
NOT ALL STATISTICS ARE CREATED EQUAL

D. James Greiner*

Replying to Steven L. Willborn & Ramona L. Paetzold, *Statistics Is a Plural Word*, 122 HARV. L. REV. F. 48 (2009), http://www.harvardlawreview.org/media/pdf/willborn_paetzold.pdf.

Dean Steven Willborn and Professor Ramona Paetzold have written a thought-provoking response¹ to my article, *Causal Inference in Civil Rights Litigation* (“*Causal Inference*”).² There is much common ground among us: All proof of causation, including especially statistical proof in the civil rights litigation context, is “messy, indirect, uncertain, and subject to varying interpretations.”³ Any use of statistics is “part art, part science,”⁴ requires “discretion and judgment”⁵ as well as “complicated (and potentially controversial) choices,”⁶ and relies on a “host of underlying assumptions.”⁷ Direct evidence (assuming this term to be well defined in civil rights) of discrimination is rarely available,⁸ and statistical proof should not be held to higher standards than nonstatistical proof.⁹ The conceptualization of discrimination should not be limited to “conscious” discrimination (again, to the extent that this term is well defined).¹⁰ Dean Willborn and Professor Paetzold’s Response, *Statistics Is a Plural Word* (“*Plural*”), thus makes clear the extent of our agreement.

Plural also makes clear our disagreements, and in doing so, illustrates many of the reasons why potential outcomes is a superior approach to regression as the latter is currently used in civil rights litigation. Statistics is a plural word, but not all statistics are created equal.

* Assistant Professor of Law, Harvard Law School. Ph.D., Department of Statistics, Harvard University, 2007; J.D., University of Michigan Law School, 1995. My sincere thanks to Jamie Lynn Dodge for providing feedback on this reply. As usual, all mistakes are my own.

¹ Steven L. Willborn & Ramona L. Paetzold, *Statistics Is a Plural Word*, 122 HARV. L. REV. F. 48 (2009), http://www.harvardlawreview.org/media/pdf/willborn_paetzold.pdf.

² D. James Greiner, *Causal Inference in Civil Rights Litigation*, 122 HARV. L. REV. 533 (2008).

³ Willborn & Paetzold, *supra* note 1, at 48.

⁴ *Id.* at 56.

⁵ *Id.* at 55.

⁶ *Id.* at 51.

⁷ *Id.* at 56.

⁸ *Id.* at 59.

⁹ *Id.* at 60.

¹⁰ *Id.* at 59 n.40.

I proceed by posing two questions that allow me to discuss the bulk of my disagreement with Dean Willborn and Professor Paetzold. First, what is a causal effect? Second, how should we think about characterizing and selecting variables? I conclude with some additional points.

I. WHAT IS A CAUSAL EFFECT?

A primary theme of *Causal Inference* was that a coherent framework for causation in the civil rights litigation context should begin by identifying the goal of the inference. We should know exactly what we want from a statistical analysis before we begin crunching numbers. Section I.C of *Causal Inference* demonstrated how many of the problems associated with regression as it is currently used in civil rights litigation stem from a failure to begin with a definition of a causal effect.¹¹ Part II of *Causal Inference* showed how, in contrast, the potential outcomes framework does begin with such a definition, and how much follows naturally from there.¹²

In defending regression as it is currently used in civil rights litigation, Dean Willborn and Professor Paetzold “agree wholeheartedly . . . that statistical evidence must be ‘honed by an insistence on a definition of a causal effect.’”¹³ Given this agreement, and given the centrality of this point to *Causal Inference*, one might expect to find a definition of a causal effect in *Plural*, or a citation to the regression literature containing one. No such definition appears. The closest thing to such a definition occurs when Dean Willborn and Professor Paetzold discuss what regression “produces” in a gender discrimination lawsuit: “a coefficient for sex that is either significant or not.”¹⁴ The singular, “a coefficient,” is not accidental, and it is telling. As Part I of *Causal Inference* explained, implicit in a definition of a causal effect in terms of a single regression coefficient are a host of unrealistic, undesirable, and unnecessary assumptions, including especially the idea that being perceived to be female hurts an upper-level manager the same as a file clerk, in the teeth of experimental, experiential, and judicial evidence to the contrary.¹⁵

The failure to begin with a definition of a causal effect (or, alternatively, the use of a definition based on a single regression coefficient) infects much of *Plural*'s subsequent discussion. For example, Dean

¹¹ Greiner, *supra* note 2, at 556–57.

