CAUSAL EFFECTS OF PERCEIVED IMMUTABLE CHARACTERISTICS

D. James Greiner and Donald B. Rubin*

Abstract—Despite their ubiquity, observational studies to infer the causal effect of a so-called immutable characteristic, such as race or sex, have struggled for coherence, given the unavailability of a manipulation analogous to a "treatment" in a randomized experiment and the danger of posttreatment bias. We demonstrate that a shift in focus from actual traits to perceptions of them can address both of these issues while facilitating articulation of other critical concepts, particularly the timing of treatment assignment. We illustrate concepts by discussing the designs of various studies of the role of race in trial court death penalty decisions.

I. Introduction

WE discuss the prerequisites for the design of an observational study to infer the causal effect of a so-called immutable characteristic, such as race or sex. Despite their ubiquity, such studies have struggled for theoretical coherence because of, among other things, the impossibility of manipulating such traits in a way analogous to administering a treatment in a randomized experiment and the danger of posttreatment bias stemming from the fact that almost all variables on which a researcher would like to condition are determined after an individual's conception. Because of these issues, prominent scholars (Holland, 1986a; Winship & Sobel, 1999; Freedman, 2004; Berk, 2004) contend that it is inappropriate to conceptualize a person's actual race, sex, or national origin as a treatment in an observational study and that attempts to infer the causal effects of such traits are incoherent.

Here, we articulate how to draw causal inferences with respect to immutable characteristics. We argue that a shift in focus from actual traits to perceptions of them allows a researcher to address both of these problems while simultaneously facilitating sharp articulation of necessary assumptions. Our discussion has practical consequences. For example, we demonstrate that one cannot coherently study the effect of victim or defendant race on jury assignment of capital punishment without including acquittals in the analysis data set, despite numerous studies that fail to do so. Similarly, we suggest that even certain kinds of randomization will not allow causal inference of the effect of a judge's sex on case outcomes despite studies attempting to do so.

We ground our discussion within the potential outcomes framework of causation ((Neyman, 1990) in the context of randomization-based inference in randomized experiments, and Rubin (1974, 1978) more generally, including for Bayesian inference in randomized studies), also called the Rubin causal model (Holland, 1986a). We nevertheless believe the concepts discussed, particularly the presumptive timing of treatment assignment we identify, are equally applicable and useful to other causal paradigms (Pearl, 2000; Heckman, 2005). Although we draw our examples primarily from antidiscrimination law, the issues we discuss are applicable to studies in economics, sociology, political science, and other fields.

A preview and slight oversimplification of our thinking, using race as an example, is as follows. When assessing whether intentional discrimination (what the law labels "disparate treatment") has occurred, researchers are often interested in whether the decisions of what we call a "decider"—a particular person or an institutional actor—are made without regard to race. Under this formulation, perceived race is not an immutable characteristic because one can hypothesize interventions that might change the decider's perceptions. Such interventions are of only indirect concern, however, because the researcher does not contemplate manipulating perceptions of race and therefore does not contemplate implementing the treatment whose effects are under study. A key benefit of this formulation is that it identifies the presumptive timing of treatment assignment as the moment the decider first perceives a unit's race. This moment, which occurs long after the unit's conception, allows (indeed, requires) the researcher to consider pretreatment variables as covariates. And the entire structure allows sharp articulation of critical assumptions.

For the remainder of this paper, we assume familiarity with the following concepts: units, treatments, timing of treatment assignment, outcome variables, covariates (observed and unobserved), intermediate outcomes, assignment mechanism, ignorability, posttreatment-adjustment bias, the stable unit treatment value assumption (SUTVA; Rubin, 1980a), and the fundamental problem of causal inference. For readers unfamiliar with any of these terms, summaries are available in Rubin (1978, 2006b), Holland (1986b), Imbens and Rubin (2006), Gelman and Hill (2007).

II. Previous Arguments

For analysts from a variety of fields, the intensely practical goal of causal inference is to discover what would happen if we changed the world in some way (Holland, 2003). This understanding has several consequences discussed in detail elsewhere: the maxim, "No causation without manipulation" (Rubin, 1975); the randomized experiment's status as the gold standard for causal inference; and the imperative to analyze observational data by reconstructing a hypothetical

Received for publication March 12, 2009. Revision accepted for publication March 5, 2010.

* Greiner: Harvard Law School; Rubin: Harvard University.

We extend great thanks to Richard Berk, Gabriella Blum, Rachel Brewster, Stephen Fienberg, Franklin Fisher, Andrew Gelman, Adam Glynn, Don Greiner, Ellen Greiner, Sam Gross, Daniel Ho, Louis Kaplow, Michael Kellerman, Gary King, Jennifer Lewis, Adriaan Lanni, Xiao-Li Meng, Kevin Quinn, Jed Shugerman, Ben Sachs, Matthew Stephenson, Jeff Strnad, Elizabeth Stuart, Jeannie Suk, Chris Winship, and Alan Zaslavsky for helpful comments on earlier versions of this paper. No one mentioned necessarily endorses the statements herein. We also thank two anonymous reviewers for their constructive comments.

The Review of Economics and Statistics, August 2011, 93(3): 775–785

randomized experiment by (among other things) separating covariates from intermediate outcomes and balancing covariates between treatment and control groups (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1983; Rubin, 2007a).

The emphasis on manipulation has led some scholars (Holland, 1986a; Winship & Sobel, 1999; Freedman, 2004; Berk, 2004) to contend that it is inappropriate to conceptualize a person's actual race, sex, or national origin as a treatment in an observational study. Holland (2003) in particular distinguishes "properties" or "attributes," such as race and sex, from "causes," such as a pill. The objection to studying causal effects of attributes has two aspects. First, attributes are not subject to change by intervention. Second, some properties (including immutable characteristics) are determined at a person's conception, and thus almost all measurable variables specific to a unit are posttreatment: "For example, because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black" (Holland, 2003). Note that such arguments may implicitly adopt a biological definition of immutable characteristics.

Meanwhile, other scholars have explored the idea that perceptions of immutable characteristics, not the "actual" traits (to the extent that the latter are well defined), are what matter and that perceptions are manipulable. Social psychologists have experimented with manipulations involving race in the lab (for a review and an analysis of cognitive mechanisms, see Cosmides, Tooby, & Kurzban, 2003). Researchers have also discussed (Berk, 2004) and implemented manipulations of variables closely associated to traits. For example, analysts have manipulated names associated with political speeches (Sapiro, 1981–1982) or résumés (Bertrand & Mullainathan, 2004), although these variables may also be closely associated with other and less legally problematic variables such as socioeconomic status or education. Other scholars have explored the effect of interventions that remove a decider's ability to perceive a trait (for example, placing a screen in front of auditioning musicians, Goldin & Rouse, 2000). Thus, the idea that perceptions matter and can be manipulated is not new. Nevertheless, nothing in this literature articulates the definitions and assumptions needed for causal inference regarding the effects of perceived immutable characteristics via an observational study.

