal effect.

***Z* mediates the effect of *X* on *Y*.** A final relationship between *X*, *Y*, and *Z* involves *Z* as a mediator. In that case, conditioning on *Z* can have two effects, one likely familiar to readers. Consider Figure 3. First, conditioning on *Z* will redefine the effect of *X* on *Y*—that effect is now the direct effect. The indirect effect refers to the effect of *X* that manifests itself through its effect on *Z*. When one conditions on *Z* (ignoring for the moment *W*), one again breaks the relationship between *X* and *Y* in this figure: There is no direct effect of *X* on *Y* (Baron & Kenny, 1986; Judd, Kenny, & McClelland, 2001).

The usefulness of the DAG (and the complexity of causal inference) is perhaps most apparent when more than three variables are involved, especially when one is unmeasured. Consider *W* in Figure 3, an unobserved determinant of the mediator. This problem could occur even in a clinical trial, where the treatment has been randomly assigned and so is free of unobserved confounding. The mediator in this case is a collider—it is caused by

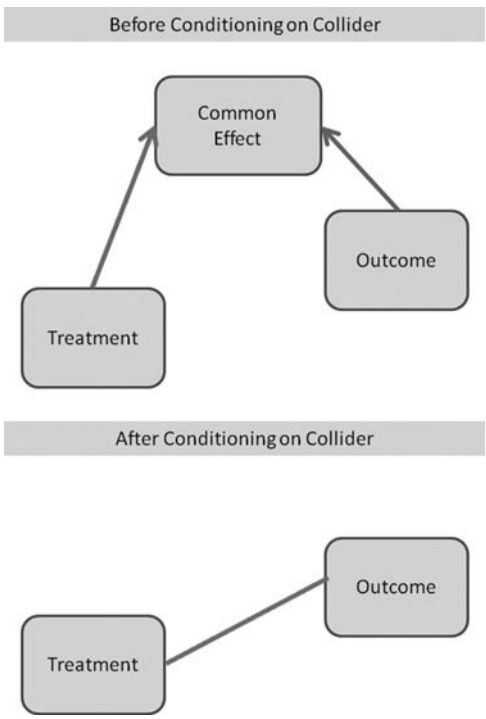

Figure 2. Conditioning on collider.

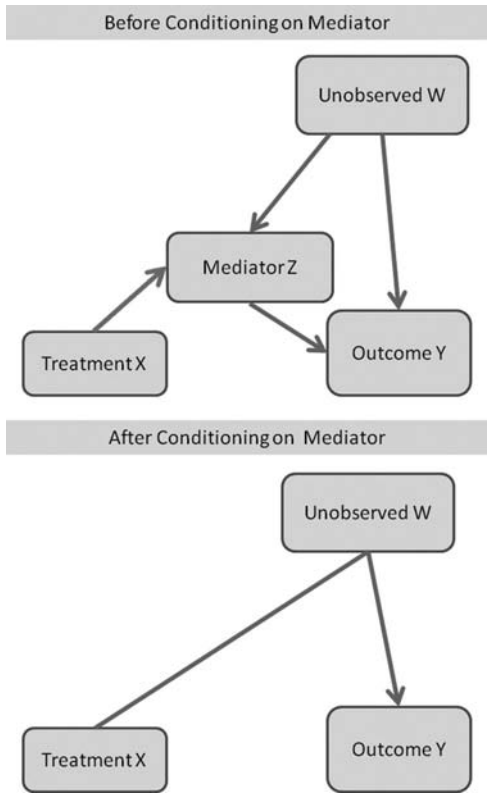

Figure 3. Conditioning on mediator.

the treatment as well as by *W*. Conditioning on the mediator creates a spurious relationship between treatment and *W*. This backdoor path would bias the estimated direct effect of treatment (Sobel, 2008).[5] In that case, conditioning on *Z* does not properly partition the effect of *X* on *Y* into direct and indirect components.[6]

With covariates chosen, the stage is now set to consider alternative statistical methods (such as regression, matching, and stratification) for conditioning on *X* and *Z* (i.e., forming conditional expectations of the outcome). To understand how these tools relate to causal inference and to one another, a second set of tools is helpful.

Tool 2: The Potential Outcomes Framework

The counterfactual lies at the heart of causal inference. What would happen to an individual or entity if one changed its exposure? This perspective defines the effect of interest: the difference between what did happen and what would have happened had the individual not experienced the treatment.[7] Even in instances where the treatment cannot be manipulated (a person's race or gender), causal inference is still possible, and it rests on the existence of a counterfactual. To capture the effect of a child's race, one has to assess how that child would have fared if he or she had been born Black instead of White. (Researchers are not in complete agreement about this point. Some would argue that causal inference is not possible if a treatment cannot be manipulated [Holland, 1986]. On the other hand, some experts [e.g., Pearl, 2000] clearly take it as a given that one could look at the effects of gender.)

One can express the counterfactual more formally using the potential outcomes framework (Holland, 1986; Rubin, 2005). In that framework, one is interested in a treatment (or exposure or characteristic; *D*) and some outcome (*Y*). $D = 1$ or 0 for the treated and untreated groups, respectively. One can think of each individual as having two possible outcomes, $Y_{D=1}$ and $Y_{D=0}$, corresponding to the individual's outcomes if treated or not. One can characterize each individual by three variables (Y_1, Y_0, D): the outcome if treated, the outcome if not treated, and the treatment status. The fundamental problem is that one does not observe (the joint distribution of) all three random variables. Rather one observes D for each individual (was he or she treated or not?) and either Y_1 or Y_0 for the treated ($D = 1$) and untreated ($D = 0$), respectively. One can write that $Y_{\text{obs}} = (D \times Y_1) + (1 - D) \times Y_0$, where Y_{obs} is the observed outcome.

The person subscript has been omitted to this point, but these outcomes differ across individuals, because individuals differ in the treatment they receive and because of other factors (differentiating even those who receive the same treatment). What one would like to know is the treatment effect—that is, $\tau = Y_{1,i} - Y_{0,i}$—ideally for every individual *i*. One cannot calculate this term for a given person.

Can one get traction on this problem if one reduces one's goal? What if one wanted to know only $\tau = E[Y_{1,i} - Y_{0,i}]$, the mean of this distribution of effects? (The *E* identifies the expectations operator or the average. When one writes $E[Y|X = x]$ or $E[Y|x]$, this notation refers to the average value of *Y* for a specific value or range, *x*, of *X*. Following convention, random variables are denoted with uppercase variables, and specific values of those variables are lowercase.) Because one chose the mean and not some other characteristic of the distribution (such as the median), one can see that $\tau = E[Y_{1,i}] - E[Y_{0,i}]$.

The problem is (still) that neither of the two terms is observed. In particular, one observes not $E[Y_{1,i}]$ but $E[Y_{1,i}|D_i = 1]$, and not $E[Y_{0,i}]$ but $E[Y_{0,i}|D_i = 0]$. One sees the outcome under treatment only for those treated and the converse for untreated individuals. These quantities are related to the terms of interest:

$$E[Y_{1,i}] = p(D_i = 1)E[Y_{1,i}|D_i = 1] + p(D_i = 0)\mathbf{E}[\mathbf{Y_{1,i}}|\mathbf{D_i} = \mathbf{0}]. \quad (1)$$

The average outcome for all individuals were they all treated is the weighted average of the treated outcome for those who received treatment and those who did not. The weights are the proportions of the sample that do and do not receive treatment. The bolded term is unobserved—that expectation and the expression as a whole cannot be calculated without some additional assumption. This problem is the fundamental problem of causal inference.

Similarly,

$$E[Y_{0,i}] = p(D_i = 1)\mathbf{E}[\mathbf{Y_{0,i}}|\mathbf{D_i} = \mathbf{1}] + p(D_i = 0)E[Y_{0,i}|D_i = 0]. \quad (2)$$

The bolded term is again unobserved: One does not know the average untreated outcome for the individuals who received treatment.

Given these two terms, one can rewrite τ as

$$\tau = \left(p(D_i = 1)E[Y_{1,i}|D_i = 1] + p(D_i = 0)\mathbf{E}[\mathbf{Y_{1,i}}|\mathbf{D_i} = 0] \right) - \left(p(D_i = 1)\mathbf{E}[\mathbf{Y_{0,i}}|\mathbf{D_i} = 1] + p(D_i = 0)E[Y_{0,i}|D_i = 0] \right) \quad (3)$$

One can rewrite this term as

$$\tau = p(D_i = 1)\left[E[Y_{1,i}|D_i = 1] - \mathbf{E}[\mathbf{Y_{0,i}}|\mathbf{D_i} = \mathbf{1}] \right] + p(D_i = 0)\left[\mathbf{E}[\mathbf{Y_{1,i}}|\mathbf{D_i} = \mathbf{0}] - E[Y_{0,i}|D_i = 0] \right]. \quad (4)$$

The treatment effect is the weighted sum of the treatment effect for those treated and those not treated. One labels the first as the *average effect of treatment for the treated* (ATT) and the second as

[5] Methods do exist for incorporating mediators that suffer from unobserved confounding. These include principal stratification (for details, see Barnard et al., 2003; Frangakis & Rubin, 2002).

[6] Other relationships between *X*, *Y*, and *Z* are possible. In another situation, *X* is subject to unobserved confounding but *Z* is not. In that case, one can obtain appropriate estimates of *X*. See the discussion of the "front-door criterion" in Pearl (2000).

[7] Researchers disagree on the origins of the potential outcomes framework. Some argue that the ideas can be found early in the work of Jerzy Neyman (1923; as cited in Splawa-Neyman, Dabrowska, & Speed, 1990) or Roy (1951). The potential outcomes framework is generally implicit in these early articles, and whom researchers cite generally reflects their discipline. My assessment is that the clearest and most influential (re)statement of these ideas can be found in Rubin (1974, 1975).

the *average effect of treatment on the untreated* (ATU). τ is labeled the *average effect of treatment* (ATE).

Economists describe this problem using somewhat different language. In particular, the ATE is *not identified*. Equation 4 represents one equation but includes two unknowns; it is not possible to generate a unique estimate of the treatment effect. This problem would exist no matter what the sample size: It is a problem of logic rather than an empirical problem per se. At this point, one has reached a dead end unless one can add some additional data or assumptions of some sort, information that would allow one to identify the model.

Before proceeding, one wants to make clear three assumptions made implicitly so far. The first assumption is *stable unit treatment value assumption* (SUTVA; Rubin, 1980).[8] This assumption requires that one's counterfactual states ($Y_{0,i}$ and $Y_{1,i}$) do not depend on the treatment status of other individuals. One can note that in the math above, there is no interference among individuals: An individual's outcome Y_i does not depend on the treatment received by person *j*. As discussed below, many problems in developmental science may not fit this assumption.

The second assumption is *positivity*. This assumption requires that the probability that a given individual receives each level of treatment is positive for every combination of treatment and covariates. This assumption eliminates illogical possibilities, such as men developing uterine cancer. As discussed below, this assumption has an empirical counterpart.

A third assumption is *consistency* (Cole & Frangakis, 2008). (Economists label this policy invariance [Heckman & Vytlacil, 2007a].) This assumption implies that the outcome of treatment does not depend on the assignment mechanism. For example, this assumption means that the returns to enrolling a child in day care are the same for all regardless of the mix of incentives that led to that choice. Some families may enroll their children because of a government subsidy. On the other hand, others may enroll because they perceive large benefits. Consistency means that the benefits of early childhood education do not depend on the mix of incentives.

An Aside: The Value of Random Assignment

As noted above, one's effort to identify the three treatment effects is stalled. One can move forward by making assumptions about the treatment assignment mechanism. One alternative, however, is random assignment. Because of developmentalists' long tradition of experimental studies, this article pauses to consider random assignment.

Random assignment solves the identification problem (Rosenberger & Lachin, 2002). One can replace the unknowns above with terms one can calculate. In particular, in the equations, one can freely substitute $E[Y_{0,i}|D_i = 0]$ for $E[Y_{0,i}|D_i = 1]$ and $E[Y_{1,i}|D_i = 1]$ for $E[Y_{1,i}|D_i = 0]$. The only remaining uncertainty is statistical uncertainty stemming from sampling. The uncertainty due to the treatment mechanism no longer exists; that mechanism can be effectively ignored. Ignorability is not an assumption but a feature of the design.

The essence of random assignment is "exchangeability": One can exchange the experiences of individuals not receiving the treatment (which are observed) for those of individuals currently receiving treatment were they not to receive treatment (which is not observed). One can be confident that no confounding occurs; no possibility exists for confusing the effect of treatment with preexisting differences among individuals actually receiving different treatments. This lack of confounding is true for both the variables one can measure and those one cannot or has not (Rosenberger & Lachin, 2002). The latter is unobserved confounding, and removing it is among the principal benefits of randomization.

Moving Forward Without Random Assignment: Ignorability

One often cannot randomize individuals to treatments of interest. The solution is to make stronger assumptions about treatment assignment.[9] One such assumption is ignorability, or the ignorable treatment assumption. Ignorability assumes that among individuals with an equivalent profile of covariates, treatment assignment is as if random assignment spontaneously occurred.

This assumption can be written in different ways, but the assumption stipulates that

$$E[Y_{1,i}|X_i = x, d_i = 0] = E[Y_{1,i}|X_i = x, d_i = 1]. \quad (5)$$

Conditional on the set of covariates, *X*, the value of the outcome at a level of treatment is unrelated to treatment chosen. In essence, one can exchange the expectation one does not observe (the left-hand side) with the expectation one does observe (the right-hand side). One can see, therefore, why another name for ignorability is exchangeability.