¹² Specifically, a causal effect is a comparison (often a difference, meaning a subtraction) between the value of some outcome variable (such as salary) as actually observed and the counterfactual value that outcome variable would have taken had a woman been perceived to be a man, an African American perceived to be white, and so on. *See id.* at 558–59.

¹³ Willborn & Paetzold, *supra* note 1, at 56 (quoting Greiner, *supra* note 2, at 543).

¹⁴ *Id.* at 59.

¹⁵ *See* Greiner, *supra* note 2, at 548–51.

Willborn and Professor Paetzold argue that the goal in statistical analysis should be “to equalize the two [treated versus control] groups to make . . . the two groups as comparable as possible.”¹⁶ I agree. But regression does not purport to make treatment and control groups comparable. Instead, it attempts to account for systematic differences between the two groups via a mathematical model. For this to work, the mathematical model has to have not just the treatment variable (race, gender, and so forth), but also all of the other variables, well specified, meaning the math has to be very close to the truth. Part I of *Causal Inference* shows how easy it is to get the math wrong, and how badly things can go when that happens.¹⁷

Plural makes much of the contention that implementation of the potential outcomes framework involves a model, perhaps a regression model. Depending on how one defines the term “model,” this is not actually true, but grant the point for now.¹⁸ *Causal Inference* addressed

¹⁶ Willborn & Paetzold, *supra* note 1, at 57–58.

¹⁷ This distinction between attempting to make treated and control groups as comparable as possible, versus attempting to account for differences between the two groups via a mathematical model, is critical to causal inference. The former should constitute the “design” phase of an observational study; the latter should take place in the “analysis” phase. Potential outcomes separates these two phases, placing primary importance on the former, which can and should be accomplished without access to the outcome variable. Once the design phase has made the treated and control groups as comparable as possible, the analysis phase (which may or may not involve a stochastic model) can proceed. If the design phase has worked properly, much less should turn on the technique employed in the analysis phase because when treated and control groups are comparable, different analysis techniques tend to give similar answers. In this regard, one of the odder passages of *Plural* is the accusation that I conflated the design and analysis phases of a statistical analysis. See Willborn & Paetzold, *supra* note 1, at 49–50. A strength of the potential outcomes framework is the maxim “design before analysis,” meaning that in an observational study one should balance covariates by isolating subsets of comparable treated and control observations before analyzing outcomes. *Causal Inference* made explicit the distinction between design and analysis, with the former illustrated by Figure 2, Greiner, *supra* note 2, at 567, and the latter illustrated by Figure 3, *id.* at 569. The distinction is made particularly clear by noting that design can proceed without access to the outcome variable, while analysis cannot. For more detail here, see Donald B. Rubin, *For Objective Causal Inference, Design Trumps Analysis*, 2 ANNALS APPLIED STAT. 808 (2008).

In contrast, regression as it is typically used in modern civil rights litigation does not attempt the design phase, and thus does not attempt to make groups comparable. Instead, everything turns on an analysis model relating outcome to covariates and treatment. This single-step approach requires a high degree of confidence that one can express the mathematical relationship among covariates, treatment, and outcome correctly. Such confidence seems questionable when a statistician has the full range of specifications available, but it is particularly problematic when the statistician limits herself to models involving “a coefficient.”

¹⁸ Alexis Diamond and Professor Jasjeet S. Sekhon’s paper, Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies (June 12, 2008) (unpublished manuscript), available at <http://sekhon.berkeley.edu/papers/GenMatch.pdf>, cited in *Causal Inference*, Greiner, *supra* note 2, at 575 n.91, proposes a method of balancing without an explicit stochastic model. Thus, one need not proceed by propensity scores and hypothesis testing, as is suggested by *Plural*, Willborn & Paetzold, *supra* note 1, at 53 & n.17. To be clear, however, the analyst typically must employ a one-dimensional distance measure that

this point head-on; the claim was that the process of making treated (for example, male) and control (for example, female) groups comparable, a process called “balancing,” can and should be implemented without access to the outcome variable, allowing the analyst to make the judgment calls required in any statistical procedure without knowing what the results will be. I have been unable to conceive of a better way to bolster analyst expert witness neutrality. And bolstering neutrality seems particularly important in the litigation context, which provides powerful incentives for an expert to cherry-pick techniques that will favor the side that hired her, even subconsciously.¹⁹