To our knowledge, Fienberg and Haviland (2003) provide the first explicit discussions of perceptions as opposed to actual traits in a causal inference framework. They note that perceptions may be manipulable and discuss both the need to observe variables sufficient to justify an ignorability assumption, as well as the advantages of matching methods in this area. Some excellent work that we have seen on the subject we address here is Kaufman (2008), who discusses the importance of a decision maker (whom we call the "decider") and distinguishes inferences about the effect of race on such a person or entity's decisions, which may be rigorously conceptualized, with inferences about other aspects of race, which may not be. We developed this idea independently from Kaufman (see Greiner, 2007), and it is a cornerstone of proper thinking in this area. It is not the only cornerstone, however. Work remains to be done, among other things, to answer Holland's objections articulated above; to address the timing of treatment assignment; separate covariates from intermediate outcomes; and articulate the meaning of the stability assumption.

III. Perceptions

A shift in emphasis to perceptions of immutable characteristics allows some well-defined causal questions to be posed and, within the limits of observational studies, inferences to be drawn. At a minimum, it is possible and useful to identify a set of assumptions that allows causal analysis. If some find the assumptions we state too strong to accept as plausible, we sympathize. We do not intend to minimize the difficulties involved.

To answer the first primary objection to considering immutable characteristics as "treatments" as opposed to "attributes," the unavailability of a manipulation, one must ask the following question: Why does causal inference by the potential outcomes framework typically require the researcher to specify an intervention, real or hypothetical? The answer is that a researcher typically has some desired state of the world he wishes to realize, and that whether a contemplated intervention will further progress toward that state of the world can depend critically on the intervention's precise nature. For example, analysts attempt to study the effect on test scores of sending children to Catholic schools (Coleman, Hoffer, & Kilgore, 1982). Here, the "no causation without manipulation" maxim reminds us that such studies can be abused. Politicians might rely on them to push school vouchers, but vouchers may induce only certain types of children to attend private schools or induce children to flock to private schools in overwhelming numbers (Morgan, 2001). Thus, the danger is that analysts may study the effect of a variable that might bear an uncertain relationship to a contemplated policy intervention, whereas proper causal inference counsels study of the effect of the proposed intervention itself (Mealli & Rubin, 2003).

Immutable characteristics as causes are different. When studying the causal effect of traits, particularly in law, we ordinarily do not contemplate an intervention taking the form of manipulating these attributes or even perceptions of them. Rather, the goal is to decide whether to compensate victims (as in employment discrimination) or whether to alter a governmental system in light of any discrimination found (such as by suspending the death penalty). In short, we draw inferences as to causal effects of perceptions of traits to decide whether to intervene in some remedial way, not to study what would happen if we did intervene to alter these perceptions. Thus, an inability to manipulate actual immutable characteristics may not be fatal.

This is not to say that identifying possible manipulations has little value. To the contrary, such an exercise can illuminate several important aspects of an observational study.

We identify two such aspects. First, when a researcher identifies what would be manipulated, the mechanisms by which a person (below, we define this person, the decider) perceives another's race (or sex, or something else) are thereby identified. If, as we believe, race is a social construct, then specifying the mechanisms by which the decider perceives another's race helps to define that construct for the purposes of the study. Second, and relatedly, specifying possible manipulations can assist in the process of identifying covariates needed to make an ignorability assumption plausible.

We discuss ignorability in greater detail below, but to illustrate this point briefly, consider Bertrand & Mullainathan (2004). This study sought to manipulate perceived race by manipulating names on résumés (for example, "Lakisha" for black, "Emily" for white). As Bertrand and Mullainathan recognized, however, the hiring authorities in the firms they studied might use names as a shorthand for other variables, such as education, skills, or socioeconomic class. For this reason, they sought to provide information on these variables in the résumés they sent. Thus, a focus on the variable to be manipulated helped these analysts identify possible confounders to the effect they wanted to measure. Identifying possible manipulations can serve the same role in an observational study.

Note that the fact that the analyst might identify several possible manipulations affecting perceived traits (for example, names, clothes, hair length, body shape) does not pose definitional problems. Treatment variables need not be 0 or 1. And when proceeding Bayesianly, so that missing counterfactual values are conceptualized as random variables with a distribution, an analyst could define a probability distribution for the various manipulations. The distribution for each unit's missing counterfactual value would then be represented by a stochastic mixture of the distributions induced by the various combinations of manipulations. Such a conceptualization might pose inferential challenges and assumptions of some strength may be required to proceed, but the basic framework remains coherent.

The second primary objection to considering immutable characteristics as treatments as opposed to attributes is that at-conception assignment of treatment renders almost all relevant variables posttreatment. This is a serious problem, but it does not prevent all progress. For example, many antidiscrimination mechanisms, particularly those implemented as a result of litigation, turn on whether a particular person or institutional actor, a decider, has behaved in a trait-neutral manner. (We distinguish the decider, who controls the outcome of interest, from the decision maker in Rubin, 2008, who may control treatment assignment.)

In the employment context, for example, a race discrimination lawsuit focuses on whether a firm has administered some benefit, such as hiring without regard to race. Hypothetically it might be that African Americans applying for jobs at a firm suffer lower education achievement (as compared to Caucasians) because of past governmental school segregation and that the employer, by hiring based on education achievement, perpetuates the effects of this discrimination. Under the law, however, as long as the employer makes decisions on the basis of educational achievement alone (meaning without regard to race), no liability attaches. In colloquial terms, the employer is responsible only for avoiding its own discrimination, if any. (We remind readers that here we discuss intentional discrimination, not what the law calls disparate impact, which focuses on disparities on different groups' outcomes allegedly caused unintentionally.) Thus, in the employment example, we are not simply allowed to condition on all variables whose values are determined prior to the moment of first interaction between a set of job applicants and an employer; we are compelled to do that conditioning. Many relevant variables should accordingly be thought of as pretreatment.

Much here depends on a willingness to exonerate the decider from responsibility for prior events. In some situations, we do require the decider to respond to circumstances arguably not of its own creation. For example, a statutorily imposed duty to "affirmatively further fair housing" (42 U.S.C. §§ 1437c-1(d)(16), 5306(d)(7)(B)) may require a local housing authority to decrease the racial identifiability of neighborhoods, a circumstance that may be due in part to nongovernmental decisions in the private housing market. This situation differs from the one we study.

A. Primitives

We identify the primitives of causal inference: the units, the treatments, the timing of treatment assignment, and the outcome. Again, we use antidiscrimination (for example, in jobs or capital trials) to illustrate concepts.