Note that the specific form of ignorability depends on the treatment effect of interest. Equation 5 is enough to estimate the ATU. On the other hand, an analogous equality,

$$E[Y_{0,i}|X_i = x, d_i = 0] = E[Y_{0,i}|X_i = x, d_i = 1], \quad (6)$$

is required to estimate ATT. Both are required to estimate ATE.

In a regression context, this assumption is equivalent to the conditional independence assumption (Angrist & Pischke, 2008). However, the choice of a specific statistical model, like regression, is a separate task from identifying the causal model, which should be done first. For that reason, this article refers to the ignorability assumption in a more general form.

Return of the DAG: Insights Into the Counterfactual

The DAG can illustrate key points about the counterfactual. First, randomization essentially erases lines into the treatment and eliminates backdoor paths. Randomization accomplishes this task for both observed and unobserved correlates of treatment status.

Second, ignorability can be expressed in terms of the DAG. Within strata defined by (or otherwise conditioning on) the ob-

[8] This is yet another use of the term *stability*, but this use is indeed related to the term as used in describing the DAG. If treatment effects spilled over across individuals, then a broader change in the environment (and the corresponding structure of relationship among variables in the DAG) might occur.

[9] There are other assumptions that are weaker but do not produce point estimates. For outcome measures with maximum and minimum values, one can generate a range of possible estimates for the treatment effects. (That range is not probabilities: No estimate in the range is more or less likely than any on estimate in that range.) A point estimate is not possible, however (see Manski, 1997, 2007; Manski & Nagin, 1998).

served covariates, ignorability assumes away any additional backdoor paths involving unobserved variables.

Third, the discussion above is fairly vague about how one might chose the X variables. In that task, the value of the DAG complements the mathematical apparatus of the counterfactual. As discussed above, the DAG can help one identify variables that should—and should not—be included in the vector of covariates, X. One should block backdoor paths and not open others by conditioning on colliders. One also can see that in some instances, one need not condition on all potential confounders. In Figure 1, for example, one could condition on either Z_1 or Z_2.

Linear Regression as a Tool for Causal Inference

The discussion of the counterfactual focuses on the conditional expectation of the outcome, Y. To say more about the counterfactual and how it relates to other variables, one must write this expectation as a function of the covariates, X. There are many ways to write such an expectation, but one way is the ubiquitous linear regression. Ordinary linear regression is arguably the Swiss army knife of the social sciences. However, regression has features that need to be considered for good causal inference. To the extent standard practice ignores these features, regression may be less than ideal for causal inference. (Note that the use of ordinary least squares or analysis of variance in this article includes path analysis as well as recursive structural equations models, that is, models in which the relationships all point in one direction.)

The Strengths and Weaknesses of Linear Regression

Linear regression provides a way to form conditional expectations. The key issue in calculating the causal effects is that one needs the conditional expectation for an outcome in the counterfactual state. One needs $E[Y_1|D = 0,X]$, for example. Regression essentially provides that by estimating a slope for the variable X using information from those who did receive treatment ($D = 1$)—more specifically, $Y = XB$, estimated using data on the treated. In instances where one assumes the slope of X does not depend on treatment status, one can use data from the treated and untreated observations to form the expectation.

One should recognize that statistics provides many ways to form such expectations. Perhaps the best way to recognize the strengths and weaknesses of parametric, linear regression is to compare that method with an alternative, such as locally weighted scatterplot smoothing, or nonparametric regression. Locally weighted scatterplot smoothing involves estimating the expected value of a function by approximating the outcome, Y, at each possible value of X, the covariate. In approximating the function at X, one would rely on values of Y for observations with values of X in the neighborhood of a given value, X_0.

The approximation depends on three key factors. The first is the bandwidth. How large would the neighborhood be around X_0 that would contribute to the estimation of the function at X_0? Smaller neighborhoods around X_0 would mean the approximation would be based on fewer cases, but the values of X involved would be relatively close to X_0. If X and Y are related, a smaller neighborhood would mean that the values of Y included are closer to the true value of Y_0 as well. One can see the trade-off involved. Smaller neighborhoods mean that the observations used have values of X closer to X_0; of course, as the neighborhood shrinks, the number of observations used shrinks, and the resulting estimates become less precise.

A second issue involves weighting. One could imagine placing more importance on those cases within the neighborhood that are closer to X_0. In essence, the bandwidth implies weighting. Observations outside the bandwidth receive a zero weight. One might specify a weighting function to weight observations within the neighborhood differentially.

Third, one might make some assumptions about the nature of the function in that neighborhood. For example, one might assume that with the neighborhood around X_0, the function does not make any jumps (a finite second derivative). Or one might assume the function is linear within that neighborhood. One might connect the lines estimating within the neighborhoods to approximate the function as a whole. The fitted line would consist of small linear sections (splines), but the overall line could be quite nonlinear (Wood, 2006).

One could use this fitted approximation for a variety of purposes. For example, one might estimate the value of the function (i.e., of Y) at points for which no data were observed. X might range from 0 to 1, and though no datum was observed at $X = .50$, one might want to generate a predicted value of Y there. If there were closely adjacent points, one could imagine interpolating across these points (say from .51 to .49). For the purpose of prediction, one might generate a prediction for points outside the observed range. For example, one might want a prediction at $X_0 = 1.1$. One can see that generating such predictions is risky. As one moves farther and farther from the observed range, the prediction depends increasingly on points farther and farther away. Moreover, the extrapolation to these points would become increasingly reliant on the presumed function at the ends of the observed range. The resulting extrapolation would have considerable uncertainty associated with it.

In this light, ordinary regression makes rather extreme assumptions. Regression uses all the data to estimate the line at every point on the range of X. That line is assumed to have the same linear shape at every point. Observations vary in their importance in calculating the line: Those with greater leverage have a larger influence on the fitted line. Leverage reflects how far a given observation's values for the explanatory variables are from the mean. Some observations contribute more to the estimated slope, but that contribution does not vary across the other observations in the data, no matter how similar those observations are.[10] As a result, the key observations for predicting Y at a specific value of

[10] To see this point clearly, think of a small data set with five observations. Both Y and the single covariate X have mean 0, so one can ignore the intercept. Without loss of generality, one can imagine these points spread evenly between -1 and 1 (-1, $-.5$, 0, .5, 1). In that case, the prediction of Y_i equals

$$\frac{Cov(X,Y)}{Var(X)}X_i = \frac{X_1Y_1 + X_2Y_2 + X_3Y_3 + X_4Y_4 + X_5Y_5}{X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2}X_i .$$

The contribution of X_5 to the denominator will be larger than that of X_3 and, depending on the value of Y_5, may be larger for the numerator as well. The key point is that the contribution of X_5 will be the same for X_4 and X_1 even though X_4 is much closer to X_5 than X_1.

X may not be those close at hand but those with extreme values of X.

Regression has many positive features. The resulting relationship will be relatively easy to describe, and the approximation at X_0 will depend on more data than in nonparametric regression. As a result, the standard error of that approximation may be smaller. If the functional form is correct, one can produce predicted values for values of X_0 beyond the observed range. The problem, of course, is that the functional form may be incorrect or the estimated line may reflect a few outlying cases.

Linear Regression and Estimating a Treatment Effect

To really understand regression, one needs to write the model in mathematical form. One presents the standard regression model below, but first one writes it in a more elaborate form to tie it to the causal inference framework. In particular, one can write the standard regression framework as

$$Y_{1,i} = \beta_0^1 + \beta_1^1 X_{1,i} + \beta_2^1 X_{2,i} + \varepsilon_i^1$$

$$Y_{0,i} = \beta_0^0 + \beta_1^0 X_{1,i} + \beta_2^0 X_{2,i} + \varepsilon_i^0$$

$$Y_i = Y_{1,i} d_i + Y_{0,i}(1 - d_i). \quad (7)$$

Subtracting the second line from the first implies variation in the difference in Y between the treated and untreated states. This framework naturally allows for treatment moderation. As noted above, the key issue for causal inference is that a given individual contributes to the estimation of the equation for Y_1 or Y_0 but not both.

Typically, analysts simplify Equation 7 by assuming $\beta_j^1 = \beta_j^0$ and by redefining the regression coefficients. Doing so results in the standard regression model:

$$Y_i = \gamma_0 + \tau d_i + \gamma_1 X_{1,i} + \gamma_2 X_{2,i} + \delta_i. \quad (8)$$

The τ is generally interpreted as the average treatment effect or just the treatment effect without specifying for whom. Graphically, the model assesses the treatment effect as the distance between two parallel planes. (If there is only one explanatory variable, two parallel lines.) Because the model does not specify any variation in the effect, the three effects (ATE, ATT, and ATU) are assumed equal.

This assumption points one toward the first problem. If there is variation in the so-called treatment effect, τ does not correspond to any of the key estimates in the causal inference. In particular, the estimate of τ does not represent the ATE, ATU, or ATT (Morgan & Winship, 2007). This problem reflects the estimator generated by ordinary regression or analysis of variance.

To see this, imagine strata formed by combinations of the explanatory variables. The relationship between those strata, the outcome, and the treatment can be completely flexible. (The model can be saturated.) The ordinary least squares estimate is the weighted within-strata, between-groups difference. The weights reflect not only the distribution of cases across the strata (as they would to estimate one of the three treatment effects of interest) but also an added factor: the variance of the treatment within strata. Strata where the observations are split equally between the treatment and comparison groups have greater weight. This feature of the model is useful for explaining the variance in the outcome but not so useful for understanding the effect of treatment.

As a result, the estimated effect can differ quite a bit from the ATE, ATT, and ATU. The regression estimate can approximate the ATE under two circumstances. The first occurs when the weight implied by the propensity score is similar across strata, that is, if the likelihood of treatment does not depend on the covariates. Of course, in that situation, adjusting for the covariate is unnecessary. The second occurs when there is no variation in the treatment effect. Such an assumption is strong and unlikely.

A second problem lies in linearity. Model misspecification (such as nonlinear relationships) creates confounding; the misspecification of the model essentially represents an omitted variable that creates a dependency between the treatment and the covariate involved. That effect can also spill over to and involve the other covariates in unknown ways.

A third problem also involves the distribution of the covariates. Values of the covariate where both treated and untreated cases can be found are particularly important. In some instances, the area of overlap may be small; there may be ranges of the distribution of the covariate where all individuals with those values of X are either treated or untreated. The distribution of the covariates is known as the *support*, and when the overlap is small, one has a support problem.[11]

How regression handles those cases depends on the model specification. In the case of a saturated model with a flexible specification, regression recognizes that no estimate of the treatment effect is possible; those observations do not contribute to the estimated treatment effect (Angrist & Pischke, 2008). In the case of a less flexible specification (e.g., a linear model), regression essentially extrapolates the treatment effect from the range where overlap exists. Perhaps even more worrisome is that these cases may have an especially strong influence on the slope of the fitted straight line. (These outlying cases often have high leverage [Belsley, Kuh, & Welsch, 2004].)

In either case, regression provides no indication of having taken this step. As a result, the estimated treatment effect applies to only a portion of the sample. (Of course, one can assume this problem away by assuming the treatment effect is constant, but that assumption seems untenable given the discussion above.)

One could argue that these problems are not inherent to regression but reflect standard practice. For example, one could identify and address the problems driving the distorted results in at least three ways. First, one might consider whether the underlying regression function is common across the treatment and comparison groups. One can test in any of several ways, including simply interacting treatment status with the other covariates.

Second, one could test whether within each group the relationship between X and Y is linear. Such tests are relatively simple. One could take the residuals from the linear regression model and regress them on X itself. If the model specification is correct, then X should explain no variation in the residuals. Third, one could identify the disjuncture in the distribution of X between the two

[11] It is striking that James Heckman, best known for his work on handling selection on unobservables, emphasis the support problem as one of the major sources of bias in estimating treatment effects (Heckman, Ichimura, & Todd, 1998).

groups. This task would involve simply comparing the distribution of the covariates across the treated and untreated groups. Researchers typically present descriptive tables comparing the means and standard deviations across the groups. Simply comparing other percentiles would go a long way toward identifying potential problems.

Alternatives to Regression: Matching and Stratification

What are the alternatives to forming conditional expectations? One option is matching. If one can find cases that have similar values for the covariates, one can calculate the difference between the matched cases. One then can average those effects to compute the overall effect of treatment.

In terms of the counterfactual, in matching, one forms the $E[Y_{1,i} \mid D_i = 0, S(X_i = x)]$ (the average outcome in the treated state for the untreated) using $E[Y_{1,i} \mid D_i = 1, S(X_i = x)]$. S refers to the members of the sample where $X_i = x$, a vector of specific values for x. (One could indicate the vector with an underscore.) Ignorability is still the key: Once again, one is assuming $E[Y_{1,i} \mid D_i = 0, S(X_i = x)] = E[Y_{1,i} \mid D_i = 1, S(X_i = x)] = E[Y_{1,i} \mid S(X_i = x)]$.

One might condition, for example, on gender and calculate treatment effects for boys and girls. If the sample were half boys and half girls, one could average the estimated effects for an overall effect. If boys were overrepresented among the treatment effect, one could average using the gender mix among the treated to get an ATT. Similarly, one could calculate the ATU.