Plural's principal response on this point is surprising: All is well with expert witness neutrality. Everything is fine the way it is. The “standard criteria used for assessing regression models” together with the adversary process are enough,²⁰ so nothing new is needed. I confess that my strong sense that not all is well in the world of expert witnesses, that courts and litigators do perceive experts as advocates masquerading as witnesses (that is, whores²¹), is an impression based on my experience as a litigator, not a conclusion from a scientific survey. Thus, I must leave it to readers to decide which is more persuasive, Dean Willborn and Professor Paetzold's contention that the existing system does an adequate job of assuring expert witness neutrality, or my view that the system could use reform.²²

allows comparison of how similar observations are to one another in their covariate vectors. Thus, I grant Dean Willborn and Professor Paetzold's point here because this seems close enough to a “model” if the term is used loosely.

¹⁹ The references cited in *Causal Inference*, Greiner, *supra* note 2, at 544 n.33, include discussions by highly ethical quantitative analysts, experienced expert witnesses, who report feeling the urge to produce an analysis favorable to “their” sides. See Franklin M. Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. AM. STAT. ASS'N 277, 285–86 (1986); Joseph B. Kadane, *Ethical Issues in Being an Expert Witness*, 4 LAW, PROBABILITY & RISK 21 (2005).

The charge that I “minimize[]” the “role of expert judgment” when I discuss the potential outcomes approach, Willborn & Paetzold, *supra* note 1, at 52, is logically inconsistent with the charge that I overstate the benefits of proceeding with the design phase of an observational study without access to the outcome variable, Willborn & Paetzold, *supra* note 1, at 53–54. It is precisely because all applications of statistical techniques to data require the exercise of judgment that it is so important that this judgment be exercised blind to the results wherever possible. If statistical analysis were a rote and cookbook exercise, there would be no place for discretion, and thus investigator bias, to operate.

²⁰ Willborn & Paetzold, *supra* note 1, at 54.

²¹ See J. Morgan Kousser, *Are Expert Witnesses Whores? Reflections on Objectivity in Scholarship and Expert Witnessing*, PUB. HISTORIAN, Winter 1984, at 5 (cited in Greiner, *supra* note 2, at 544 n.31). Note that dissatisfaction with expert witness neutrality goes back at least a century and a half, as is colorfully documented in Samuel R. Gross, *Expert Evidence*, 1991 WIS. L. REV. 1113, 1114–1116.

²² *Plural* also suggests that an opportunity for expert bias exists in deciding what variables are covariates. See Willborn & Paetzold, *supra* note 1, at 54–55. I demonstrate below that this contention rests on confusion regarding exactly what variables are covariates as opposed to intermediate outcomes, confusion the potential outcomes framework does much to eliminate.

To summarize: Dean Willborn, Professor Paetzold, and I agree that we should insist on a definition of causal effect and that statistical analysis should proceed by making treated and control groups as comparable as possible. The potential outcomes approach pursues these objectives explicitly and directly. Regression as commonly used in modern civil rights litigation does not purport to pursue these objectives at all. Most of our disagreement stems from these points.

II. HOW SHOULD WE THINK ABOUT CHARACTERIZING AND SELECTING VARIABLES?

Causal Inference explained that a primary benefit of the potential outcomes framework is its insistence that the analyst specify a timing of treatment application, with the presumption in the civil rights context²³ being that treatment application occurs at the moment the person or entity whose decisions we are attempting to assess first perceives a unit's sex, race, and so forth. The focus on timing in turn facilitates separation of "covariates," variables whose values are unaffected by treatment, from intermediate outcomes, variables whose values are a function of treatment. Covariates are the "background" variables; in the running gender/salary example, we have to make men and women comparable on these background variables in order to isolate a gender effect. If a variable's value is set prior to the time that treatment is administered, then the potential outcomes framework tells us that the variable is a covariate. The process of making treated and control groups comparable, discussed above, involves balancing covariates, something an analyst can and must do. Balancing a variable affected by treatment (called an "intermediate outcome"), in contrast, can result in the wrong answer, as *Causal Inference* makes clear using an example involving smoking, lung cancer, and death.²⁴

Dean Willborn and Professor Paetzold's response on this subject demonstrates how much we need the guidance that the potential outcomes framework provides. They point out that in my gender discrimination example, I use years of education as a covariate, reasoning that a unit's years of education will typically have been set prior to the

²³ See Greiner, *supra* note 2, at 539 n.21 (clarifying that I am speaking of intentional discrimination, not disparate impact).

