6

A DIAGNOSTIC ROUTINE FOR THE DETECTION OF CONSEQUENTIAL HETEROGENEITY OF CAUSAL EFFECTS

*Stephen L. Morgan**
*Jennifer J. Todd**

Least squares regression estimates of causal effects are conditional-variance-weighted estimates of individual-level causal effects. In this paper, we extract from the literature on counterfactual causality a simple nine-step routine to determine whether or not the implicit weighting of regression has generated a misleading estimate of the average causal effect. The diagnostic routine is presented along with a detailed and original demonstration, using data from the 2002 and 2004 waves of the Education Longitudinal Study, for a contested but important causal effect in educational

This research was supported by a grant from the American Educational Research Association, which receives funds for its AERA Grants Program from the U.S. Department of Education's National Center for Education Statistics of the Institute of Education Sciences, and the National Science Foundation under NSF Grant #RED-0310268. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies. We thank Youngjoo Cha, Ken Frank, David Harding, Sean Reardon, Chris Winship, anonymous reviewers, and the Editor for helpful comments and suggestions.

Direct correspondence to Stephen L. Morgan, Department of Sociology, 358 Uris Hall, Cornell University, Ithaca, NY 14853 (slm45@cornell.edu) or Jennifer J. Todd, Department of Sociology, 362 Uris Hall, Cornell University, Ithaca, NY 14853 (jjt24@cornell.edu).

*Cornell University

research: the effect of Catholic schooling, in comparison to public schooling, on the achievement of high school students in the United States.

1. INTRODUCTION

With the growth of interest in counterfactual approaches to causal analysis, simple regression estimates of causal effects are often considered suspect. At the same time, the appeal of matching and natural experiment approaches has increased substantially. One negative consequence of these changes in research practice is that some investigators may move too quickly to the estimation of matching and instrumental variable models, under the justification that these models are either more intuitive or more novel. In the end, however, the alternative estimates obtained from such models may prove to be either worse than the regression results on various statistical criteria or nearly equal to the regression estimates that were dismissed in preliminary analysis.

To minimize the risk of such missteps in analysis, it would be helpful if simple and reliable routines were available to assess, after specifying a regression model, whether the causal effect estimate obtained is consistent with the most rigorous justification for regression as a causal effect estimator: a case in which, conditional on the specification of adjustment variables, individual-level variation in the causal effect of interest is completely random. The goal of this paper is to lay out one such routine based on the literature that has developed the counterfactual approach to causal analysis.

The methods embedded within the proposed routine are not new and have diverse and overlapping origins—inverse probability weighting in survey statistics (see Kish 1965, 1987; Thompson 2002), missing data imputation and survey nonresponse adjustment via weighted complete-case analysis (see Little 1982; Little and Rubin 2002), weighting procedures in multiple regression analysis for data from stratified samples (see DuMouchel and Duncan 1983), propensity score models and general methods for modeling the probability of treatment assignment (see Rosenbaum and Rubin 1983; Rubin 2006; Rubin and Thomas 2000), direct adjustment estimators (see Rosenbaum 1987, 2002), program evaluation methods in econometrics (see Heckman and Robb 1985; Heckman and Vytlacil 2005; Imbens 2004), models of causal effect heterogeneity in econometrics (see Angrist 1998; Angrist and Krueger 1999; Heckman,

Urzua, and Vytlacil 2006), and inverse probability of treatment weighting in epidemiology (see Robins and Ritov 1997; van der Laan and Robins 2003). Similar methods have been used in applications in sociology (e.g., Brand and Halaby 2006) and reviewed in methodological work (e.g., Morgan and Winship 2007).

Given the diversity and technical depth of this background literature, our primary goal in this paper is to distill from this material a simple and accessible routine that can reveal to an analyst whether a causal effect estimate from a regression model can be given a warranted average causal effect interpretation, given the identification assumptions that one is willing to maintain. This last "given" clause is crucial, as we will present a routine that is general enough to be interpretable under alternative sets of identification assumptions for the same application (such as full ignorability, partial ignorability, or nonignorability, as we describe later). This generality promotes consideration of these alternative assumptions, which thereby facilitates clarity of thinking about their alternative appropriateness.