When one wants to match on a single variable, matching has several advantages, such as the absence of any presumed functional form between the covariates and treatment. Problems arise, however, when one wants to match on multiple variables. Seldom is the sample large enough to match individuals on exact values of a large number of variables. In that case, matching involves finding cases that are close to one another, and the wrinkle involves a definition of *close*. The latter involves specifying a weight function that combines differences across variables into a measure of closeness.

Similar issues arise with stratification. Stratification represents a form of matching in which individuals are grouped into strata or combinations of explanatory variables. The problem is that as the number of variables (and their possible values) increases, the stratum become homogeneous in terms of the treatment: Each includes only individuals who did or not receive the treatment. In that case, the assumption of positivity no longer holds.

The propensity score represents a means of addressing this problem.

A Summary of the Covariates: The Propensity Score

The propensity score is a statistical convenience, no more and no less, and its key properties are mechanical. It is the predicted probability of receiving the treatment. One can calculate the propensity score in a variety of ways, but most common is using a parametric model (like logistic regression). The propensity score is a weighted sum of the covariates in which the weighting reflects the strength of the association between the covariates and treatment status, that is, the potential for the covariates to confound the relationship between treatment. The original propensity score article demonstrates that when conditioning on the propensity score, any remaining between-groups variation in the covariates included reflects chance alone (Rosenbaum & Rubin, 1983).

Note that the propensity score addresses the behavior of the observed covariates only. Like regression, propensity-score-based matching assumes away unobserved confounding; it does not eliminate it. One still needs to think carefully about which covariates to include and which to omit. As the DAG illustrates, including some covariates may make (unobserved) confounding worse.

One advantage of the propensity score is that it eliminates many of the practical problems associated with matching or stratification. However, its value may principally lie in its use as a diagnostic tool. Before considering the analysis of outcomes, one considers what might be learned from the estimation of the propensity score itself.

The Propensity Score as a Diagnostic Tool

The first diagnostic task involves the effectiveness of the propensity score itself as a summary of the covariates (and implicitly the adequacy of linearity in describing the relationship between the covariates and the treatment). The second involves identifying treatment cases that have no counterpart in the comparison group (and vice versa).

Diagnostic 1: Assessing balance. In technical terms, the propensity score is one of several possible *balancing scores*. This label refers to the fact that the distribution of the covariates—conditional on the propensity score—no longer varies meaningfully between the treatment and comparison cases. In other words, the propensity score captures all of the variation between the treatment and comparison groups in X that is relevant for understanding the treatment effect. By assumption, any remaining between-groups differences in X are noise. This property, however, holds only as a sample grows in size to infinity. In a finite sample, significant differences in the distribution of covariates may remain even after cases are matched on the propensity score (Hansen, 2008). Like model misspecification in regression, this imbalance can produce additional confounding in finite samples.

The key task, then, is to try to refine the propensity score to eliminate this residual confounding. This task is known as *assessing balance*. How might one do so? Simply, one needs to compare the covariates across groups in the same way one will assess the difference in the outcomes. For example, as discussed below, a popular means of analyzing outcomes is to stratify the data based on the propensity score (into, say, quintiles). For analyzing outcomes, the key issue is to determine whether the outcome of interest varies *within* these strata. Similarly, what one wants to know about the balance of the covariates is whether the distribution of the covariates differs between the treatment and comparison groups within strata. To fully assess distributional differences, one should check both the means and the variances of the explanatory variables.

What is one to do if such differences are statistically significant (i.e., if the covariates are unbalanced)? The answer is to modify the predictive model until the covariates balance. One might add square terms to the model or interactions and recheck balance. The specific steps required depend on the analysis. In their own work in child development, I often find it quite helpful to estimate separate propensity score models by race and gender.

The educated reader, of course, will worry about overfitting the data. Indeed, that is exactly what the analysis is doing. Note, however, that one is estimating the model for a specific purpose: to capture variation in the covariates relevant to estimating a treatment effect. One is not attempting to predict treatment involvement in a new sample. For that reason, one should not offer interpretations for the coefficients of the propensity score equation. They are simply an empirical summary of the covariates (Morgan & Harding, 2006; Morgan & Winship, 2007).

A more general option for improving the propensity score estimation might be to abandon parametric regression functions such as the probit model. One could, for example, employ the tools of data mining (Hand, Mannila, & Smyth, 2001; Hastie, Tibshirani, & Friedman, 2001; Sumathi & Sivanandam, 2006). Models such as neural networks can be more effective in classifying cases into groups, such as the treatment and comparison groups. One wrinkle, however, is that such models may find increasing numbers of cases whose predicted probabilities are either 0 or 1 (or very close). These cases can be classified quite effectively: The model is suggesting that all cases with that profile of explanatory variables are in either the treatment or the comparison group (propensity scores equal to 1 or 0, respectively). This possibility points the reader to a second diagnostic issue: identifying cases that are unique to one group or the other.

Diagnostic 2: Unmatched cases. In some instances, the relationship between treatment and a covariate is strong. In such instances, one may receive a message about "perfect prediction" from the software package estimating logistic regression. This message corresponds to covariates that perfectly predict treatment status. For example, Stata screens variables for this issue and omits them from the analysis. No propensity score is produced for the cases involved (i.e., those with that profile of the explanatory variables).[12]

What can one do with such cases? The best solution is to drop these cases from the analysis and describe them separately, especially when few cases are involved. On the other hand, the issue is tricky if a substantial number of cases are involved. Dropping these cases begins to change the meaning of the estimand itself. For example, if one ethnic group does not participate in the treatment, then the treatment estimate will not generalize to this group (Rosenbaum & Rubin, 1985). In many instances, the group dropped will have a combination of characteristics, making them difficult to describe and, by extension, difficult to describe the group to whom the estimated effect applies.

The implications of dropping cases depend on the treatment effect of interest (Pearl, 2000). For example, dropping cases that are exclusively comparison cases has no implications for the ATT: That effect captures the difference between treated and untreated cases for the latter. Similarly, for estimating the ATU, the mismatched treatment cases can be dropped without generating problems. Because the ATE is a function of the ATU and the ATT, dropping either group has potential negative implications for that estimand.

Hopefully, the number of cases that have propensity scores exactly equal to 0 or 1 is rather small. However, predicted probabilities in the neighborhood of 1 or 0 are much more likely. Most propensity score software packages have an arbitrary rule for discarding such cases. For example, one might discard treatment cases with a propensity score above the highest value in the sample for the comparison case.

Having identified the mismatched cases and balanced the covariates, the analyst then selects a method for analyzing the outcome, that is, for forming conditional expectations using the propensity score to capture the covariates.

Note that both diagnostics identify potential problems with regression. Checking the support identifies circumstances where ordinary regression would rely heavily on the functional form employed in the outcome regression. Given that most analysts generate the propensity score with logistic regression, checking balance effectively involves checking whether an additive, linear function describes the relationship between the covariates and (the log-odds of) treatment. For that reason, even a researcher who chooses to rely on a standard regression model for assessing the treatment effect would benefit from propensity score diagnostics.

Having formed the propensity score, one can analyze treatment and outcomes in a range of ways. The Appendix describes the various alternatives for analyzing data using propensity scores. A key point for the reader at this point is to recognize that there is no single method known as "propensity score analysis." All, however, retain the ignorability assumption, and some would argue that ignorability is inherently implausible.

Alternatives to Ignorability

Economists, in particular, would argue that ignorability is unrealistic. Their models of behavior are grounded in the assumption that economic agents (e.g., consumers) are well informed—indeed, better informed than the researcher—and act rationally (Heckman & Vytlacil, 2007a). For example, in determining whether to enroll a child in preschool, an economist would work from the assumption that a parent knows more about the child than a researcher could ever hope to capture with his or her measures.

However, one cannot simply relax ignorability but must replace it with another assumption. This trade-off reflects the fundamental problem of causal inference and is a cost of being unable to assign treatment status randomly.

Modeling Unobservables

One alternative to ignorability is to recognize that unobservables do indeed affect both treatment and outcomes and make some assumptions about those relationships. Although the unobservables are not measured (by definition), one can assume they follow some distribution and can specify a relationship between them and the (potential) outcomes.

[12] In the standard application, the researcher is left to wonder whether perfect prediction is a result of substantive considerations or chance. For example, perfect prediction might result because some characteristic rules out or rules in participation in the treatment (e.g., gender and breast feeding). On the other hand, perfect prediction can occur by chance alone and corresponds to sparse cells. In an analysis of breast feeding, for example, none of the six Vietnamese women in the study may have breast-fed their children. This distinction is important for conceptual applications, but in the case at hand—summarizing covariates—the underlying motivation makes no difference.

Perhaps best known among these models are selection models. These models have existed for decades, and much is known about their statistical properties. They replace ignorability with statistical and conceptual assumptions (Heckman & Vytlacil, 2007a, 2007b). The selection model involves a model for determining the conditional expectation of the outcome, a model determining treatment status and a specification of the relationship between the two equations. Economists generally stipulate that the consequences of the treatment differ across individuals, and those who benefit most (or are harmed the least) select the treatment. (This feature is known as *essential heterogeneity* [Heckman & Vytlacil, 2007a].) Because those returns vary for reasons that are both observed and unobserved, essential heterogeneity implies a violation of ignorability.

The model of the treatment choice embodies a set of additional assumptions. One must identify the decision makers responsible for selecting a treatment and the basis for that choice. The latter includes information about whether and how much the outcome varies in each treated state and whether agents maximize well-being in each state. (An alternative, for example, would be a strategy in which agents rank the states according to the worst outcome in each state.) One must also give some thought to whether the treatment choice reflects subjective perceptions and whether those perceptions are correct. One also needs to stipulate a specific form linking treatment and *X*, and like the outcome model, misspecification has a variety of negative effects.

Having specified the full model, one can estimate the parameters of the two equations jointly, allowing for the interdependency of the unobservables in each. One can see the challenge involved. One must distinguish two processes when each is shaped by an unobserved variable. As a result, parameter estimates can be sensitive to the functional form of the error distributions, and slight changes in model specification can influence estimates of the treatment effect. Research demonstrates that the performance of selection models is improved dramatically by an "exclusion restriction," a variable that affects the treatment directly but only influences the outcome through the treatment. The variable involved is known as an *instrumental variable*, and by assumption, the treatment fully mediates the effect of the instrumental variable on the outcome. One can see the value of the instrumental variable in the selection model: It provides added empirical leverage for distinguishing the causes and consequences of treatment because it generates variation in the former unrelated to the latter.

As Heckman and Vytlacil (2007a) and others describe, the selection model performs well with a valid exclusion restriction. When embedded in the broader selection model framework, the instrumental variable can be used to estimate all the treatment effects of interest as well as the full distribution of treatment effects. (Note that the discussion here is oversimplified. It is indeed true that the selection model can be identified without an exclusion restriction [an instrumental variable]. However, the text here is accurate in terms of the selection models commonly available to developmental psychologists in statistical packages such as Stata. For a full discussion of identification, see appendices in Heckman & Vytlacil, 2007a, as well as Abbring & Heckman, 2007. Some approaches are semiparametric, and this approach has some advantages even in the presence of a valid exclusion restriction [Abadie, 2003]. A nice introduction and overview of the various forms of the selection model can be found in Blundell & Dias, 2009.)

Instrumental Variables Estimation

Some statisticians and others (such as epidemiologists) remain skeptical of the selection model and often prefer ignorability to alternative assumptions. Still other statisticians, however, do employ instrumental variables without the full apparatus of the Heckman model. Known as *instrumental variables estimation*, this method has a long history in statistics and econometrics and allows one to estimate the effect of a treatment even with unobserved confounding.

The intuition is relatively simple. Suppose one has a dichotomous instrumental variable. In the case of maternal employment, the instrumental variable might be whether the state has a higher minimum wage than the federal requirement. In that case, instrumental variables estimation substitutes two indirect comparisons for one direct one. That is, rather than compare the outcome of children living with mothers who do and do not work, instrumental variables estimation involves comparing children living in high- and low-minimum-wage states and adjusting for differences in the employment of mothers living in the two groups of states. In this way, one avoids comparing the children of mothers who do and do not work, which may potentially suffer from unobserved confounding.

How does instrumental variable estimation compare with the selection model? First, both instrumental variables estimation and selection models depend on the validity of the instrument. If the instrument somehow affects the outcome directly or is correlated with unobserved determinants of the outcome, the desirable statistical properties of instrumental variables estimation (and selection models) are lost. In the hypothetical example above, states with more generous minimum wages might differ in other, child-friendly ways. In essence, such a correlation would open a back-door path from the instrument to the outcome. Another possible problem involves a "weak instrument," an instrumental variable that is only weakly related to the treatment of interest. In that situation, an instrumental variable can produce estimates with worse statistical properties than the standard regression model (Bound, Jaeger, & Baker, 1995).

This issue points to a key property of instrumental variables estimation: All of its desirable statistical properties are asymptotic (i.e., involving samples approaching infinity). In small samples, instrumental variable estimates are biased (meaning that the confidence interval for the treatment effect is not centered on the true value). The best one can say about instrumental variables estimation is that it is consistent: As the sample size grows to infinity, the confidence interval collapses on the true value. As a practical matter, these properties suggest that using instrumental variables estimation with 75 cases is a bad idea.