A unit is typically a person in some defined role, such as an applicant for a job or a capital defendant. The treatment is the unit's immutable characteristic as perceived by the decider—an employer or the jury in a capital case (see Pierce & Radelet, 2002, for analogous reasoning). The timing of treatment assignment is presumptively the moment the decider first perceives the unit's immutable characteristic. Conceptualizing treatment as occurring at the moment of first perception captures the fact that variables whose values are determined after that moment may be affected by the perception. For example, in an evaluation of an applicant's job interview, an employer may rate a unit perceived to be male higher than an otherwise functionally identical unit perceived to be female, so it makes sense to conceptualize the evaluation as an intermediate outcome. This critical issue of the timing of treatment assignment has been unaddressed in statistical work regarding immutable characteristics (National Research Council, 2004).

Two aspects of the primitives deserve further explanation. First, to make it possible to imagine a unit's counterfactual outcomes and avoid the problem of a treatment administered at birth, the decider must be a relatively discrete person or institutional actor. For example, juries in capital cases play a

discrete role and interact with homicide victims over a defined period of time, from jury selection to punishment verdict. It may be possible, then, to imagine a capital jury's perception of a victim's race being different from what it actually was, that is, to visualize a counterfactual. In contrast, if the decider is the set of all employers in the United States, as might be in a nationwide study of the effect of perceived race on wages, we have difficulty visualizing stable potential outcomes for study units, a point to which we return in our discussion below on the limits of our proposal.

Second, the emphasis on perceptions addresses problems that previous researchers have not considered. One such problem addressed is the question of defining "true" race. Many view race as a social construct that evolves over time, as opposed to a biological concept (American Anthropological Association, 1998; Lopez, 1996; Holland, 2003). For example, one court dispute involved the following question: "Is a high caste Hindu of full Indian blood, born at Amrit Sar, Punjab, India, a white person within the meaning of" U.S. naturalization law (*United States* v. *Thind*, 261 U.S. 204, 206, 1923)? This question's combination of socioeconomic status, religion, ancestry, geography, and ethnicity suggests that attempting to define "true" race in terms of biological characteristics could be futile. A focus on perceptions demonstrates that such an attempt may also be unnecessary. Instead, for causal inference to proceed, we must believe that U.S. society has constructed classifications called, say, American Indian or Alaska Native, Asian American, Black or African American, Native Hawaiian or Other Pacific Islander, and White (U.S. Office of Management and Budget, 1997), and that these classifications exist in the mind of the decider. If so, then biological definitions are irrelevant.

B. SUTVA

Previous work (Rubin, 1980b) has emphasized the criticality of SUTVA in defining causal effects in terms of an $N \times 2$ table of the units' potential outcomes, where N is the number of units. Both prongs of SUTVA—that for each unit there is only one form of the treatment that the unit did not receive (for example, only one kind of active pill for a unit that receives placebo) and that the treatment one unit receives does not affect a different unit's potential outcomes—require careful evaluation.

The assumption that there is only one form of each counterfactual treatment for each unit will ordinarily involve at least three different aspects: one touching on unit characteristics, the second involving the extent of the interpersonal interaction between the units and the decider, and the third consisting of a concept we label "invariance." With respect to unit characteristics, the assumption means, for example, that for a unit actually perceived as male, a potential employer would not base hiring decisions on degree of "womanliness" were that unit perceived to be female. Rather, for this unit who is actually perceived male, we must imagine there could have been in the employer's mind a single essential, counterfactual state of "woman." With respect to the extent of interpersonal interaction, in most settings, a researcher will be forced to assume that the nature and extent of the relationship between the units and the decider does not change the potential outcomes. For example, in a death penalty setting in which the defendant's race is at issue, a researcher will ordinarily have to assume that it makes no difference whether a defendant exercises a Fifth Amendment right not to testify, testifies for a short period of time, or occupies the witness stand for a week.

To understand what we call invariance, one might ask how jurors discover a homicide victim's race. A defense attorney may exercise his client's constitutional right to inform the jury pool of the victim's race and question potential jurors as to their prejudices (*Turner* v. *Murray*, 476 U.S. 28, 37, 1986). Alternatively, jurors may see pictures of a corpse, or they may draw an inference about the victim's race based on their perceptions of the races of the victim's relatives who are called to testify. The critical assumption here is that how the perception is created does not matter, that is, the counterfactual potential outcome is "stable," invariant to the nature of the evidence on which the decider's perception is based. Something analogous to this assumption is implicit in any causal study. For example, whether a potential vaccine shot is administered in a unit's right or left arm is ordinarily not recorded. On the other hand, some differences in application (for example, intermuscular versus subcutaneous administration of a drug) may make a serious difference.

With respect to the second prong of SUTVA, noninterference, much depends here on the choice of decider and the question to be studied. For example, in the death penalty context, with the jury as the decider, the legal steps taken to ensure jury neutrality and independence may render a noninterference assumption plausible. In contrast, with the prosecutor as the decider, the researcher must think carefully about whether, for example, resource constraints allow the prosecutor's charging decisions to be independent from case to case.

As is true with many assumptions, SUTVA could be relaxed if the researcher has access to information beyond that typically available. For example, if an accurate measurement of maleness were possible and recorded, where accuracy here focuses on how well the measurement mimics the governmental or socioeconomic actor's perceptions, then the treatment might be conceptualized as taking on multiple values. In the ordinary study, however, the researcher will have access only to "M" or "F." Similarly, and as discussed in section II, the invariance assumption may be relaxed if the analyst proceeds in a Bayesian fashion and hypothesizes a probability distribution over several possible hypothetical manipulations. Proceeding in this manner in an observational study would require substantial quantities of data regarding which manipulations units actually received, however, and may not be feasible in many observational contexts.

C. *Further Assumptions for Implementation*

The primitives, together with SUTVA, allow a researcher to construct a well-defined framework for causal inference. In practice, two additional assumptions, ignorability and accurate perceptions, will typically be necessary when applying the framework to a data set.

Ignorability. In complicated transactions in which a unit interacts with multiple parts of an overall socioeconomic or governmental system, a researcher may have more than one choice of decider to study. In such situations, the researcher may have to balance the need to make an ignorability assumption plausible against a desire to detect the effect of perceived immutable characteristics in all aspects of the system. By focusing on a decider who perceives the unit's immutable characteristic "late" in the interaction, the researcher implicitly chooses a later timing of treatment assignment, rendering more measured variables pretreatment and thus properly characterized as covariates. That in turn can make an ignorability assumption more plausible. But by treating such variables as covariates (and thus conditioning on them in the analysis), the researcher forgoes the detection of any prior discrimination that may have affected the values of these covariates.