Our presentation strategy is to offer up front a complete demonstration of the diagnostic routine, showing all steps of the routine in considerable detail. Only thereafter do we present foundational material from the counterfactual model of causality. Our hope is that this presentation strategy advances our primary goal, which is to convince analysts unfamiliar with the fine points of the counterfactual model that (1) they already have much of the technical know-how that is required to implement the diagnostic routine and (2) having been given a clear understanding of the routine from a practical data analysis perspective, they can then better understand how the routine is grounded in and justified by the literature on counterfactual causality.

2. MOTIVATION

As a departure point, consider a general multiple regression model of the form

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e, \tag{1}$$

where Y is an interval-scaled outcome variable and X_1 through X_k are predictor variables. Estimation of the slope parameters b_1 through b_k via least squares can be motivated in a variety of ways, depending

on the goals of the analysis. The model can be estimated as part of a descriptive analysis where the goal is to obtain a best-fitting linear approximation to the population-level relationship between Y and X_1 through X_k. Alternatively, the model can be estimated as part of a full causal analysis where the goal is to identify the expected shifts in Y that would result from what-if interventions on the values of X_1 through X_k.

For the main body of this paper, we consider an intermediate case between these two extremes that has also become the focal model of the literature on counterfactual causality. For this model, the variable X_1 in equation (1) is a dummy variable D, as in

$$Y = a + b_d D + b_2 X_2 + \cdots + b_k X_k + e, \tag{2}$$

and the goal of the analysis is to estimate the causal effect on Y of shifting D from 0 to 1. In this case, the variables X_2 through X_k are considered adjustment variables that are entered into the regression equation solely to aid in the effective estimation of the causal effect b_d. Accordingly, estimates of b_2 through b_k are of secondary interest and are not typically given a causal interpretation. Moreover, under this motivation it is generally presumed that individual causal effects may vary, such that the effect on Y of shifting D from 0 to 1 is not the same for all individuals.

Given this setup, we can now pose the challenge that the diagnostic routine of this paper is designed to address. Does a regression estimate of a causal effect represented by b_d in equation (2) mask important variation in the individual-level causal effect? In particular, is a regression estimate of b_d from equation (2) interpretable as

1. the average of individual-level causal effects for all individuals,
2. the average of individual-level causal effects for individuals with $D = 1$,
3. the average of individual-level causal effects for individuals with $D = 0$,
4. none of the above?

In the counterfactual causality literature, where those for whom $D = 1$ are considered members of the treatment group and those for whom $D = 0$ are considered members of the control group, this question is stated as: Is a regression estimate of b_d interpretable as

1. the average treatment effect (ATE),
2. the average treatment effect for the treated (ATT),
3. the average treatment effect for the controls (ATC),
4. none of the above?

The diagnostic routine presented here is designed to help an analyst determine whether or not the answer to this question is "1, 2, and 3." If this is not the answer, then the answer will be "2," "3," or "4." In this case, consequential heterogeneity exists and the regression equation that gives rise to an estimate of b_d masks this heterogeneity.[1]

Notice that our conception of consequential heterogeneity is model-dependent, as it is defined with reference to a specific regression equation and its associated set of adjustment variables. Heterogeneity of this form will arise from two sources: (1) variability in the causal effect across individuals that is related to at least one of the variables in X_2 through X_k that also predicts D and (2) variability in the causal effect across individuals that is related to an unobserved variable embedded in e that predicts D, conditional on X_2 through X_k. This second source of heterogeneity can itself be separated into two categories: (1) variability that is a function of a known omitted variable (such as mental ability for models that seek to estimate the causal effect of education on earnings) and (2) variability that induces some individuals to self-select into one of the two values of the causal variable (such as an accurate individual-level forecast of the gains from participation in a program being evaluated). This last source of heterogeneity is referred to as "essential heterogeneity" by Heckman et al. (2006).[2]

In the remainder of this paper, we first offer background on the demonstration of the routine that we will present as well as the data source utilized. We then present and demonstrate each step of the routine with minimal justification from the methodological literature. Thereafter, we explain the routine using the literature on counterfactual

[1] If the answer is "1, 2, and 3," then some heterogeneity of individual-level causal effects may still exist, but it is inconsequential for most research questions because it is random with respect to D.