Another disadvantage of instrumental variables estimation involves the nature and interpretation of the estimated treatment effect. In the presence of essential heterogeneity, the estimated effect is neither the ATT, the ATE, nor the ATU. The resulting estimate is known as the *local average treatment effect*. This effect pertains to a rather difficult-to-describe group: those individuals who would begin working if they moved from a low-wage to a

high-wage state. This group cannot even be observed, but the estimated treatment effect pertains to them.[13]

Instrumental variables estimation can be found in several forms. Two-stage least squares is one form; regression discontinuity, another (Hahn, Todd, & Van der Klaauw, 2001). The latter is growing in popularity, especially among researchers in education. The method involves a covariate that predicts both treatment status and the outcome; that variable does not meet the requirement for an instrumental variable. However, the distribution of the variable is characterized by a discontinuity or jump that influences only treatment status. That jump in the regressor is the instrumental variable, and the method works by comparing cases on either side of the discontinuity adjusting (typically using regression) for the "main" effect of the covariate. For an example, see the analysis of the effect of Head Start in Ludwig and Miller (2007).

The Natural Experiment

One source of instruments involves so-called natural experiments. Such experiments involve circumstances where an event or policy causes a shift in participation in the treatment of interest. One example involves the effect of income and children's development. Poor and other families differ in a wide range of ways that likely influence the effect of poverty on child development (e.g., maternal education). Researchers correctly worry that direct comparisons of children in poor and other families capture the effect of poverty as well as unmeasured characteristics. In other words, the ignorability assumption may very well not apply. The instrumental variables estimate seems like a natural solution. The wrinkle, of course, is finding an exclusion restriction.

The literature on the effect of income on child development offers an illustration of a natural experiment. Costello, Compton, Keeler, and Angold (2003) estimated the effect using the natural experiment created by the opening of a casino in a community where they were conducting a study. Comparisons of children's development before and after the casino opening represent instrumental variables estimation: The casino affected family income but, by assumption, did not affect child development directly. One could argue about the validity of this assumption. A key point is that some assumption is necessary. The key question is whether the assumption embedded in the natural experiment is more or less plausible than ignorability.

Further reflection is required regarding the key assumptions identified above. Regarding SUTVA, is it reasonable to assume that the effect of an income increase does not depend on whether others in that community receive that increase as well? Regarding consistency, does the effect of an income increase not depend on its source (a lottery vs. some other source)?

Agnosticism: Do Unobservables Even Matter?

Many readers may feel left without options. They can either assume unobserved confounding away or can deal with the difficulties involved in adjusting for it. Perhaps an instrumental variable is not available for a problem of particular interest, or the assumptions involved in selection modeling are objectionable.

A third option, however, is available. That option involves assessing the potential effect of unobserved confounding, were such confounding to exist. Different versions of such assessments exist, and a full review is beyond the scope of this review. However, one such method can be illustrated here.

Rosenbaum (2002) originally proposed a method for assessing unobserved confounding that builds on a nonparametric test of treatment effects when the data have been grouped into matched pairs. Suppose one has 300 matched pairs and that a higher outcome is preferred. In that case, in the absence of a treatment effect, one would expect the treatment observation in each pair to have the better outcome half the time. Suppose that in the actual data, one observes 180 pairs in which the treatment outcome "wins." Can one reject the null hypothesis? If one assumes ignorability, then the probability of observing this outcome is less than .01. Such evidence would appear to strongly favor rejecting the null hypothesis.

However, one fears that unobserved factors remain that favor the treatment observations. Suppose those differences imply that the chance a treatment observations wins is actually .51. How strong is the evidence in that case? Still strong. The corresponding *p* value is less than .01. As one continues toward greater levels of imbalance, one's statistical confidence continues to erode. If the imbalance is such that the treatment is preferred in 55% of the pairs (even after matching), the significance level is reduced to .05.

How large is the actual confounding? Is the probability of .55 reasonable? One has to assess how likely any remaining confounding is—above and beyond the covariates included—and how large it is likely to be based on the nature of the other covariates included and, more generally, the outcome of interest.

A Special Word on Timing and Lagged Dependent Variables

Finally, this article considers one other way to adjust for unobserved confounding. One discusses this issue separately because it frequently involves the timing with which the key variables are measured. Many articles in developmental psychology rely on lagging one variable in an analysis, with the lagged variable representing the cause and the other variable representing the effect.

Temporal ordering may be quite informative or even essential in establishing causality (Stewart, 2003). Indeed, "temporality" is recognized as one of the key criteria for establishing causality. An early, influential article in epidemiology by A. B. Hill (1965) identified having causes precede the effect as a key standard for assessing whether a reported relationship identifies a cause-and-effect relationship (Rothman & Greenland, 1998). A seminal figure in causal inference, Judea Pearl (2000) stated in his classic book, *Causality: Models, Reasoning, and Inference*, that "temporal precedence is normally assumed to be essential for defining causation, and it is undoubtedly one of the most important clues that people use to distinguish causal from other types of associations" (p. 42).

This relationship should appear to be good news for a field much concerned with development, time, and timing. Many, if not

[13] This interpretation relies on assumed "monotonicity," which rules out counterintuitive behavior. In particular, if a woman moved from a low-wage to a high-wage state, her chance of working could not decrease (Kaufman, Kaufman, MacLehose, Greenland, & Poole, 2005).

most, articles appearing in key journals involve longitudinal data. A key limitation for developmental psychology is that timing is most useful when events are involved. For example, suppose one wants to know the relationship between my losing a tennis match and a bike accident in which I was involved. Knowing which event preceded the other can be quite informative. If I crashed my bike before the match, perhaps an injury impaired my ability to play effectively. On the other hand, if the accident was after the match, perhaps I was aggravated by the loss and riding my bike carelessly. Because events are involved, timing plays a key role in understanding causality.

Unfortunately, timing in developmental psychology does not typically involve events but, instead, measurements. Generally, the researcher understands that two concurrent processes are at work at any time point and simply measures one construct (e.g., parenting) before the other (e.g., children's behavior). Such timing may tell the reader something about the researcher (that he or she believes parenting causes bad behavior) but offers little insight into the processes of interest.

Nonetheless, timing of measurement may play a helpful role in understanding causal relationships and even in dealing with unobserved confounding. Consider a common situation in developmental research. Suppose that one is interested in the effect of parenting on children's behavior and that one recognizes that parenting may reflect children's behavior as well. In that instance, developmentalists frequently condition on a lagged value of the dependent variable. For example, one might regress the outcome at Wave 2 on the outcome at Wave 1 and a measure of the treatment of interest. (See the top panel of Figure 4.) Such analyses are problematic. As has long been known in economics and other fields, in the presence of autocorrelation (a relationship between unobservables over time), the resulting estimates have poor statistical properties (Allison, 1990; Liker, Augustyniak, & Duncan, 1985). If one considers the DAG, one can see that the correlated error terms open a backdoor path from the treatment to the outcome.[14] (Indeed, there are more complications than commonly recognized in developmental psychology. Generally, the models involved assume "stationarity"—that the covariance of errors over time is stable. This assumption depends on a set of initial conditions being met, namely, that the first observation represents a steady state observation of the developmental process [Arellano, 2003]. This assumption merits careful consideration by developmentalists. The nature of the initial conditions also influences the level of bias caused by autocorrelation.)

A panel data structure, however, can be quite helpful for eliminating some of forms of unobserved confounding. Consider the bottom panel of Figure 4. The model still presumes a panel data set, but now there is no causal effect of the outcome at Wave 1 on the outcome at Wave 2. Rather the two covary because they share common, time-invariant, unobserved characteristics. Because that unobserved characteristic also influences treatment status, there is a backdoor path between, for example, treatment and outcomes at Wave 2.

Because of the panel nature of the data set, one can effectively condition on the unobservables by giving each unit (an individual in this case) its own intercept (Arellano, 2003). (This model is known as the least squares dummy variable model, and the intercepts are known as *fixed effects*.)[15] Alternatively, one can remove the confounding by "differencing"—by subtracting one record from the record for the preceding period. One then can regress the change in the outcome on the change in the regressor. The key issue is that this method relies only on within-person changes over time: Time-invariant unobserved characteristics that confound comparisons across individuals are effectively removed. Also removed is any confounding association between time-varying unobservables and covariates that is stable over time (Arellano, 2003).

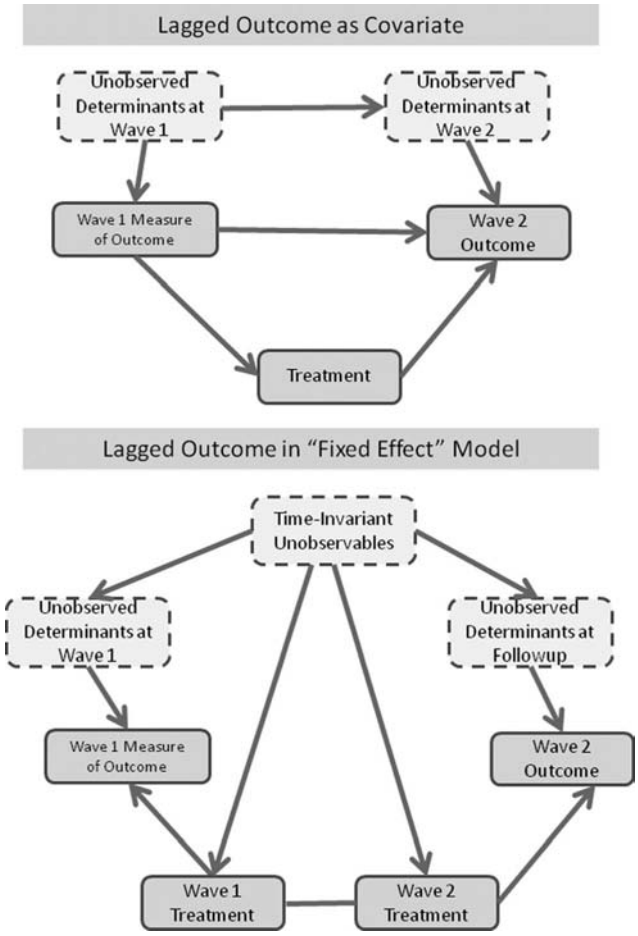

Figure 4. Lagged outcome as covariate.

Developmentalists have long distrusted change-score models, but the underlying technical reasons have been addressed (Rogosa,

[14] The solution is to find an instrumental variable for the lagged dependent variable. One can find a discussion of this issue in any elementary econometrics textbook, such as that by Kmenta (1971). This model has other problems, some conceptual. Generally, developmentalists do not believe in a state dependent, where today's behavior (e.g., parenting) causes future behavior. Rather they often believe merely that the forces that cause a process at one point are stable over time. This perspective is much more consistent with the alternative model in the bottom panel of Figure 4.

[15] There are some added complexities to using this model with categorical outcomes, such as logistic regression, but estimation is still possible (for details, see Chamberlain, 1980, and for an illustration, see Hoffman et al., 1993).

1988). The fundamental problem is that if the underlying constructs do not change, change scores are indeed unreliable. In that case, however, it is not clear what longitudinal model is really appropriate: Modeling problems run deeper than whether to use a change score.

The fixed-effects model does have other limitations. One is that the method depends on model specification. For example, the procedure depends on the fact that "lagged" treatment (at Wave 1) does not have a lingering effect on outcome at Wave 2 and that treatment does vary over time. This approach will not work for treatments that do not change over time, such as race, or work poorly for those that change little, such as parental education. Other limitations include the loss of power. Whether one adds dummy variables or differences the data, numerous degrees of freedom are lost. (The number lost is equal to the number of higher order units, such as individuals.)

One reviewer mentioned the possibility of random-effects or growth-curve models in this situation. It is important to note that the growth curve does little if anything to improve causal inference. The reason is that the model assumes the time-invariant unobservables are uncorrelated with observed variables, including the treatment. Their potential as confounders, therefore, is assumed away. One is left with the same set of assumptions as ordinary regression for identifying causal effects. (Indeed, the ordinary regression coefficient estimates themselves have good statistical properties under the random-effects assumptions [Greene, 2008].) Ignoring the nested structure of the data, however, results in incorrect standard errors and confidence intervals. The random-effects model remedies this problem.

Note that the fixed-effects estimation is not limited to just panel data; any nested structure permits a form of fixed-effects estimation. For example, comparisons of sisters involve girls nested within families (Foster, 1995; Hoffman, Foster, & Furstenberg, 1993). In that case, fixed-effects estimation represents a way to control for unobserved family-level effects.

An Illustration: Family Structure

Although a full empirical example is beyond the scope of this article, one can outline the steps involved in the analysis of a hypothetical question, the effect of family structure on children's outcomes. Table 1 provides an overview of these steps.

Table 1
Steps in Causal Inference

Step 1	Carefully define the treatment.
Step 2	Carefully think about the mechanism determining the treatment (i.e., the agents, motives, incentives, costs, and information shaping treatment choice).
Step 3	Define which treatment effect is of interest: (a) Address whether and how positivity applies. (b) Address whether and how stable unit treatment value assumption applies.
Step 4	Draw a directed acyclic graph that links treatment; the outcome and potential confounders, both observed and unobserved; and colliders.
Step 5	Select a statistical method for estimating the treatment effect under ignorability.
Step 6	Consider methods for relaxing or replacing the ignorability assumption: (a) Think carefully about possible instrumental variables. (b) Think carefully about other means of controlling for unobservables (such as fixed-effects estimation). (c) Calculate Rosenbaum (2002) bounds or other forms of sensitivity analyses.