For example, in homicide investigations, at least six distinct actors play a role in the administration of a case from discovery of a corpse's race to a possible death sentence: the police, the prosecutors, the witnesses, the defense team, the judge, and the jury. Jurors perceive the victim's race late in the administration of the case, typically on or after the first day of jury selection, allowing a researcher to consider as covariates any variables whose values are determined before jury selection begins. That makes an ignorability assumption more plausible but also renders beyond the scope of the study any discrimination prior to voir dire by the police, the prosecutor, the defense counsel, the witnesses, or the judge. In contrast, a researcher studying prosecutors may have fewer pretreatment variables, but if an ignorability assumption is nevertheless plausible, discrimination (if any) in charging decisions may be estimable.

Accurate perceptions. In a typical biomedical or social science study, there are few conceptual issues associated with recording the treatment to which a unit is assigned and whether there is full compliance (for example, a unit takes either an active ingredient or a placebo). In the context of perceived immutable characteristics, the treatment is in the mind (perhaps the collective minds) of the decider and is thus technically unobserved. What is typically observed or recorded is someone else's perception of each unit's immutable characteristic, often the unit's self-report. Whatever the source of the recorded value, its value must be "accurate," for example, agree with the decider's perception. This assumption may often be plausible, but the recent increase in the numbers of persons with mixed race or ethnic self-identification, however, may make this assumption questionable.

IV. Where the Framework May Work

We believe that the set of circumstances in which inferences about the causal effect of perceived immutable characteristics can reasonably be attempted is nonempty. For example, a researcher might be interested in the effect of the victim's race on the jury's decision to impose death or life imprisonment in cases in which no member of the victim's family takes the witness stand (note that we focus on the victim here, in contrast to our earlier focus on the defendant; as we will explain, capital punishment studies of race typically concern the races of the victim and the defendant). Under these circumstances, the jury's perception of the victim's race will ordinarily be based on evidence that one can imagine a researcher manipulating, such as defense counsel's statements to the jury pool during jury selection, the juror's observation of the victim's physical characteristics as depicted in photographs, or the victim's name. Defense counsel could be asked to change a statement, photographs could be altered, or the victim's name (as reported to the jury) could be changed. The fact that such hypothetical manipulations are illegal does not distinguish this kind of observational study from other observational studies in which a contemplated intervention is unethical or prohibitively expensive. In any event, as explained in section II, precisely specifying manipulations is less essential in the immutable characteristics context because no one contemplates actually implementing them. Finally, legal safeguards such as sequestration may ensure that juries do not interfere with one another, and variables with values set prior to jury selection are covariates, thereby making SUTVA and ignorability more plausible.

Contrast this relatively straightforward situation with that confronting a researcher who attempts to assess the effect of the victim's race on the entire capital criminal process. Here, the decider is the entire criminal justice system, including the police, who ordinarily perceive victim race on discovery of a corpse. Many variables a researcher might record, from the severity of the circumstances surrounding the crime to some of the mitigating characteristics of the defendant's life (as discovered by investigations), are revealed posttreatment. Although some underlying facts are pretreatment, the facts as they are recorded in available sources (such as a police report) are not. If police investigate majority-race homicides with greater vigor than others, then the treatment may affect what exists in the case file. This issue requires careful thought.

Thus, we have identified some circumstances under which causal inference as to the effect of immutable characteristics can proceed under plausible assumptions and some in which, under the current state of our knowledge, such inference might require implausible assumptions.

V. Familiar Patterns

Clarifying these principles has several practical benefits. In particular, problems arising in research regarding immutable characteristics are structurally identical to issues confronted,

and solved, elsewhere. Numerous examples exist; we provide one illustration from the death penalty context. In response to the Supreme Court's decision in *Furman* v. *Georgia* (408 U.S. 238, 1972), states adopted multistage trials for capital prosecutions. For simplicity, we consider a two-phase process used in some states, including Georgia. In the first, or guilt, phase, the jury decides whether the defendant committed a death-eligible crime. If it convicts on a death-eligible offense, the same jury then decides, in a penalty phase that ordinarily includes additional evidence, whether the defendant should be executed or sentenced to life in prison. The penalty phase includes a variety of procedural safeguards designed to reduce consideration of irrelevant factors, including race.

Under such a two-stage, single-jury system, courts and researchers often ask whether race (of the defendant) plays a role in the jury's sentencing decision, an inquiry that directly engages the procedural safeguards mentioned above. For illustrative purposes, we assume for the moment that all defendants are either white or black. Jurors typically perceive a defendant's race during voir dire, and certainly by the first day of the guilt trial. Thus, the treatment is administered prior to the jury's guilt verdict. But for the death penalty to be an option at the sentencing phase, the jury must first have convicted the defendant of a death-eligible offense (usually first-degree murder). Thus, an effect of defendant race on sentencing is defined only for capital cases in which the jury would have convicted the defendant of a death-eligible offense under both treatments: perceived as white and perceived as black. A researcher who fails to isolate this set of units for analysis could make one of many mistakes, among them the attribution of an effect in fact occurring at the guilt phase to an effect on sentencing. Structurally this problem is identical to one labeled "censoring due to death" or "truncation due to death" in the biomedical context (Zhang & Rubin, 2003; Rubin, 2006a). Note that here, alas, "death" in the biomedical context is an acquittal (or a conviction of a lesser offense) in the legal one. Greiner (2008) discusses further details. We return to this concept in our discussion of the death penalty literature.

VI. Capital Punishment Studies

To illustrate concepts, we had initially thought to reanalyze data from one of the numerous recent (post-1990) studies on race and capital punishment. After reviewing dozens of papers in the area and paying particular attention to study design (Rubin, 2008), we could find no preexisting capital punishment data set that would allow causal inference by the potential outcomes framework to proceed. As explained below, we take heart in the fact that we do believe it possible to gather a data set that would allow a successful study of, at a minimum, jury decision making. Our reading of the literature suggests, however, that no such data-gathering effort has been accomplished so far, and we thus review capital punishment study designs to illustrate our ideas. We also reanalyze some data from the most famous empirical study in this area (indeed, perhaps the most famous empirical study in the law): the Baldus Charging and Sentencing Study of Georgia during the 1970s (Baldus, Woodworth, & Pulaski, 1990). But because we are unable to articulate a well-defined causal question answerable with the Baldus study data, we keep our discussion brief.

We discuss each of two categories into which fall the majority of studies of trial-level adjudication of death sentences: analyses that follow an approach we believe to be due to Gross and Mauro (1984) and those that structurally resemble the Baldus study. We include in the latter a discussion of the results of our reanalysis of a portion of the Baldus study data.