²⁴ See *id.* at 564. Again, a primary goal of the potential outcomes framework here (and more generally) is removal of bias. Thus, the paradigm is not, as *Plural* would have it, "a statistical approach for reducing sources of variation." Willborn & Paetzold, *supra* note 1, at 49. *Plural*'s confusion on this point mirrors Professor Paetzold's suggestion in other work, that post-treatment variables be included in a regression so long as multicollinearity remains under control. Lack of multicollinearity concerns variance, meaning a focus on avoiding inexact answers. Avoiding post-treatment bias involves a focus on avoiding wrong answers, inexact or not. See Greiner, *supra* note 2, at 546–47, 547 n.40 (citing Ramona L. Paetzold, *Multicollinearity and the Use of Regression Analyses in Discrimination Litigation*, 10 BEHAV. SCI. & L. 207, 227 (1992)); *id.* at 564.

moment in which a firm first perceives an employee's gender.²⁵ They then contend that years of education might not really be a covariate because it "may be perceived or valued differently once the employer is aware of an employee's race or gender," and thus that "[d]istinguishing covariates from intermediate outcomes is a complex and uncertain task."²⁶

This discussion evidences substantial confusion that the potential outcomes framework clears up.²⁷ The number of years of education is a covariate. According to my hypothetical, employees finished school before applying for jobs at the firm. Thus, the number of years of education each employee actually received was decided before the firm perceived whether the employee was male or female, and an analyst should "balance" on the variable by identifying groups of men and women with similar educational backgrounds for comparison. In contrast, the abstract value an employer gives to each employee's educational background is not a covariate; this value is assigned after the employer perceives each employee's gender and, as *Plural* points out, may depend on gender. Thus, an analyst should not try to "balance" on value-assigned-to-education-by-the-employer (assuming that could be measured) because doing so could mask discrimination at work. If a firm values an additional year of education for a man more than for a woman, the firm is discriminating against women by valuing the same variable in a gender-specific way. The idea here is the same as the one in the smoking, lung cancer, and death hypothetical mentioned earlier. If, when deciding whether smoking kills, one attempts to "control for" or "balance" lung cancer incidence, then one will control or balance away deaths that smoking causes by means of lung cancer. That will make smoking look less deadly than it actually is. Analogizing the two cases, being a woman is smoking; the actual number of years of education is something that predates the decision to smoke, such as a genetic predisposition to cancer; lung cancer is the value the firm places on years of education; and failure to hire is death. The potential outcomes framework beats a clear path.²⁸

²⁵ See Willborn & Paetzold, *supra* note 1, at 54.

²⁶ *Id.*

²⁷ The confusion stems from a failure to separate causal mechanism from causal effect. See generally Paul W. Holland, Comment, *Causal Mechanism or Causal Effect: Which Is Best for Statistical Science?*, 3 STAT. SCI. 186 (1988).

²⁸ To use another example from a well-known paper: Fictional resumes listing educational achievement can be sent to employers with randomly assigned names at the top, names designed to signal blackness or whiteness (they may also signal something else, but put that aside for now). In this situation, the educational achievement levels listed on the resumes are covariates, even though the employer may attach disparate value to the listed achievement according to whether the name at the top appears to belong to an African American or a Caucasian. See Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991, 991 (2004).

More generally, *Plural* contends that I did not address how to “incorporate the full range of intermediate outcomes into the analysis.”²⁹ It is true that *Causal Inference* does not include a separate section on the subject, but the basics are there. The first question is not how to incorporate intermediate outcomes into an analysis but whether to do so. In my view, we may not always need to model all aspects of a causal mechanism or process to produce probative evidence regarding a causal effect. To illustrate: Suppose in the gender and salary discrimination example, we have access for each unit to measurements of all relevant covariates as well as to a subjective supervisor evaluation score. Stipulate also that for present purposes this evaluation score could be affected by whether the worker is perceived as male or female (that is, it is an “intermediate outcome” in the terminology of *Causal Inference* and thus should not be part of the balancing process).³⁰ Further, suppose that, after balancing on all the relevant covariates but not the evaluation score, statistical analysis suggests that men are receiving higher salaries than women in a way unlikely to be due to chance. Is this, by itself, useful evidence? Or does such evidence lack probative value unless and until we understand whether the firm is discriminating by assigning women lower evaluation scores, by undervaluing years of education for women, or by some other (unknown) mechanism? I suggest the former — that is, we do not need to understand the precise role played by every measured variable to produce probative evidence. In a perfect world, it might be desirable to have such an understanding, in the sense that the statistical presentation may be more convincing if the analyst can explain exactly how the discrimination is occurring. But tackling this question may involve strong and unrealistic assumptions, and these can lessen the potential power of evidence of statistically significant disparities (evidence which is available in the absence of a complete understanding of mechanism).³¹