Step 1

The first step is to carefully define the treatment. A primary thrust of the discussion above is to highlight the importance of thinking carefully about the treatment before one even obtains data or starts a software package. In the case of family structure, one might think of treatment as "living in a two-parent family." This choice, however, is not obvious; there are various aspects of family structure, such as the presence of other adults (e.g., a grandmother), and other aspects of the definition (e.g., are the two parents the biological parents?). Further complicating matters is the role of marriage and time. For example, does one want to differentiate treatment according to whether the parents are married or cohabiting? And does one want to add a time element? Does one mean living together now or most of the time? Or when the child was born? One could (and should) spend a considerable amount of time just defining the treatment. More refined treatments may indeed be more interesting (and realistic), but statistical power is reduced as observations are spread across more categories.

Note that the definition of the treatment has important implications for everything that follows, such as the plausibility of ignorability. More nuanced or distinct treatments involve more choices by more individuals. For example, one might differentiate single-parent families by whether a grandmother is present. However, in that case, one needs to think carefully about how the decision to live with one's daughter and grandchild is made and whether that choice reflects factors unobserved by the researcher, especially if those factors potentially influence outcomes of interest as well.

For this discussion, suppose treatment is simply binary: currently living with two biological parents or not (regardless of whether the parents are married). *Currently*, in this case, refers to the timing of the outcome measurement. This choice also reflects trade-offs in terms of facilitating and complicating causal inference. Lagging family structure, for example, might represent a means of clarifying the direction of causality. However, doing so means that the lagged value is further removed from the actual conditions in which the child lives and is presumably less salient.

Step 2

The second step involves thinking carefully about the mechanisms that determine the treatment—in this case, the reasons why families form and dissolve. As Rubin (2008) noted, the potential outcomes framework includes both a model of the effects of treatment and a model of treatment "assignment." His more recent work emphasizes identifying the information available to those making the decision as a key step in causal inference. As noted above, this theme has long been emphasized in the economic literature where "information" refers to the incentives that shape the behavior of the decision maker. This approach reflects the broader economic framework in which consumers, workers, or firms optimize their behavior, and this framework may very well not fit many or most developmental processes (Foster, 1993,

2002). But it surely would fit some, such as the decision to enroll a child in preschool, and in any case, psychologists would benefit from making it clear who makes the decisions that shape treatments, their goals as decision makers, and the information they use to make those decisions.

In many instances, developmental psychologists should look outside their discipline to understand treatment choice. For example, in the case of family structure, a large literature in demography and sociology considers how and why families form. A key factor may be the availability of marriageable men (Guzzo, 2006). Articles in developmental psychology often ignore such contextual forces. Causal inference is inherently interdisciplinary.

Step 3

The third step involves defining the treatment effect of interest: the ATT, the ATE, the ATU, or yet another treatment effect. This choice should reflect the research questions of interest. For example, if one is interested in the effect of extending preschool to more children, the ATU is of interest.

This task should stimulate the researcher to think more generally about whether the potential outcomes framework applies. Two aspects of that framework—SUTVA and positivity—seem particularly salient to family structure. Both should spur the analyst to think carefully about the population for which one is estimating the treatment effect. One might be interested in minority children living in poor families in very poor neighborhoods. In those neighborhoods, virtually none of the children may be living with two parents; at that point, it is important to recognize that positivity may not hold, and the only way to identify the effect of family structure would be through extrapolating from other communities. Such an extrapolation would require that the processes shaping the causes and consequences of family structure do not fundamentally vary across communities.

The issue of SUTVA is also relevant. Does a child's outcome depend on his or her own family structure or on that of his peers as well? Does the presence of a father in a poor neighborhood help all children, not just his own (Wilson, 1987)? One can imagine a multidimensional treatment defined by whether one lives in a two-parent family and the prevalence of such families in one's environment. Such an analysis would have to rely on new methodological tools of causal inference (discussed below). To keep matters simple, for the purposes of this discussion, one defines the treatment by whether a child him- or herself lives in a two-parent family. (The discussion below considers the neighborhood prevalence of two-parent families as a covariate.)

With some strong limitations, therefore, the potential outcomes framework would seem to apply to the hypothetical illustration above. In that case, ATT represents the benefits of living in a two-parent family for those actually living in such families; the ATU represents the benefits for those not living in those families (were they moved to a two-parent family). The two may not be equivalent. Suppose, for example, that single mothers are less educated (because education levels influence the likelihood of giving birth out of wedlock). In that case, the presence of the child's father may be especially beneficial (depending on the causal mechanism) for those women and their children. In that case, the benefits of living in a two-parent family might be even larger for those not currently living in those families than for those children who are (i.e., larger than the loss the latter would suffer were they moved to a single-parent family).[16] One could argue that all three effects are interesting and plan analyses to differentiate the different effects. On the other hand, one might be interested in the potential effects of a program to encourage residence by biological fathers in poor families.[17] In that case, the ATU would seem most applicable: One would like to know what the benefits of treatment would be for children currently living apart from their fathers.

Step 4

The fourth step involves drawing a DAG that links treatment; the outcome and potential confounders, both observed and unobserved; and colliders (see Figure 5). That figure can help establish several key features of the analysis, such as the likelihood that ignorability applies and, if so, with which covariates. What does ignorability mean in the context of this empirical example? Equation 5 implies that children currently living in single-parent families (the untreated) would fare no better or no worse if living with two parents than would those actually living in those families (the treated) who have the same values for the covariates (e.g., the level of maternal education). Similarly, Equation 6 stipulates that children currently living in two-parent families would fare no better or no worse if living in a single-parent family than would those actually living in those families who have the same values for the covariates (e.g., the level of maternal education).

As discussed above, the plausibility of ignorability depends on the choice of covariates, and those choices should reflect prior research. Broadly put, the goal is to condition on shared determinants of family structure and child outcomes. However, as the discussion of colliders makes clear, some covariates may create more problems than they solve.

In drawing a DAG for this exercise, one might start with an obvious choice, maternal education. Maternal education seems like a certain confounder. Less educated women may be more likely to form single-parent families, and they may spend less time reading with their children as well. One concern that arises immediately is whether education is correlated with unobserved factors, such as ability. More educated mothers also may be more capable, and they may influence child outcomes as well. In this case, the DAG reveals that as long as those unobserved factors do not influence family structure directly, no backdoor path is opened. That education reflects ability and is included as a covariate prevents a relationship between ability and family structure. (Of course, if ability affects family structure directly, apart from education, then potential confounding remains.) An essential element of the DAG is that the arrow runs from education to family structure and not vice versa.

Other confounders are tricky, especially those that represent mediators as well, such as family income. In many instances,

[16] Because this variation is related to a variable one can measure, it does not inherently violate ignorability.

[17] In that case, still further reflection on the causal model would be required. Consistency, for example, would require that the effect of a father's presence would not depend on whether he decided to live with his children on his own or whether that choice was encouraged or even coerced by a government policy.

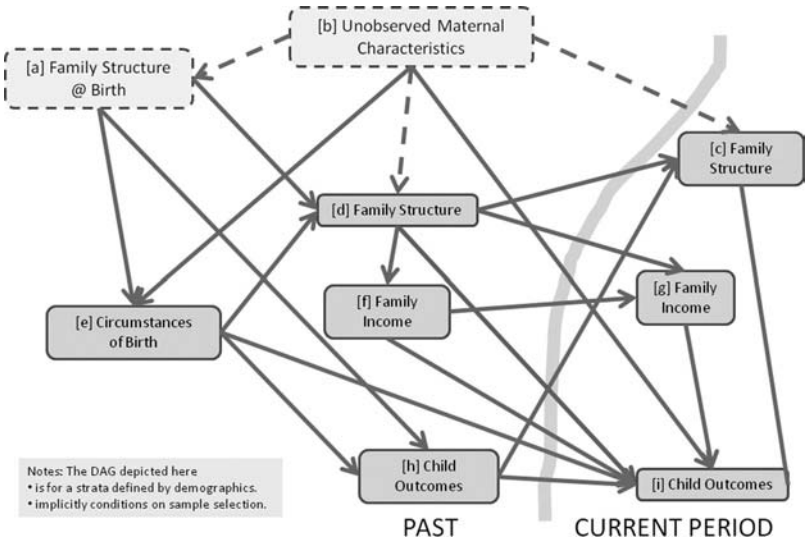

Figure 5. Family structure directed acyclic graph (DAG), before conditioning.

researchers are interested in controlling for poverty status as they recognize that family structure covaries with poverty status for a variety of reasons (e.g., poverty is a stressor on relationships). One implication is that the coefficient on family structure now represents a direct effect, the effect not mediated by poverty status (Baron & Kenny, 1986). A more serious problem is statistical: The mediator is also a collider, and conditioning on it creates an association between family structure and unobserved correlates of poverty. (Baron and Kenny, 1986, implicitly assumed ignorability applied to the mediator as well.) Most researchers do not appear to realize this, yet ignorability as it applies to the mediator needs equal attention to ignorability as it applies to the treatment (Sobel, 2008). Adding a sensible covariate to control for confounding can create other potential problems.

What about the prevalence of two-parent families in the neighborhood? One decided to treat that variable as a confounder and add it as a covariate. The implications must be assessed in light of the potential confounding of this covariate as well: Families living in neighborhoods with few two-parent families may differ in a range of other ways. The effect of neighborhood family structure is likely correlated with other neighborhood characteristics (e.g., such as employment opportunities), some of which are unmeasured. As a DAG makes clear, the effects of such confounding on a covariate likely spill over to the treatment of interest.

What are the implications for the question at hand? The key issue involves the relationship between the structure of the child's own family and neighborhood family structure. If the arrow runs from family structure to the neighborhood variable, then the latter is a collider, and conditioning for it creates a confounding relationship between the unobserved determinants of neighborhood residence and family structure. On the other hand, if the neighborhood family structure influences the child's own family structure, then including the neighborhood family structure as a covariate makes sense.

Given a suitable set of covariates to include (and exclude), is the ignorability assumption plausible? Assessing that assumption using the data is difficult or impossible. Ultimately, the plausibility depends on one's understanding of how individuals select the treatment (i.e., on the factors that influence family structure). Ignorability also depends on how well the data analyzed maps onto that theory. If key determinants of family structure are not measured, then the potential for confounding remains.

Step 5

The fifth step involves selecting a statistical method for estimating a treatment effect under ignorability. As discussed, one might use any of several alternatives, including regression, stratification, or matching. The literature is not especially helpful in guiding the researcher to one method over others. The discussion above highlights the limitations of ordinary regression, but careful consideration of the underlying assumptions can redress many of those problems. What may matter more than the choice of method is the quality of implementation. Regardless of the method, for example, one should give careful attention to covariate balance and the amount of overlap in the support (or range) of the covariates. As one has seen, the propensity score can facilitate the diagnostic process.

Step 6

The sixth and final step involves analyses that relax or replace the ignorability assumption. Two issues are key. First, one might consider whether additional information is available that can be used to understand the determinants of the treatment, such as an instrumental variable. In the case of family structure, one might consider the generosity of cash assistance that influences family structure but not children's outcomes directly (Foster & Hoffman, 2001). Finding instrumental variables can be difficult, but in general, if one does not look, one is sure not to find them.

Another strategy would rely on features of the data. For example, researchers have compared children of sisters as a means of estimating the effect of mother's age at first birth on children's outcomes (Geronimus, Korenman, & Hillemeier, 1994). Such

fixed-effects analyses were made possible by the availability of a large national database that included sisters and their children.

As a final means of addressing the role of unobserved confounding, one might consider sensitivity analyses, such as those proposed by Rosenbaum (1988) and others (e.g., Imbens, 2003). Such analyses seem especially important in instances where methods for directly dealing with unobserved confounding are not available.

Discussion and Conclusion

Recent developments offer exciting possibilities for developmental psychologists to further unravel key causal processes and their origins and consequences. To do so, however, developmentalists need to realize that "the days of 'statistics can only tell us about association, and association is not causation' seem to be permanently over" (Rubin, 2004, p. 161). One naturally wonders whether this claim is overstated. A large and growing literature considers whether observational methods can reproduce the results of experiments (for a review, see Cook, Shadish, & Wong, 2008). At this point, a balanced assessment of this issue is that "it depends." The answer largely depends on the careful checking of data and model specification. The former have been highlighted here, but one should not neglect the latter. In responding to a well-known "failure" of propensity score methods (J. A. Smith & Todd, 2005), Dehejia (2005) demonstrated that propensity score methods can perform satisfactorily. However, as the author noted, his findings are quite sensitive to the specification of the propensity score equation (see also J. L. Hill, Reiter, & Zanutto, 2004).[18]

Best practice at this point likely involves presenting multiple estimates involving alternative approaches. To the extent the alternative estimates cohere, one is left reassured that his or her findings are not spurious outcomes of arbitrary modeling choices. To the extent they do not, the resulting uncertainty may represent part of the costs of the inability or failure to randomly assign individual to treatment. Consistency across alternative methods is one standard by which analyses of observational data can be assessed (Rosenberger & Lachin, 2002).