A. SHR Studies

To our knowledge, Gross and Mauro (1984) pioneered the practice of linking covariates from Supplementary Homicide Reports (SHRs) with outcome data from other sources to examine death penalty administration (for subsequent similar analyses, see Radelet & Pierce, 1991; Brock, Cohen, & Sorensen, 2000; Williams & Holcomb, 2001; Lenza, Keys, & Guess, 2005). Local law enforcement agencies file, with the FBI, SHRs reporting the sex, age, and race of the both victim and suspected killer (if information on the latter is available); the date and place of the homicide; weapon used; contemporaneous felonies; and a code for the victim-suspect relationship. The Gross and Mauro (1984) template is to match SHR cases to files from another data source (Gross and Mauro used an NAACP database) containing information on whether a defendant was sentenced to death. Because SHRs do not include unique identifiers for the victim or a suspect (the latter may not exist at the time the report is filed), the linking process can be difficult and time-consuming. Once linking is complete, cross-tabulation or model-based analysis can proceed using the variables contained in the SHR, often with the four treatment groups described by the four victim/defendant black/white combinations.

The decider implicitly chosen in these studies is difficult to ascertain. To limit analysis to SHR data with reasonably complete information, Gross and Mauro (1984) use only those SHR cases for which an age of the suspect is available, suggesting that for the remainder, there being no suspect, there was "insufficient information . . . to form the basis for official action" (p. 98). In other words, these cases never became part of the criminal justice system. But entry into the criminal justice system is itself a decision made by deciders, particularly the police and, to the extent involved in the investigation, the prosecutor. Whether these deciders find and declare a suspect could depend on the nature of their investigations, which might depend on the perceived racial characteristics of the victim and possible suspects.

Perhaps the SHR studies implicitly choose as the decider the criminal justice homicide system from the prosecutor's office to penalty phase, thus internalizing an assumption that any posttreatment bias in prosecutorial behavior did not affect the covariate data in the SHRs. If one concedes that such a focus leads to a well-defined study, a second difficulty

with SHR studies is whether the covariates in the SHRs are rich enough to justify an ignorability assumption. To their credit, Gross and Mauro (1984) discuss this issue explicitly; we remain skeptical. The set of factors in the death penalty statutes of the states under study includes more variables than are coded in SHRs. For instance, there appear to be no SHR variables addressing mitigating circumstances. Nor did we encounter a study that attempted a sensitivity analysis (Cornfield et al., 1959) to check for the possible effect of unmeasured covariates.

B. Analyses Resembling the Baldus Study

Perhaps the mostly highly praised empirical study in the law (Committee on Racial and Gender Bias in the Justice System, 2001, collects favorable reviews) is the Baldus Charging and Sentencing study. This study formed the backbone of a claim, eventually rejected 5 to 4 in the Supreme Court (*McClesky* v. *Kemp*, 481 U.S. 279, 1987), that despite procedural reforms, race continued to play an unconstitutional role in Georgia's administration of the death penalty during the late 1970s (Baldus, Woodworth, & Pulaski, 1990). The Baldus study's data set consists of a stratified, multistage sample of cases in which Georgia defendants had been convicted of murder or voluntary manslaughter. Thus, the sample frame did not include cases in which a defendant was indicted for murder but the prosecution was dropped, the defendant was acquitted, or the defendant was convicted of (or pled to) a lesser offense other than voluntary manslaughter (such as reckless homicide or involuntary manslaughter, or attempted murder). Files for cases in the sample were reviewed and a rich set of variables recorded, so for the moment, we assume that an ignorability assumption might be made plausible for a study of some decider.

According to the Baldus study authors, "The primary objective of the [study's] discrimination analyses presented to the court was to estimate racial disparities in death-sentencing rates among defendants indicted for murder. Such disparities would reflect the combined effect of all decisions made from the point of indictment through the jury's decision." (Baldus et al., 1990, pp. 313–314). To assess this claim in light of the causal principles discussed in sections I to IV, we begin by identifying primitives, assuming for the moment that we are interested in race-of-the-victim effects. The unit is a homicide victim; the treatment is the victim's race as perceived by the deciders, meaning several actors within the criminal justice system, including the prosecutor, judge, defense team, and any jury; and the outcome is whether the defendant receives the death penalty (we ignore for the moment the issue of multidefendant and multivictim cases). What is the timing of treatment assignment? If randomization of perceived victim race were possible, when would it occur?

Suppose interest focuses on assessing prosecutorial behavior. Then, as discussed above, many variables a researcher records from a homicide case file are determined after the decider's first perception of the victim's or the defendant's race and are thus potentially influenced by the treatment, requiring the researcher to pause before classifying such variables as covariates. Particularly suspect are variables coding subjective judgments. Paternoster and Brame (2003), for example, study prosecutor behavior with models that use as a covariate whether a crime was "particularly gruesome" or "unusually repulsive/horrific." As Radelet and Pierce (1985) note, so-called objective facts might be subject to manipulation based on the race of the victim or the accused. We do not suggest that researchers should never condition on any variable the value of which is determined after the moment of first perception. We do suggest, however, that subjective evaluations are particularly suspect and that all posttreatment variables require careful evaluation and thought before they are used as covariates.

In our view, the Baldus study did not recognize this point. By way of illustration, one of the predictors conditioned on in the Baldus study's 39-variable core logistic regression, discussed further below, was whether a coperpetrator to the homicide received a lesser sentence. But whether a coperpetrator receives a lesser sentence may depend on whether this individual's testimony contributed to an outcome that the prosecution considers favorable, and thus this variable is not just posttreatment but postoutcome, making its characterization as a covariate questionable.

Identifying the primitives also highlights difficulties posed by the fact that the Baldus data set was collected according to (and limited by) the offense for which the defendant was convicted (see Berk, Asuza, & Hickman, 2005). If the goal is to draw causal inferences about the criminal justice process from the beginning of prosecutorial involvement to the penalty phase, then the sampling frame should have included all cases potentially chargeable as homicides in which a prosecutor became involved, that is, all units subject to randomization in a hypothetical experiment. The practice of selecting cases for study on the basis of some disposition, often a disposition by the decider under study, is common in this area (in addition to papers cited above and below, see Radelet and Pierce, 1985; Klein & Rolph, 1991; Baldus et al., 1998; Weisburd & Naus, 2001; Baime, 2001). The data set in Paternoster and Brame (2003) (reanalyzed by Berk et al., 2005) comes the closest we could find to avoiding this problem, but again, the data were limited to cases that a research panel found to be "death eligible."

We are not the first to make the particular point regarding selection of data by final adjudication (see Pierce & Radelet, 2002, and Radelet & Pierce, 1991), but our approach demonstrates the inadequacy of the retorts most often made to this point. Some defend selection by final adjudication as a way to limit the study to cases that are "death eligible" (Unah & Boger, 2001; Baldus et al., 2002; Weisburd & Naus, 2001), but under the framework proposed here, death eligibility is a choice made by a decider and thus is an outcome variable, not a covariate to use for case selection. Others (Keil & Vito, 1995; Pierce & Radelet, 2002) may be attempting to avoid the problem of selection on final adjudication by limiting

the question of interest to the decision of whether to impose death conditional on a conviction for a death-eligible offense. The apparent thought is that to proceed in this manner, the researcher needs only cases in which the defendant has been convicted of a death-eligible offense, all of which may be (and are, in the Baldus study) included in the data set. This attempt founders when identification of the decider and the timing of treatment assignment uncover the "censoring due to death" issue discussed in section IV. In short, absent strong assumptions, one cannot coherently study the causal effect of race on the sentencing phase of capital trials without including acquittals (and convictions for lesser included offenses) in the data set and sampling frame.