Thus, in some situations, the wiser course may be not to incorporate intermediate outcomes into the analysis. In other situations, how-

²⁹ Willborn & Paetzold, *supra* note 1, at 54.

³⁰ In fact, the evaluation score may or may not be a covariate, depending on whose decisions are being examined as possibly discriminatory. See Greiner, *supra* note 2, at 581–83 (discussing this “tug-of-war” problem). I put this concern aside for now, but I do note that using regression does not allow an analyst to avoid the tug-of-war problem.

³¹ Note that if an analyst does decide to attempt to incorporate intermediate outcomes (such as subjective performance evaluations) into an analysis, the potential outcomes framework is flexible enough to allow the analysis to attempt to do so. For a detailed and technical discussion of intermediate outcomes in the potential outcomes framework, see Greiner, *supra* note 2, at 587 n.109 (gathering sources). These and subsequent articles refute *Plural*’s suggestion that “[t]he treatment of intermediate outcomes is less developed in the potential outcomes approach.” Willborn & Paetzold, *supra* note 1, at 55.

ever, the analyst must incorporate them. *Causal Inference* provides an example of this situation in the discussion of race and capital punishment. There, the causal effect of the perceived victim's race on the jury's death or life decision at the penalty phase of a capital trial was defined only for cases in which the (same) jury would have convicted for a death-eligible offense under both treatment and control (that is, victim-perceived-black and victim-perceived-white). Thus, the analyst had to incorporate the intermediate outcome (conviction of a death-eligible-offense) into the analysis.³² Otherwise, the analysts can mistake an effect of race on conviction for an effect of race on sentencing. *Causal Inference* demonstrated how to proceed here, suggesting that the analyst should focus on filling in the missing counterfactual values of the intermediate outcome before isolating those observations for which a causal effect is well defined, and proceeding with the analysis for these units.

III. ADDITIONAL POINTS

Two additional minor points: First, *Plural* correctly points out that but-for causation is not the only relevant causal concept in civil rights litigation.³³ True. *Causal Inference* never argued otherwise. My point was that the potential outcomes approach, with its focus on counterfactuals, should feel familiar to those with legal training because of its resemblance to but-for causation. Perhaps more importantly, the distinction between but-for causation versus substantial factor and other causation concepts often used in tort law will typically make little difference to a quantitative analyst. Statistics requires data representing repetitive events (for example, a series of hiring decisions, or a set of capital punishment determinations). When an analyst focuses on a set of repeated events as opposed to a single fact situation, a decision maker's repeated use of an improper reason will ordinarily become indistinguishable from but-for causation.

Second, *Plural* re-highlights what *Causal Inference* had explained in detail, that the process of "equaliz[ing treated and control] groups to make . . . the two groups as comparable as possible,"³⁴ which all agree should occur, involves recognizing (in my gender discrimination example) that for certain men, there are no comparable women, and vice versa. As a litigator, I tip my hat to Dean Willborn and Professor Paetzold's rhetorical flourish when they say that this process involves "discard[ing] some actual data on employees" and "creating a *hypothet-*

³² See Greiner, *supra* note 2, at 583–88.

³³ See Willborn & Paetzold, *supra* note 1, at 56.

³⁴ *Id.* at 57–58.

ical set of data that is not the *actual* set of data,”³⁵ but the argumentation threatens to obscure the truth. The truth is that the data being “discarded” have no information; worse, they can be highly misleading in the absence of a near-perfect model. If, for some men, there are no comparable women, what purpose is served by including these men in a statistical analysis focusing on gender? In fact, including men with no relevant female counterparts is dangerous. The running example in *Causal Inference*, particularly Figures 1-3 and accompanying text, addressed and illustrated precisely this point.³⁶

³⁵ *Id.* at 55.

³⁶ See Greiner, *supra* note 2, at 551–53, 566–71.