The discussion above touches on the benefits of experiments for establishing the true causal effect. Of course, for one to implement an experimental design, the characteristic or program has to be manipulable, both conceptually and as a practical matter. However, the discussion here is not without implications for experimenters. Particularly salient is the framework for understanding and naming the causal effect of interest. In many experiments, the estimated effect (e.g., a clinical trial) may not be the ATT or ATU for a community-based population. In that sense, the benefits of an experimental design in terms of internal validity may be exceeded by the costs in terms of external validity. This article is hardly the first to make this point, but the discussion here highlights that a strategy that blends experimental and observational approaches may best answer important questions in developmental psychology. Observational data, for example, might provide a way to weight variation in causal effects obtained from the experiment to better reflect the likely effects of the program in a broader population. Such approaches have been used successfully in medicine (Goldman et al., 1999; Phillips et al., 2000; Tice et al., 2001; Weinstein et al., 1987).

Plausibility and Other Values

This article identifies plausibility as a key criterion for selecting one's approach to causal inference. Likely no researcher would aim for implausible inference, yet the literature on children and families is filled with work that makes implausible causal claims, such as television viewing causing attention-deficit disorder (Christakis et al., 2004). A bit of reflection could deter much of this work (Foster & Watkins, 2010). In other situations, the choices are more complex, and plausibility offers some guidance. For example, adjusting for family income cannot prove whether the association between television and attention problems is causal, but it can make such claims more plausible.

Plausibility is not the only value, however. Model simplicity also may be valued. Simplicity may indeed have a variety of benefits, such as increasing the policy impact of one's work. Plausibility and simplicity, however, are not synonymous. Econometric models are often quite complex yet may be more plausible because they do not assume ignorability. The nature of the underlying problem may result in increasing plausibility at the cost of greater complexity. A simple, plausible method may not exist for some empirical problems. Again, this trade-off may be the true cost of an inability to randomize.

Another value is robustness, and its importance varies across fields. Biostatisticians place great value on getting the same answer with different methods, and they are distrustful of a truth that is only revealed by a complex model. On the other hand, economists are generally much more comfortable with complex models if those models are grounded in economic theories of behavior.

Another value at work involves the use of and returns to the research. As noted above, one can generate ranges for treatment effects under very weak assumptions (Manski, 2007), and in some instances, simply knowing that a treatment does no harm is useful. For example, one might be interested in the effects of a low-cost intervention that participants inherently enjoy (e.g., a mentoring program). In that case, using an elaborate statistical model to produce a point estimate may be unnecessary.

Economists and others employing complex models often emphasize that models can generate insights not available with simpler methods. Selection models, for example, can generate estimates not available even in experiments. Even as noneconomists embrace instrumental variables estimation, there is some backlash in economics against them. Some would argue that the local average treatment effect describes a treatment effect that is just not very useful to policymakers (Deaton, 2009). As noted, however, in some instances, the local average treatment effect is *the* treatment effect of interest.

In the end, some judgment is required to balance these considerations. The particular balance for a given problem depends on the specifics of the substantive problem. Ignorability may indeed be implausible for understanding job training; workers may behave rather rationally in deciding about such training and may indeed be better informed than researchers in making those decisions. On

[18] There are other practical issues as well, ones that have not been highlighted here. Heckman, Ichimura, and Todd (1997) considered the circumstances under which one might draw a comparison group from a secondary data source. They highlighted the role of differences in survey instrumentation across data sources as a source of bias.

other hand, decisions about teenage pregnancy (and sexuality) may be less rational, and ignorability may hold.

The final balance also will reflect disciplinary training. At this point developmentalists may not be willing or even able to embrace and employ econometric tools. That individuals maximize the returns to different choices seems natural to economists but may seem rather alien to psychologists. (Indeed, psychologists have produced some of the strongest criticisms of the rational choice model [V. L. Smith, 1991].) Nonetheless, it seems essential that psychologists give considerably more thought to why people select the treatments they do. For example, why do parents enroll their children in day care? How important are the benefits for parents versus benefits for the children? If the choice is driven by perceived benefits, how accurate are those perceptions and how are they formed? How many of these factors are measured in the available data? The lesson of the econometric approach is that one cannot understand the effects of day care without understanding these issues.

The Causal Challenge for Developmental Psychologists

If Rubin (2004) is correct that associations can produce causal insights, then developmentalists arguably are lagging behind in meeting the challenge posed. Causality represents a challenge for any quantitative social scientist, and developmental psychologists do bring strengths to the task. For example, better methods are not a substitute for strong theory; in the best analyses, the two work in tandem. As the DAG makes clear, selecting covariates involves hard choices that theory can inform.

If developmental psychologists are to meet this challenge, business as usual will need to change. For example, too often studies in developmental science involve unrepresentative samples. Such samples may have been necessary in studies implementing randomization. However, such data are inadequate—or arguably fundamentally flawed—for causal inference. A prime example of this problem is the National Institute of Child Health and Human Development's Study of Early Child Care and Youth Development. Participants were an unrepresentative subsample of those giving birth at major medical centers. Such a design may have had benefits (such as reducing study costs), but regardless of the reasons, such sample selection influences all analyses that follow. Simply acknowledging the sample as a limitation on external validity and referring readers to appendices is inadequate. Even a decade after the Early Child Care study began, sampling weights still have not been developed. In fairness to those investigators, their work represents some of the best work in the field, and their data are typical. After all, there are many strange samples in developmental psychology, and the issue receives far too little attention. Data from the multimillion-dollar Fast Track evaluation were described incorrectly for more than a decade. (The study long described the participants as from the most aggressive decile in the population sampled; subsequent analyses revealed that the children were drawn from the top three deciles and that the procedures varied across the four study sites [compare Conduct Problems Prevention Research Group, 1999, with Conduct Problems Prevention Research Group, 2007].)

The point here is not to admonish a few investigators for anomalous behavior. Rather these studies represent the norms and values of the field. The lesson here is that sampling issues are not secondary or peripheral: Sample representativeness is central to the task of understanding the processes of interest. One can best see this issue using the DAG. In effect, the sampling frame for the Early Child Care study adds a variable to the DAG: delivering in a major medical center. That variable would seem to reflect a variety of factors, both observed and unobserved. For some questions, the variable may serve as a collider, and conditioning on it has implications for all causal questions.

To respond to the challenge of causal inference, methodology in developmental psychology will have to improve, in both depth and breadth. Authors need to spend more time describing and justifying their methods and models. Currently, authors race from a rather lengthy theory section of their article to their results, describing their measures in great detail but pausing only briefly to describe their model. Often that model is identified only by name (or even an acronym) with little specificity about the assumptions made. Describing a model in detail requires mathematical specificity. For that reason, the models need to be written in their mathematical form. Doing so will require a substantial number of developmentalists to achieve mathematical literacy. Understanding the expectations operator well enough to assess Equations 5 and 6, for example, is not difficult, and any quantitative social scientist should be able to do so. At this point, many developmental psychologists appear to me as lacking these skills.[19]

Responding to the causal challenge also will involve other changes in practice. For example, it may appear that regression takes a beating in the discussion above. However, good causal inference is indeed possible with regression. (After all, regression is a matching estimator.) But better causal inference requires careful model checking of the sort that is seldom done in any social science (e.g., checking for outliers). Especially when working with large data sets, developmentalists would do well to think in terms of a saturated model. Such models would incorporate key covariates more flexibly. For example, in including parental education, years of schooling is likely a poor choice vis-à-vis including a series of dummy variables reflecting alternative levels of schooling.

Developmentalists need better method training. Currently, advanced training in methods largely focuses on a niche area of applied statistics. Judging from the submissions to *Developmental Psychology*, one would see that much of developmental psychology centers on generalized linear latent variable models, such as those implemented in Mplus (Muthén & Muthén, 1998). These models are highly parameterized and assumption-laden; in many cases, the models are fundamentally underidentified. For example, a model with $k + 1$ latent classes can be represented with k latent factors—the data offer no way to distinguish the two (Bartholomew, 1987). The best case situation is that such models offer improved description of developmental phenomena. The worst case is that the resulting estimates are ones chosen ad hoc from a large set of equally valid alternative estimates.

As discussed above, complex models are not bad models. What seems clear, however, is that a necessary condition for using such models is a full understanding of their assumptions. Such an understanding extends beyond the options of a software package to

[19] This observation also reflects my 5 years as an editor at *Developmental Psychology* (for further discussion, see Foster, in press).

include a mathematical representation of the model and its statistical properties. Similarly, a diagram (be it a path diagram or a DAG) is no substitute for a mathematical model.

Finally, unobservables likely play a far greater role than developmentalists generally believe. For example, in supplemental tabulations based on data from the Child Development Supplement of the Panel Study of Income Dynamics, I determined that the correlation between siblings' scores on a well-known IQ test is .67: Siblings share the genes of their parents and a great deal of their home environment. The wrinkle is that standard measures of family background explain little of the relevant factors. For example, I generated estimates of the sibling fixed effects and regressed them on a wide array of maternal characteristics, including mother's age at first birth, her education, her family's income at the time of the births, enrollment in the Medicaid program, and her own IQ. Admittedly, the estimated fixed effects are noisy, but this wide range of covariates explained less than 6% of the variance sibling share. And the list of covariates included is far larger than the list included in the standard article in developmental psychology. Ignorability may be plausible, but the list of covariates required is likely far greater than common practice suggests.

At the same time, one can see that including covariates may cause unanticipated problems. The DAG may seem more useful for ruling out covariates than validating their inclusion. Beginning to think about the process shaping the explanatory variables does indeed open a can of statistical and conceptual worms. A reasonable conclusion to draw from the DAG is that the coefficients on all the covariates need to represent causal effects or none do. In that case, the analyst may feel rather conflicted: More covariates are needed to block backdoor paths, yet covariates may open other backdoor paths. The net effect may be the feeling that in picking covariates, less is more. If so, unobserved confounding may seem unavoidable. In that case, developmental psychologists will need to turn to methods with which they currently have little experience or training.

Recommendations for Further Reading

In this article I have tried to balance breadth versus depth. For full breadth and depth, one would need to consult a range of additional materials. A first area for further reading involves a third framework for causal inference neglected here, the sufficient component causal model. This perspective recognizes that cause-and-effect relationships are often complex and combine multiple causal mechanisms. From this perspective, no single cause is necessary or sufficient for an effect to occur. This perspective emphasizes that risks work in combination with one another and makes it clear that a given cause may have different effects depending on the presence of other risk factors. As a result, there is no *singular* causal effect. Rather the effect of one construct on another is likely to vary across individuals and with other constructs at work. This approach fosters a consideration of mechanisms that complements both of the frameworks considered here (Greenland & Brumback, 2002; Prince, Stewart, Ford, & Hotopf, 2003; Susser, Schwartz, Morabia, & Bromet, 2006).

Other areas for further reading involve extensions to the counterfactual model. For example, the SUTVA assumption may be untenable in some cases. One could argue that intensive intervention to improve a child's behavior might have spillover effects onto other children in the same classroom. In those instances, one could expand the model to include more elaborate treatment effects (Hong & Raudenbush, 2006; Hudgens & Halloran, 2008). Other types of policy spillovers are possible, especially for broader, large-scale treatments affecting children and families. For example, an effort to provide better training to day care providers in a community may influence the market for day care and affect the return to training itself. Economists label these phenomena as *general equilibrium effects* (Brock & Durlauf, 2001, 2002).

The presentation here devotes considerable attention to propensity scores. As discussed, the emphasis on this method reflects its growing popularity rather than any role it plays in revolutionizing causal inference. The emphasis is also on propensity scores because they generalize in many ways to more complex treatments, such as ordered and unordered categorical, continuous and dynamic treatments (Hirano & Imbens, 2004; Imbens, 2000). In some instances, these newer methods allow one to assess ignorability more effectively or to relax the specific form required. Models of dynamic treatment, marginal structural models, assume a weaker form of ignorability (sequential ignorability) that only the *current* state of the treatment is independent of current potential outcomes. Past values of the treatment may be confounded with unobservables (Bodnar, Davidian, Siega-Riz, & Tsiatis, 2004; Hernán, Brumback, & Robins, 2000; Robins, Hernán, & Brumback, 2000). Dynamic treatments have promising potential for understanding developmental phenomena (such as the effect of family structure) and are important in their own right (Foster & Gibson Davis, 2010). The marginal structural model is only one of several approaches for examining dynamic treatments (Abbring & van den Berg, 2004; Heckman & Navarro, 2007; Murphy, 2003; Murphy, van der Laan, Robins, & Conduct Problems Prevention Research Group, 2001).

Another key area for developmentalists involves mediation. In many instances, developmentalists are quite interested in mediators as a means of understanding the causal mechanism linking the treatment to the outcomes of interest. As noted above, it appears that much of current empirical work in this area involves a collider and so is incorrect (Sobel, 2008). At the very least, researchers should examine whether ignorability applies to the mediator. Further reading, however, would reveal that new methods have been developed for assessing the role of mediators without assuming that individuals are effectively randomly assigned to levels of mediation (Barnard et al., 2003; Frangakis & Rubin, 2002; VanderWeele, 2008; Zhang, Rubin, & Mealli, 2008).

Finally, my own assessment is that the sensitivity analyses are particularly promising, especially in circumstances where no instrumental variable is available. One then is left with the broader framework of the Heckman selection model. Methodological advances continue to weaken the assumptions required for those models. For example, semi- and nonparametric approaches have been developed (Das, Newey, & Vella, 2003). Still, noneconomists in particular may find that approach objectionable for conceptual or methodological reasons. For that reason, the sensitivity analyses seem particularly useful: If unobserved confounding is a possibility, but findings are robust to its presence, then no argument about the selection models is necessary. In discussing sensitivity analyses, I have highlighted the work of Rosenbaum (2002), but a range of these methods are becoming available (Altonji, Elder, & Taber, 2005; Imbens, 2003).