A third retort is that tracking all homicides that reach, say, the prosecutor's office is not feasible (Pierce & Radelet, 2002). Assuming this to be the case, in our view, a less ambitious causal project might be wise. For example, with jury as the decider, the first step would be to gather a data set that includes acquittals and convictions for lesser offenses (indeed, any case that reached voir dire) and excludes cases that did not reach the jury selection stage with a death-eligible charge intact, such as cases in which a pretrial plea bargain was reached in which the maximum charge was manslaughter or in which the case was heard by a judge. (Note that the Baldus study data set used for the 39-variable core regression includes cases decided by plea bargain.) Costs to this approach are discussed above, but the study of jury behavior is worthwhile.

A final retort we have heard informally is that few cases that are indicted for murder or that reach a jury result in an acquittal or a conviction for an offense lesser than voluntary manslaughter. Empirically we are uncertain that this claim is true; Bortner and Hall (2002), for example, report that for first-degree murder cases in Arizona, approximately 20% of indicted cases and 14% of cases reaching trial fall into this category. But even if the claim were true, the issue is that a small number of cases can have a substantial influence in a well-designed observational study, depending on where those cases fall in the covariate space. Because in observational studies randomization cannot balance covariates, a researcher must typically achieve balance by using some method (say, propensity scores) to isolate subsets of data in the treated and control groups with similar covariate distributions. Rarely do these subsets span the covariate space; in other words, comparisons are possible only for certain subsets of cases in which the covariate distributions of treated and control units overlap and not all observations are members of these subsets (Fienberg & Haviland, 2003). Thus, a critical issue in selecting cases on their final adjudication is not how many acquittals (or convictions for lesser offenses) there are, but where they fall in the covariate space. When some cases selected by outcome have zero probability of being sampled, a researcher might have to rely on implausible assumptions regarding either the covariate distributions of such cases or the correctness of a parametric model (or perhaps a classification technique; see, Morton & Rolph, 2000) relating covariates to outcome.

To illustrate concepts further, we obtained the Baldus study data from ICPSR and replicated the point estimates of the centerpiece model: the 39-variable weighted logistic regression that formed the primary basis for the empirical argument in the McClesky case. A principal finding of the 39-variable model was that the coefficient for race of the defendant had a "perverse sign" (black defendants were less likely to receive capital punishment), although this coefficient was not statistically significant. (On the basis of this finding, one of us was instructed in a first-year criminal law class that the race of the defendant played no role in the trial court administration of the death penalty.)

After dividing the data into four treatment groups (black/white for victim/defendant), we examined how we might have proceeded had the 39 predictors been true covariates and had a well-defined causal question been answerable. We observed the following results. First, no comparisons involving white defendant, black victim cases were feasible, as there were too few such cases (27, with only two death sentences). Thus, the two relevant studies were a study of the effect of the victim's race (white versus black, that is, treated versus control) in black defendant cases, and a study of the effect of the defendant's race (black versus white, treated versus control) in white victim cases. Second, 38 of the Baldus study 39 predictors were binary, and the 39th (number of prior felony sentences) was nearly so, with over 90% of units having values of 0 or 1. Third, there was a difference of several standard deviations (however estimated) between the treated and control groups in the occurrence rates of several covariates. For example, in the victim-race study, the rates of homicides committed with armed robberies for treated versus control were .31 to .07; the corresponding rates for the involvement of a coperpetrator were .42 versus .12. Regression adjustment in such situations can result in biased estimates (Cochran, 1965; Cochran & Rubin, 1973). Meanwhile, other covariates had extremely low incidence rates (for example, 2 out of 657 victims in the potential victim-race study died by drowning).

These observations, and a desire to keep things simple, led us to a categorical matching strategy (Rosenbaum & Rubin, 1985), and we focus here on the defendant race study (the one using only white victim cases) as better illustrating concepts. For each of the 139 treated (black defendant) cases, we identified its 39-variable predictor vector, then searched among the 382 control (white defendant) cases for one with an identical 39-variable vector. If we found one, we stopped the search with respect to that treated unit. If we found no 39-variable match, we searched for a 38-variable match, stopping if we found one; if we found none, we looked for a 37-variable match, and so on, until we had one or more best available matches for each treated unit, where "best" was measured by the number of identically valued variables. (If there was more than one best available match, we averaged the corresponding outcomes.) The results appear in table 1. If we accept a

TABLE 1.—CATEGORICAL MATCHING RESULTS FOR WHITE VICTIM CASES, TREATED (BLACK DEFENDANT) VERSUS CONTROL (WHITE DEFENDANT)

Floor	# Trt. Units Mtchd.	% Trt. Dths.	% Cntrl. Dths.	#Unique Cntrl. Obs. Used	Max Cntrl. Wt.
Exact	6	0	0	31	.5
38	29	17.2	10.3	61	3
37	58	**24.1**	**15.9**	90	4.8
36	93	30.1	25.1	120	5.8
35	123	34.1	27.3	137	6.8
34	137	35.0	28.9	146	6.8
33	139	36.0	29.2	146	7

"Floor" reports the number (out of 39) predictors exactly matched and thus measures the quality of the match between the treated (black defendant) and control (white defendant) cases. "# Trt. Units Mtchd." denotes how many cases pass the bar set by "Floor" to be included in an analysis data set. The "% Trt. Dths" and "% Cntrl. Dths." columns detail the percentage of death sentences for each type, where "treated" represents black defendant cases and "control" represents white defendant cases. The "# Unique Cntrl. Obs. Used" column shows how many different white defendant cases were involved in the matching process. The "Max. Cntrl. Wt." column shows the largest number of times a single white defendant case was used in matching. The fractional values are due to our use of averages when more than one match of acceptable quality was available to a particular treated (black defendant) unit.

match on 37 of 39 predictors as good enough (see the bold and italicized cells in table 1), a chi-squared test for a difference in means demonstrates that the disparity between a 24.1% death rate (for defendants perceived black) and 15.9% death rate (for defendants perceived white) is not statistically significant, but it is also not perverse. The disparity is about 1.1 standard deviations, and in contrast to the findings in Baldus, Woodworth, & Pulaski (1990), the difference is in the expected direction: black defendants were more likely to receive the death penalty than white defendants. This suggests that the perverse sign in Baldus et al.'s (1990) core model logistic regression may have been due to the use of a poorly fitting model, although all race-of-defendant results may also have been due to random variation.