[2] Pearl (in a personal communication, but see also Pearl 2000) maintains that essential heterogeneity does not exist since, in principle, it should be indexed by a variable in a model that obeys his Markov condition for the existence of a causal model.

causal analysis to justify its utilization. We conclude the article with a discussion of regression, matching, and instrumental variable options for additional analysis when causal effect heterogeneity has been detected.

3. BACKGROUND FOR THE DEMONSTRATION

3.1. *The Catholic School Effect on Achievement in High School*

In the 1966 government report *Equality of Educational Opportunity*, James S. Coleman maintained that differences in resources between public schools had surprisingly small effects on student achievement. Less than two decades later, Coleman and a new set of colleagues presented evidence that private Catholic schools in the United States are more effective than public schools, even though they spend comparably less money on each pupil (see Coleman, Hoffer, and Kilgore 1982; Hoffer, Greeley, and Coleman 1985; Coleman and Hoffer 1987).

The original findings on the Catholic school effect were challenged immediately by other researchers (see Alexander and Pallas 1983, 1985; Goldberger and Cain 1982; Noell 1982; Willms 1985), but knowledge of the size of the causal effect remains important for educational policy and research. Many scholars, for example, contend that estimates are needed to inform current policies on school choice and vouchers. In addition, in the context of the present article, the Catholic school effect has also become a frequent example in the methodological literature—for the presentation of multilevel models (see Raudenbush and Bryk 2002), introductions to matching methods (see Morgan 2001), and critiques of regression practice (see Freedman 2005).

For the agenda of this paper, the application is appropriate because prior analysis of earlier data from the 1980s and 1990s suggests that consequential heterogeneity exists for models of the form of equation (2), where D is attendance at a Catholic school and X_2 through X_k are various family background, educational history, and demographic variables. As we will discuss later, consequential heterogeneity may result because of heterogeneity in the effect of Catholic schooling across strata of the adjustment variables and across strata of unobserved determinants of the decision to enter Catholic schooling.

3.2. *Data*

For the demonstration of the diagnostic routine, we analyze data from the 2002 base-year and 2004 follow-up waves of the Education Longitudinal Study (ELS), collected by the National Center for Education Statistics (NCES) of the U.S. Department of Education. The ELS is a nationally representative sample of students in public and private high schools, based on a two-stage sampling design that first draws a random sample of public and private high schools and then draws random within-school samples of sophomores. For the first follow-up in 2004, respondents were tracked to alternative destinations, and most respondents were high school seniors.

From among all 15,360 base-year ELS participants, we restrict the analysis to respondents who were enrolled in either a Catholic school or a public school during the 2001–2002 academic school year. Table 1 presents descriptive results for the data we will analyze: 1918 students who were enrolled in 95 Catholic schools and 12,025 students who were enrolled in 580 public schools (for a total $N = 13{,}943$).

Several practical features of our subsequent analysis require detailed explanation, which we provide in the appendix. None of the material in the appendix is essential for understanding the diagnostic routine. However, readers who are contemplating using the routine for a project with a similar structure—where a complex sampling design necessitates the use of a poststratification weight and where some models are estimated with an adjustment for panel attrition—should consult the appendix for practical advice. We show there how the weights we utilize in the main text of the article can be multiplied by poststratification weights and attrition-adjustment weights in order to account for design features of the data. We also discuss briefly our decisions for how to estimate standard errors for our various causal effect estimates, in light of the complex nature of the sampling design as well as the estimation of the weights deployed in the routine.

4. THE DIAGNOSTIC ROUTINE WITH A DEMONSTRATION

The diagnostic routine that we present in this section is simple and is accessible to all researchers who can properly estimate a regression model using the sampling weights that are typically made available with

survey data. As we will show, the routine is not a rigid test. Rather, it is a set of practices meant to guide a researcher to the conclusion that is most consistent with the data and with one's maintained assumptions.