A Positive Note

The timing is right for developmentalists to turn to causal inference. Arguably, their interest in public policy has never been greater; their influence on policy issues, never stronger. In most circumstances, those issues raise causal questions. In making decisions that affect the lives of real children and families, associations are inadequate.

Many of the methods described here are finding their way into developmental science, such as fixed-effects estimation and propensity-score-based and instrumental variable methods (Arnold, McWilliams, & Arnold, 1998; Boyle, 2002; Boyle et al., 2004; Foster & Kalil, 2007; Foster & McLanahan, 1996; Gennetian et al., 2008; Georgiades, Boyle, & Duku, 2007; Gordon, Chase-Lansdale, & Brooks-Gunn, 2004; Haviland, Nagin, Rosenbaum, & Tremblay, 2008; J. L. Hill, Brooks-Gunn, & Waldfogel, 2003; J. L. Hill, Waldfogel, Brooks-Gunn, & Han, 2005; Hong & Yu, 2008; Jenkins, Simpson, Dunn, Rasbash, & O'Connor, 2005; Lewin-Bizan, 2006; Love et al., 2005; Stormshak, Bierman, Bruschi, Dodge, & Coie, 1999; Yoshikawa, Magnuson, Bos, & Hsueh, 2003). Still, the potential these methods offer will not be reached if they are not embedded in more careful thinking about causal inference. Conditioning on a collider with a flexible, propensity-score-based method is hardly an improvement over including it as a regressor in ordinary least squares. It may even produce worse estimates.

Other developments are promising. Now, more than ever, large data sets include developmentally relevant measures. Many of the methods described above are data hungry; their good statistical properties, asymptotic. Furthermore, many of these data sets involve representative samples. For reasons described above, such samples set the foundation for good causal inference.

Finally, many of the methods described may be rather unfamiliar to developmentalists or at least not at the top of their toolbox (e.g., instrumental variables). However, one can be confident that once they start thinking about using those tools, developmental psychologists will naturally encounter applications, such as discontinuities in eligibility and other natural experiments. Many of the best applications of these tools currently involve topics of interest to developmentalists, such as the effect of Head Start, but have been conducted by nonpsychologists (Ludwig & Miller, 2007). Without some effort to catch up to the other social sciences on methodological issues such as causal inference, psychologists likely will remain marginalized in debates surrounding these and other issues affecting children and families.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics, 113*(2), 231–263.

Abbring, J. H., & van den Berg, G. J. (2004). Analyzing the effect of dynamically assigned treatments using duration models, binary treatment models, and panel data models. *Empirical Economics, 29*(1), 5–20.

Abbring, J. H., & Heckman, J. J. (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6, pt. 2, pp. 5145–5303). Amsterdam, the Netherlands: North-Holland.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In C. C. Clogg (Ed.), *Sociological methodology* (Vol. 20, pp. 93–114). Oxford, England: Blackwell.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy, 113*(1), 151–184.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Arellano, M. (2003). *Panel data econometrics: Advanced texts in econometrics.* New York, NY: Oxford University Press.

Arnold, D. H., McWilliams, L., & Arnold, E. H. (1998). Teacher discipline and child misbehavior in day care: Untangling causality with correlational data. *Developmental Psychology, 34*(2), 276–287.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association, 98*(462), 299–323.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* New York, NY: Oxford University Press.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity.* Hoboken, NJ: Wiley-Interscience.

Blundell, R., & Dias, M. C. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources, 44*(3), 565–640.

Bodnar, L. M., Davidian, M., Siega-Riz, A. M., & Tsiatis, A. A. (2004). Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology. *American Journal of Epidemiology, 159*(10), 926–934.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90*(430), 443–450.

Boyle, M. H. (2002). Home ownership and the emotional and behavioral problems of children and youth. *Child Development, 73*(3), 883–892.

Boyle, M. H., Jenkins, J. M., Georgiades, K., Cairney, J., Duku, E., & Racine, Y. (2004). Differential-maternal parenting behavior: Estimating within- and between-family effects on children. *Child Development, 75*(5), 1457–1476.

Brock, W. A., & Durlauf, S. N. (2001). Interactions-based models. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3297–3380). Amsterdam, the Netherlands: North-Holland.

Brock, W. A., & Durlauf, S. N. (2002). A multinomial-choice model of neighborhood effects. *American Economic Review, 92*(2), 298–303.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies, 47*(1), 225–238.

Christakis, D. A., Zimmerman, F. J., DiGiuseppe, D. L., & McCarty, C. A. (2004). Early television exposure and subsequent attentional problems in children. *Pediatrics, 113*(4), 708–713.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295–313.

Cole, S. R., & Frangakis, C. E. (2008). *On the consistency statement in causal inference: A definition or an assumption.* Unpublished manuscript.

Conduct Problems Prevention Research Group. (1992). A developmental and clinical model for the prevention of conduct disorders: The Fast Track Program. *Development and Psychopathology, 4*(4), 509–527.

Conduct Problems Prevention Research Group. (1999). Initial impact of the Fast Track prevention trial for conduct problems: I. The high-risk sample. *Journal of Consulting and Clinical Psychology, 67*(5), 631–647.

Conduct Problems Prevention Research Group. (2007). Fast Track randomized controlled trial to prevent externalizing psychiatric disorders: Findings from grades 3 to 9. *Journal of the American Academy of Child & Adolescent Psychiatry, 46*(10), 1250–1262.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724–750.

Costello, E. J., Compton, S. N., Keeler, G., & Angold, A. (2003) Relationships between poverty and psychopathology: A natural experiment. *Journal of the American Medical Association, 290*(15), 2023–2029.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies, 70*(1), 33–58.

Dearing, E., McCartney, K., & Taylor, B. A. (2006). Within-child associations between family income and externalizing and internalizing problems. *Developmental Psychology, 42*(2), 237–252.

Deaton, A. S. (2009). *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development* (NBER Working Paper No. 14690). Cambridge, MA: National Bureau of Economic Research.

Dehejia, R. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics, 125*(1–2), 355–364.

Foster, E. M. (1993). How sociologists should treat Becker's *Treatise on the Family*. *Sociological Forum, 8*(2), 317–329.

Foster, E. M. (1995). Why teens do not benefit from work experience programs: Evidence from brother comparisons. *Journal of Policy Analysis and Management, 14*(3), 393–414.

Foster, E. M. (2002). How economists think about family resources and child development. *Child Development, 73*(6), 1904–1914.

Foster, E. M. (2010). The ∪-shaped relationship between complexity and usefulness: A commentary. *Developmental Psychology, 46*(6), 1760–1766.

Foster, E. M., & Gibson Davis, C. (2010). *Marginal structural models and the effect of family structure.* Manuscript submitted for publication.

Foster, E. M., & Hoffman, S. D. (2001). The young and the not quite so young: Age variation in the impact of AFDC benefits on nonmarital childbearing. In L. L. Wu & B. Wolfe (Eds.), *Out of wedlock: Causes and consequences of nonmarital fertility* (pp. 173–201). New York, NY: Russell Sage Foundation.

Foster, E. M., & Kalil, A. (2007). Living arrangements and children's development in low-income White, Black, and Latino families. *Child Development, 78*(6), 1657–1674.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 3*(1), 249–260.

Foster, E. M., & Watkins, S. (2010). The value of reanalysis: TV viewing and attention problems. *Child Development, 81*(1), 368–375.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58*(1), 21–29.

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44*(2), 381–394.

Georgiades, K., Boyle, M. H., & Duku, E. (2007). Contextual influences on children's mental health and school performance: The moderating effects of family immigrant status. *Child Development, 78*(5), 1572–1591.

Geronimus, A., Korenman, S., & Hillemeier, M. M. (1994). Does young maternal age adversely affect child development? Evidence from cousin comparisons in the United States. *Population and Development Review, 20*(3), 585–609.

Goldman, L., Coxson, P., Hunink, M. G., Goldman, P. A., Tosteson, A. N. A., Mittleman, M., . . . Weinstein, M. C. (1999). The relative influence of secondary versus primary prevention using the National Cholesterol Education Program Adult Treatment Panel II guidelines. *Journal of the American College of Cardiology, 34*(3), 768–776.

Gordon, R. A., Chase-Lansdale, P. L., & Brooks-Gunn, J. (2004). Extended households and the life course of young mothers: Understanding the associations using a sample of mothers with premature, low birth weight babies. *Child Development, 75*(4), 1013–1038.

Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Greenland, S., & Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology, 31*(5), 1030–1037.

Greenland, S., & Pearl, J. (2008). Causal diagrams. In S. Boslaugh (Ed.), *Encyclopedia of epidemiology* (Vol. 1, pp. 149–156). Los Angeles, CA: Sage.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*(1), 37–48.

Guzzo, K. B. (2006). How do marriage market conditions affect entrance into cohabitation vs. marriage? *Social Science Research, 35*(2), 332–355.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201–209.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin. *Statistics in Medicine, 27*(12), 2050–2054.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York, NY: Springer.

Haviland, A., Nagin, D. S., Rosenbaum, P. R., & Tremblay, R. E. (2008). Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental Psychology, 44*(2), 422–436.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–294.

Heckman, J. J., & Navarro, S. (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics, 136*(2), 341–396.

Heckman, J. J., & Vytlacil, E. J. (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6, pt. 2, pp. 4779–4874). Amsterdam, the Netherlands: North-Holland.

Heckman, J. J., & Vytlacil, E. J. (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6, pt. 2, pp. 4875–5143). Amsterdam, the Netherlands: North-Holland.

Hernán, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology, 11*(5), 561–570.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295–300.

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4), 730–744.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L.

Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 49–60). New York, NY: Wiley.

Hill, J. L., Waldfogel, J., Brooks-Gunn, J., & Han, W. (2005). Maternal employment and child development: A fresh look using newer methods. *Developmental Psychology, 41*(6), 833–833.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). New York, NY: Wiley.

Ho, D. E., Imai, K., King, G., & Stuart, E A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236.

Hoffman, S. D., Foster, E. M., & Furstenberg, F. F., Jr. (1993). Reevaluating the costs of teenage childbearing. *Demography, 30*(1), 1–13.

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

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901–910.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology, 44*(2), 407–421.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association, 103*(482), 832–842.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, 93*(2), 126–132.

Jenkins, J., Simpson, A., Dunn, J., Rasbash, J., & O'Connor, T. G. (2005). Mutual influence of marital conflict and children's behavior problems: Shared and nonshared family risks. *Child Development, 76*(1), 24–39.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods, 6*(2), 115–134.

Kaufman, S., Kaufman, J. S., MacLehose, R. F., Greenland, S., & Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine, 24*(11), 1683–1702.

Kmenta, J. (1971). *Elements of econometrics.* New York, NY: Macmillan.

Lewin-Bizan, S. (2006). *Identifying the associations between child temperament and father involvement: Theoretical considerations and empirical evidence* (Center for Research on Child Wellbeing Working Paper No. 2006-24-FF). Princeton, NJ: Princeton.

Liker, J. K., Augustyniak, S., & Duncan, G. J. (1985). Panel data and models of change: A comparison of first difference and conventional two-wave models. *Social Science Research, 14*(1), 80–101.

Love, J. M., Kisker, E. E., Ross, C., Raikes, H., Constantine, J., Boller, K., . . . Vogel, C. (2005). The effectiveness of Early Head Start for 3-year-old children and their parents: Lessons for policy and programs. *Developmental Psychology, 41*(6), 885–901.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics, 122*(1), 159–208.

Manski, C. F. (1997). The mixing problem in programme evaluation. *Review of Economic Studies, 64*(4), 537–553.

Manski, C. F. (2007). *Identification for prediction and decision.* Cambridge, MA: Harvard University Press.

Manski, C. F., & Nagin, D. S. (1998). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociological Methodology, 28*(1), 99–137.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research, 35*(1), 3–60.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* New York, NY: Cambridge University Press.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B. Statistical Methodology, 65*(2), 331–366.

Murphy, S. A., van der Laan, M. J., Robins, J. M., & Conduct Problems Prevention Research Group. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association, 96*(456), 1410–1423.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide.* Los Angeles, CA: Author.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* New York, NY: Cambridge University Press.

Phillips, K. A., Shlipak, M. G., Coxson, P., Heidenreich, P. A., Hunink, M. G., Goldman, P. A., . . . Goldman, L. (2000). Health and economic benefits of increased beta-blocker use following myocardial infarction. *Journal of the American Medical Association, 284*(21), 2748–2754.

Prince, M., Stewart, R., Ford, T., & Hotopf, M. (Eds.). (2003). *Practical psychiatric epidemiology.* New York, NY: Oxford University Press.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550–560.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90*(429), 122–129.

Rogosa, D. (1988). Myths and longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York, NY: Springer.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika, 75*(3), 577–581.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics, 41*(1), 103–116.

Rosenberger, W. F., & Lachin, J. M. (2002). *Randomization in clinical trials: Theory and practice.* New York: Wiley.

Rothman, K. J., & Greenland, S. (1998). Causation and causal inference. In K. J. Rothman & S. Greenland (Eds.), *Modern epidemiology* (2nd ed., pp. 7–28). Philadelphia, PA: Lippincott-Raven.