The results were sensitive to the hypothetical inclusion in the sample frame of a small number of the "right" kind of cases. For example, had the analysis data set included three white defendant cases resulting in acquittals and having covariate vectors closely matching black defendant cases with an outcome of death, the essentially null result for the race of the defendant study would have $p < .05$. Again, acquittals were not included in the Baldus, Woodworth, & Pulaski (1990) data set.

We have kept our discussion brief because there is no causal question under the potential outcomes framework answerable with datasets such as are available from the Baldus study. Our suggestion that we did not find an existing data set suited to answer well-posed causal questions concerning the effect of race in capital punishment under the potential outcomes framework does not imply that we think the substantive conclusions reached in the studies are necessarily wrong or that no inquiries can be answered by these data. Other questions, including descriptive inquiries or questions focusing on associations or conditional associations, can be addressed, and consumers of quantitative legal analysis might view such answers as evidence regarding the role of race in the capital punishment system. We do not engage such issues. Our focus here is on causal inference.

VII. Limitations

A shift in focus from "true" immutable characteristics to perceptions does not mean that any and all inquiries into the effect of race, sex, and so on are well defined, even those involving some aspect of randomization. Several limits are particularly important. First, if treatments are perceptions, then someone must be perceiving something. For this reason, studies that attempt causal inference of the effect of characteristics of deciders themselves, as opposed to the deciders' perceptions, are asking questions we find difficult to define, at least at present, without further theoretical development of a causal inference framework.

For example, several studies, reviewed in Boyd, Epstein, and Martin (2010), appear to suggest that randomization of judges to cases allows randomization-based inference with respect to the causal effects of judges' sex on judicial decisions. At least currently, we are less certain. The potential outcomes in this literature appear to be case outcomes resulting from assigning a male or a female judge to a case. If researchers are truly interested in the effect of sex on judging, the potential outcomes should be case decisions (or votes) by a particular judge had that judge been male versus female. In other words, what these sex-in-judging studies may really be studying is the effect of assigning a judge from group A, all of whom happen to be female, versus assigning a judge from group B, all of whom happen to be male. Such a design might allow inferences as to the effect of the group A versus B assignment, but is this a "sex" effect? True, groups A and B differ in their sex distributions, but they likely also differ in many ways: distributions of judicial ideology, political party, and number of years spent as prosecutors, for example. Should researchers attempt to adjust for these differences by matching or some other balancing technique? Presumptively such balancing should be attempted only if these variables are pretreatment: when was sex assigned to each judge? For a male judge, would being born female have made it more likely that this judge would have been of a different political party or judicial ideology, or even been a judge at all? Meanwhile, if such balancing is necessary, what was the role or relevance of the randomization of judges to cases? Lacking a perception to consider as a treatment, with the corresponding presumptive timing of treatment assignment, it does not appear that this sort of study is asking a causal question with respect to sex. It would appear, rather, that sex-in-judging studies following this design are describing differences without a causal inquiry.

A second limit of our perspective is that the decider must be relatively discrete and of manageable size. In this paper, we have used firms or actors within the trial-level criminal justice system as deciders. The relatively discrete nature of these entities has allowed both sharp articulation of a counterfactual outcome and clear identification of a timing of treatment assignment. In contrast, for example, rigorous causal inference of the effect of race on wages in the U.S. economy (Black et al., 2006) remains a poorly defined task without implausible assumptions or further theoretical development of a causal inference framework. The decider whose behavior is to be investigated appears to be the set of all employers in the United States, a large, multifaceted, and diffuse group. What does the counterfactual look like? Is SUTVA plausible? Is treatment assigned at the moment any potential U.S. employer first perceives a unit's race?

REFERENCES

American Anthropological Association, "Statement on 'Race,'" technical report (1998).

Baime, David S., "Systemic Proportionality Review Project 2004–2005 Term," New Jersey Supreme Court technical report (2001).

Baldus, David, George Woodworth, Catherine M. Grosso, and Aaron M. Christ, "Arbitrariness and Discrimination in the Administration of the Death Penalty: A Legal and Empirical Analysis of the Nebraska Experience (1973–1999)," *Nebraska Law Review* 81 (2002), 486–756.

Baldus, David C., George Woodworth, and Charles A. Pulaski Jr., *Equal Justice and the Death Penalty Penalty: A Legal and Empirical Analysis* (Boston: Northeastern University Press, 1990).

Baldus, David C., George Woodworth, David Zuckerman, Neil Alan Weiner, and Barbara Broffitt, "Racial Discrimination and the Death Penalty in the Post-*Furman* Era: An Empirical and Legal Overview, with Recent Findings from Philadelphia," *Cornell Law Review* 83 (1998), 1638–1771.

Berk, Richard A., *Regression Analysis: A Constructive Critique* (Thousand Oaks, CA: Sage, 2004).

Berk, Richard, Azusa Li, and Laura J. Hickman, "Statistical Difficulties in Determining the Role of Race in Capital Cases: A Re-Analysis of Data from the State of Maryland," *Journal of Quantitative Criminology* 21 (2005), 365–390.

Bertrand, Marianne, and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (2004), 991–1013.

Black, Dan, Amelia Haviland, Seth Sanders, and Lowell Taylor, "Why Do Minority Men Earn Less? A Study of Wage Differentials among the Highly Educated," this REVIEW 88 (2006), 300–313.

Bortner, Peg, and Andy Hall, "Arizona First-Degree Murder Cases Summary of 1995–1999 Indictments: Data Set III Research Report to Arizona Capital Case Commission," Center for Urban Inquiry, College of Public Programs, Arizona State University technical report, http://www.azag.gov/CCC/Data%20Set%20II%20Report%20June%202002.pdf (2002).

Boyd, Christina L., Lee Epstein, and Andrew Martin, "Untangling the Causal Effects of Sex on Judging," *American Journal of Political Science* 54 (2010), 389–411.

Brock, Dean, Nigel Cohen, and Jonathan Sorensen, "Arbitrariness in the Imposition of Death Sentences in Texas: An Analysis of Four Counties by Offense Seriousness, Race of Victim, and Race of Offender," *American Journal of Criminal Law* 43 (2000), 43–72.

Cochran, William G., "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society, Series A* 128 (1965), 234–266.

Cochran, William G., and Donald B. Rubin, "Controlling Bias in Observational Studies: A Review," *Sankhya—A* 35 (1973), 417–446.

Coleman, James S., Thomas Hoffer, and Sally Kilgore, *High School Achievement: Public, Catholic, and Private Schools Compared* (New York: Basic Books, 1982).