Rothman, K. J., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health, 95*(Suppl. 1), S144–S150.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, 3*(2), 135–146.

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. (1975). *Bayesian inference for causality: The importance of randomization.* In *Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 233–239). Alexandria, VA: American Statistical Association.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test: Comment. *Journal of the American Statistical Association, 75*(371), 591–593.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31*(2), 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design,

modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322–331.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics, 2*(3), 808–840.

Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science, 2*(4), 377–395.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York, NY: Oxford University Press.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1–2), 305–353.

Smith, V. L. (1991). Rational choice: The contrast between economics and psychology. *Journal of Political Economy, 99*(4), 877–897.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33*(2), 230–251.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5*(4), 465–472.

Stewart, R. (2003). Inference 2: Causation. In M. Prince, R. Stewart, T. Ford, & M. Hotopf (Eds.), *Practical psychiatric epidemiology* (pp. 239–253). New York, NY: Oxford University Press.

Stormshak, E. A., Bierman, K. L., Bruschi, C., Dodge, K. A., & Coie, J. D. (1999). The relation between behavior problems and peer preference in different classroom contexts. *Child Development, 70*(1), 169–182.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its applications.* New York, NY: Springer.

Susser, E. S., Schwartz, S., Morabia, A., & Bromet, E. J. (2006). *Psychiatric epidemiology: Searching for the causes of mental disorders.* New York, NY: Oxford University Press.

Tice, J. A., Ross, E., Coxson, P. G., Rosenberg, I., Weinstein, M. C., Hunink, M. G. M., . . . Goldman, L. (2001). Cost-effectiveness of vitamin therapy to lower plasma homocysteine levels for the prevention of coronary heart disease: Effect of grain fortification and beyond. *Journal of the American Medical Association, 286*(8), 936–943.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters, 78*(17), 2957–2962.

Wasserman, L. (2006). *All of nonparametric statistics.* New York, NY: Springer.

Weinstein, M. C., Coxson, P. G., Williams, L. W., Pass, T. M., Stason, W. B., & Goldman, L. (1987). Forecasting coronary heart disease incidence, mortality, and cost: The coronary heart disease policy model. *American Journal of Public Health, 77*(11), 1417–1426.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy.* Chicago, IL: University of Chicago Press.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*: Boca Raton, FL: Chapman & Hall/CRC.

Yatchew, A. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature, 36*(2), 669–721.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician.* New York, NY: Cambridge University Press.

Yoshikawa, H., Magnuson, K. A., Bos, J. M., & Hsueh, J. A. (2003). Effects of earnings-supplement policies on adult economic and middle-childhood outcomes differ for the "hardest to employ." *Child Development, 74*(5), 1500–1521.

Zhang, J. L., Rubin, D. B., & Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In D. L. Millimet, J. A. Smith, & E. J. Vytlacil (Eds.), *Advances in Econometrics: Vol. 21. Modelling and Evaluating Treatment Effects in Econometrics* (pp. 117–145). Oxford, England: JAI Press.

(*Appendix follows*)

Appendix

Analyses of Causal Effects Involving Propensity Scores

Having formed the propensity score and assessed balance and the support problem, one can analyze the data in a range of ways. The seminal articles suggested that the propensity score might be used as a covariate, for stratification, and for matching. One form of matching involves using the propensity score to create probability weights. I discuss each of these in turn.

The Propensity Score as a Covariate

As noted in the original Rosenbaum and Rubin (1983) article, the propensity score serves as a summary of the covariates, and one might use the score as a regressor. Though not uncommon, this practice may strike the reader as odd. After all, one motivation for using propensity scores is to free oneself from regression. One wonders what advantage there is to including the propensity score as a regressor rather than just including the full set of covariates. (It is worth noting that most analysts do not free themselves from the linear model entirely but use it in calculating the propensity score rather than in the analysis of the outcome itself. To some extent, the choice is really a matter of where linearity does the least harm.)

Nonetheless, even as a regressor, the propensity score has some advantages. First, as noted above, the use of the propensity score may lead the analyst to check balance and to identify cases that are unique or nearly so in one group or the other. Second, regression and ordinary least squares are not synonymous. One might turn to semiparametric regression, for example, and in doing so, may make dealing with one covariate easier. Alternatively, one might turn to other forms of estimation. For example, one might apply the differencing procedure outlined in Yatchew (Yatchew, 1998, 2003). This procedure involves sorting the observations by the explanatory variable and subtracting a weighted average of preceding observations. The assumption makes very weak assumptions about the relationship between the covariate and the outcome, thereby eliminating the effect of the former on the latter under a broader array of possible models. A propensity score facilitates this task by summarizing a group of covariates down to a single variable.

Matching on the Propensity Score

Though regression represents a form of matching (Morgan & Harding, 2006), matching is easier to explain to nonstatisticians, such as policymakers. Like regression, matching forms the conditional expectations of interest by treating treatment choice as ignorable conditional on matching. As discussed, the propensity score facilitates matching by reducing the problem to matching on a single variable.

Choosing a matching strategy. Even with choosing to match based on a single variable, however, a range of algorithms exist. I highlight several dimensions of that choice. One key dimension involves the level of similarity (or proximity) one considers "close enough" to constitute a match (e.g., a propensity score difference of no more than one percentage point). In terms of the formulae in the article, this refers to how the set $S(x)$ is defined; now, however, the set is defined based on the propensity score.

The literature does not provide clear guidance on this issue. What is clear is that a looser definition has both advantages and disadvantages. Such a definition has advantages in that fewer observations are considered unmatched. On the other hand, too generous a definition of proximity raises issues of whether the matches are truly matching, leaving unadjusted differences between the treatment and comparison groups. To some extent, one can assess the alternatives empirically. One might compare the distribution of the covariates under alternative definitions of proximity. One likely has too generous a definition of proximity if the covariate differences are statistically significant.

A second key issue involves the broader structure of the matching process—namely, whether matching is "greedy." Suppose one is interested in average effect of treatment for the treated (ATT). The analyst might take each treatment case and then identify the most similar comparison case. This matching is greedy in the sense that a comparison case once used would be removed from the pool of eligible matches. Other algorithms might work through the data multiple times, shuffling matches to improve the overall balance in the sample (i.e., reduce the aggregate measure of between-groups differences). Another possibility is to use a single comparison case as a match for multiple treatment cases, or vice versa. In some instances, one might create the comparison case as a weighted average of a set of matched comparison cases. One might weight the cases in the average according to the proximity of the propensity score (see, e.g., Heckman, Ichimura, & Todd, 1998).

An extensive literature considers the best forms of matching. A good summary can be found in Rosenbaum (2002). One result established there is that an optimal matching has to involve so-called full matching. (*Optimal* here means the smallest possible differences between cases in the matched group.) This result means that one might match a case in one group to multiple cases in the other group but would never match multiple cases in one group to multiple cases in the other group.

Incorporating matching in the analysis. With cases having been matched in some way, a key choice is whether to incorporate the results of the matching in the analysis. Some analysts treat the propensity score diagnostic as a preanalytic stage where one weeds out the mismatched cases and then proceeds to analyze the data via an appropriate method (Ho, Imai, King, & Stuart, 2004). For example, one might run a regression using the outcome as the dependent variable and treatment and the covariates as explanatory variables. The propensity score is used only in the preliminary stage.

(*Appendix continues*)

Other methods incorporate matching more formally in the actual analyses. For example, one can form matched sets based on the propensity score and a matching algorithm and treat the sets as the unit of analysis. (These matching algorithms also may be used as a supplemental means of identifying cases that are mismatched.) One then might apply a simple nonparametric test. Such a test would involve counting the number of times the treatment case in the pair had the better outcome and calculating whether this figure exceeded what one would expect based on chance alone. A range of alternative tests are available that allow one to incorporate the rank ordering of cases on the outcome or matched sets of varying sizes (Wasserman, 2006).

Matching as Stratification

Another possibility for analyzing outcomes based on the propensity score is stratification. One might divide the sample into five quintiles based on the overall distribution of the propensity score (i.e., both the treatment and comparison observations). One then can calculate a measure of treatment effect (e.g., the mean difference in the outcome) for each stratum. One then can weight these estimates by the proportion of the population that lies in each stratum. The choice of weights is determined by the estimand of interest. For example, if one uses the proportion of the treatment cases in each stratum, the resulting estimate is the ATT. Weighting by the overall sample proportions or those for the comparison group produces the ATT and the average effect of treatment on the untreated (ATU), respectively.

The proportion of the population that is in each stratum is actually unknown, but one can estimate that proportion using the full sample of both treatment and comparison cases. That combined sample is assumed to represent the population as a whole. Of course, this procedure becomes problematic if no observations in either the treatment or the comparison group are in a given. Obviously, it is impossible to calculate a mean between-groups difference in a stratum when only one group is present. One can see the problem as well in terms of the weights. If a case in one group has a zero chance of being included in the sample, then the weight for the weighted average is one over zero, or infinity. In that case, one has effectively identified a group of cases that have no match in the other group, and one can handle those as discussed above.

How many strata are enough? Citing an article by Cochran (1968), analysts often use five strata. But the justification for this practice is really nothing other than convention. To know whether five or any other number of strata is adequate, one would need to assess the balance of covariates as described above. In abstraction, the more strata the better, but there is some incentive to minimize that number. The reason is that the more strata, the more likely a stratum includes only treatment or comparison cases. One cannot estimate a treatment effect for that stratum, so these cases would need to be excluded (as were the other observations with no match). Rather than assume five strata are enough, a better strategy would be to include a large number (e.g., 20) and achieve balance. Then one could collapse strata until balance is lost.

How does stratification relate to the analysis of matched sets? In essence, the method is an analysis of matched sets where the number of sets is predetermined and the definition of closeness is potentially rather loose. My opinion is that stratification should be used only initially to understand the data, and then one should turn to one of the more general methods. After all, the many-to-many nature of strata as matches disqualifies stratification as a form of full matching.

Using the Propensity Score to Form Probability Weights

One can think of stratification as weighting the observed data to represent an unobserved population. To form the ATT, for example, one reweights the comparison cases across the strata using the distribution observed for treated cases. The counterfactual now represents what the comparison group would look like were the distribution of covariates in that group equal to that observed for the treatment group.

Generalizing this idea to numerous strata, one can form probability weights. The key insight is that the observed data provide an unrepresentative sample from the hypothetical population of interest. The group actually receiving treatment, for example, represents a sample of a population in which everyone received treatment. Of course, because treatment is not randomly assigned, the sample is not representative of the population. Indeed, this is just the situation in which sampling statisticians often find themselves. In that case, the statistician estimates the probability an individual will be sampled based on a list of covariates. The statistician then forms a probability weight as one over that probability. One then can use the probability weights to make the observed data representative of the full population in terms of the variables used to form the weights. The weight is one over the probability of participation. The net result is that individuals with low probabilities receive large weights: They are underrepresented in the sample, and one needs to inflate their importance in the calculations. Similarly, individuals whose probability of participation is near one are adequately represented, and their weight is roughly one.

Now consider the case of causal inference. To make the treated group representative of the entire population, one can calculate the probability of participation—in this case, the probability of receiving the treatment or the propensity score. Similarly, one can make the comparison group representative of a full population of untreated individuals using the probability of participation; in this case, the weight is one over one minus the propensity score. Having generated those two populations, one can calculate the average treatment effect.

Similarly, one can calculate ATT and ATU. For the former, one need not weight the treated cases at all: Those individuals can represent the experience of the treated individuals were they to receive treatment. The calculation for the untreated individuals is more tricky. The weight for these individuals is the propensity score divided by one minus the propensity score. One can think of the weight as involving two steps. First, one weights by one over one minus the propensity score; as above, this weight has the effect

(*Appendix continues*)

of setting the distribution of the covariates in the untreated group to equal that for all individuals. One then multiplies the weight by the propensity score to reproduce the distribution of the covariates for the treated group. In other words, the weight standardizes the untreated group using the distribution of the covariates for the treated group.

Analogously, the weight for the untreated cases in calculating the ATU is one. The weight for the treated cases is one minus the propensity score divided by the propensity score. The weighting standardizes the distribution of the covariates in the untreated group to the distribution in the treated group.

The problem with this procedure is like that experienced by sampling statisticians. In some instances the weights can be very large. Highly variable weights inflate standard errors. For that reason, sampling statisticians often trim their weights, that is, reduce their weights in an ad hoc fashion. One could do the same thing in an analysis of causal effects. As with the number of strata, however, it would be important to check the balance of the covariates as one proceeds. A reasonable strategy would be to reduce the largest weights until the point when balance is lost.

Mixing Methods

As noted, the analyst must check to ensure that the covariates are balanced, but the final results may be sensitive to other issues. One issue is model misspecification. For example, the logit model may have limitations for estimating the propensity score: If the tails of the distribution are thicker than the logistic regression allows, the predicted probabilities generated will have poor statistical properties.

Presumably, a poor-fitting propensity score model would result in an imbalance of the covariates and would be detected in diagnostic checking. However, to guard against problems with the model and the resulting propensity scores, one might combine the strategies presented above. For example, one might use the probability weights generated from the propensity score but also include the propensity score or even the covariates as a group as a regressor (Robins & Rotnitzky, 1995). If the assumptions underlying either the regression adjustment or the propensity score weighting are correct, the resulting estimate of the treatment effect has good statistical properties.

Received March 2, 2009
Revision received April 21, 2010
Accepted April 23, 2010 ■