Committee on Racial and Gender Bias in the Justice System, "Final Report," Pennsylvania Supreme Court technical report, http://www.courts.state.pa.us/index/supreme/biasreport.htm (2001).

Cornfield, J., W. Haenszel, E. Hammond, A. Lilienfeld, M. Shimkin, and E. Wynder, "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute* 22 (1959), 173–203.

Cosmides, Leda, John Tooby, and Robert Kurzban, "Perceptions of Race," *Trends in Cognitive Science* 7 (2003), 173–179.

Fienberg, Stephen E., and Amelia M. Haviland, "Discussion of Statistics and Causal Inference: A Review," *Test* 12 (2003), 319–327.

Freedman, David, "Graphical Models for Causation and the Identification Problem," *Evaluation Review* 28 (2004), 267–293.

Gelman, Andrew, and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge: Cambridge University Press, 2007).

Goldin, Claudia, and Cecilia Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review* 90 (2000), 715–741.

Greiner, D. James, *Contributions to Law and Empirical Methods*, Ph.D. thesis, Harvard University, 2007.

———"Causal Inference in Civil Rights Litigation," *Harvard Law Review* 122 (2008), 533–598.

Gross, Samuel R., and Robert Mauro, "Patterns of Death: An Analysis of Racial Disparities in Criminal Sentencing," *Stanford Law Review* 37 (1984), 27–153.

Heckman, James J., "Rejoinder: Response to Sobel," *Sociological Methodology* 35 (2005), 135–162.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association* 81 (1986a), 945–960.

——— "Statistics and Causal Inference: Rejoinder," *Journal of the American Statistical Association* 81 (1986b), 968–970.

———"Causation and Race," Educational Testing Service research report RR-03-03 (2003).

Imbens, Guido W., and Donald B. Rubin, "Rubin Causal Model," in Steven N. Durlauf and Lawrence E. Blume (Eds.), *The New Palgrave Dictionary of Economics*, 2nd ed. (London: Palgrave Macmillan, 2006).

Kaufman, Jay S., "Epidemiologic Anaysis of Racial/Ethnic Disparities: Some Fundamental Issues and a Cautionary Example," *Social Science and Medicine* 66 (2008), 1659–1699.

Keil, Thomas, and Gennaro F. Vito, "Race and the Death Penalty in Kentucky Murder Trials, 1976–1991," *Advocate* 17 (1995), 5–15.

Klein, Stephen P., and John Rolph, "Relationship of Offender and Victim Race to Dealth Penalty Sentences in California," *Jurimetrics Journal* 32 (1991), 33–48.

Lenza, Michael, David Keys, and Teresa Guess, "The Prevailing Injustices in the Application of the Missouri Death Penalty (1978 to 1996)," *Social Justice* 32 (2005), 151–166.

Lopez, Ian F. Haney, *White by Law: The Legal Construction of Race* (New York: University Press, 1996).

Mealli, Fabrezia, and Donald B. Rubin, "Assumptions Allowing the Estimation of Direct Causal Effects: Discussion of 'Healthy, Wealthy, and Wise? Tests for Direct Causal Paths between Health and Socioeconomic Status' by Adams et al.," *Journal of Econometrics* 112 (2003), 79–87.

Morgan, S. I., "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning," *Sociology of Education* 74 (2001), 341–374.

Morton, Sally C., and John E. Rolph, *Racial Bias in Death Sentencing: Assessing the Statistical Evidence* (New York: Springer, 2000).

National Research Council, Panel on Methods of Assessing Discrimination, *Measuring Racial Discrimination* (Washington, DC: National Academies Press, 2004).

Neyman, Jerzey, "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," *Statistical Science* 5 (1990 reprint and translation of 1923 original), 465.

Paternoster, Raymond, and Robert Brame, "An Empirical Analysis of Maryland's Death Sentencing System with Respect to the Influence of Race and Legal Jurisdiction," Office of the Governor technical report, http://www.newsdesk.umd.edu/pdf/finalrep.pdf (2003).

Pearl, Judea, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2000).
Pierce, Glenn L., and Michael L. Radelet, "Race, Region, and Death Sentencing in Illinois, 1988–1997," *Oregon Law Review* 81 (2002), 39–96.
Radelet, Michael, and Glenn L. Pierce, "Race and Prosecutorial Discretion in Homicide Cases," *Law and Society Review* 19 (1985), 587–622.
——— "Choosing Those Who Will Die: Race and the Death Penalty in Florida," *Florida Law Review* 43 (1991), 1–34.
Rosenbaum, Paul, and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983), 41–55.
——— "The Bias Due to Incomplete Matching," *Biometrics* 41 (1985), 103–116.
Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (1974), 688–701.
——— "Bayesian Inference for Causality: The Importance of Randomization" (pp. 233–239), in *The Proceedings of the Social Science Statistics Section of the American Statistical Association* (Alexandria, VA: American Statistical Association, 1975).
——— "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics* 6 (1978), 34–58.
——— "Bias Reduction Using Mahalanobis Metric Matching," *Biometrics* 36 (1980a), 293–298.
——— "Discussion of Randomization Analysis of Experimental Data in the Fisher Randomization Test by Basu," *Journal of the American Statistical Association* 75 (1980b), 591–593.
——— "Causal Inference Through Potential Outcomes and Principal Stratification: Applications to Studies with 'Censoring' Due to Death," *Statistical Science* 21 (2006a), 299–321.
——— "Statistical Inference for Causal Effects, with Emphasis on Applications in Epidemiology and Medical Statistics," in C. R. Rao and Sandip Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (Amsterdam: North-Holland, 2006b).
——— "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials," *Statistics in Medicine* 26 (2007), 20–30.
——— "For Objective Causal Inference, Design Trumps Analysis," *Annals of Applied Statistics* 2 (2008), 808–840.
Sapiro, Virginia, "If US Senator Baker Were a Woman," *Political Psychology* 3 (1981–1982), 61–83.
Unah, Isaac, and John C. Boger, "Race and the Death Penalty in North Carolina: An Empirical Analysis: 1993–1997," University of North Carolina technical report, http://www.common-sense.org/pdfs/NCDeathPenaltyReport2001.pdf (2001).
U.S. Office of Management and Budget, "Revisions to the Standard for the Classification of Federal Data on Race and Ethnicity" (1997).
Weisburd, David, and Joseph Naus, "Report to Special Master David Baime Re Systematic Proportionality Review," New Jersey Administrative Office of the Courts technical report (2001).
Williams, Marian R., and Jefferson E. Holcomb, "Racial Disparity and Death Sentences in Ohio," *Journal of Criminal Justice* 29 (2001), 207–218.
Winship, Christopher, and Michael Sobel, "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology* 25 (1999), 659–707.
Zhang, Junni L., and Donald B. Rubin, "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by 'Death,"' *Journal of Educational and Behavioral Statistics* 28 (2003), 353–368.