A schematic diagram of the nine steps of the diagnostic routine is presented in Figure 1, categorized into three stages: (1) Estimate baseline regression results, (2) model treatment selection/assignment

TABLE 1
Means and Standard Deviations of Primary Variables Used in the Demonstration

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Math Test Scores				
IRT estimated number right (10th grade)	41.679	13.974	48.993	12.018
IRT estimated number right (12th grade)	47.640	15.048	56.084	12.802
Gain Score (12th–10th grade IRT estimated number right)	4.656	6.485	6.661	6.058
Female	.496		.475	
Race (White is the reference category)				
Black	.151		.061	
Hispanic	.165		.113	
Asian	.041		.043	
Native American	.010		.002	
Multiracial	.043		.040	
Urbanicity (Suburban is the reference category)				
Urban	.280		.584	
Rural	.209		.010	
Region (Midwest is the reference category)				
Northeast	.181		.311	
South	.344		.227	
West	.234		.165	
Family Background				
Mother's education (in years)	13.455	2.322	14.766	2.215
Father's education (in years)	13.587	2.587	15.253	2.567
SEI score of mother's occupation in 2002 (GSS 1989 coding)	44.975	12.865	50.551	12.850
SEI score of father's occupation in 2002 (GSS 1989 coding)	44.146	11.696	49.813	11.709
Family income (natural log)	10.603	1.092	11.233	.897
Family income (natural log) squared	113.605	19.825	126.986	17.041
Family income (natural log) cubed	1225.640	295.463	1441.326	267.984
Two-parent family	.746		.837	

Continued

TABLE 1
Continued

Variable	Public		Catholic	
	Mean	S.D.	Mean	S.D.
Past History (as reported by parent)				
Learning disability	.126		.068	
Ever held back	.134		.053	
Repeated 4th grade	.005		.002	
Years parents lived in current neighborhood	10.557	8.001	12.897	8.210

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).

Note: Data are weighted by the NCES poststratification weight (BYSTUWT). $N = 1918$ students enrolled in Catholic schools, and $N = 12{,}025$ students enrolled in public schools for all variables from the 10th grade. For the 12th grade math test scores and math gain scores, $N = 1660$ students enrolled in Catholic schools and $N = 8842$ students enrolled in public schools. For these two variables, the data are weighted by the NCES poststratification weight (BYSTUWT) multiplied by the inverse probability of remaining in the same school and not falling behind the usual grade for age.

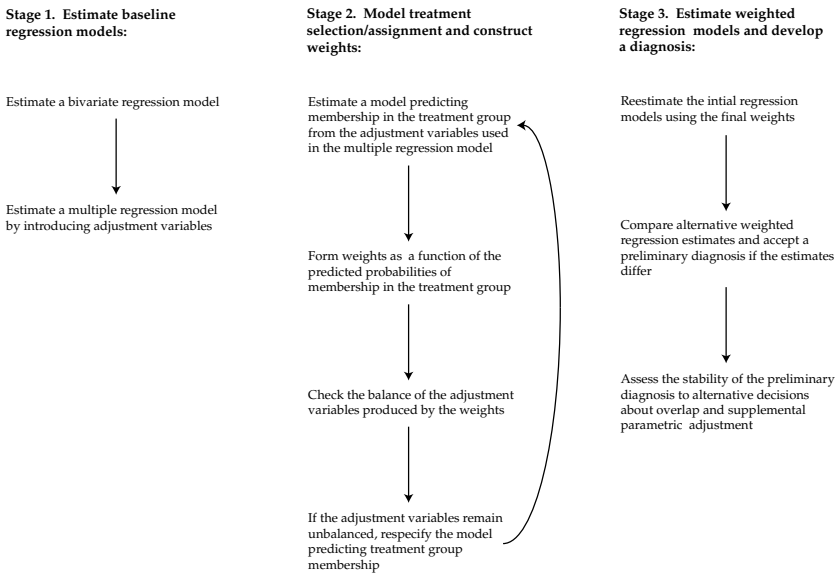

FIGURE 1. A diagnostic routine for the detection of consequential heterogeneity of causal effects.

and construct weights, and (3) estimate weighted regression models and develop a diagnosis. The ordering of the first two stages is somewhat arbitrary. We suspect that readers new to this literature on causal modeling will undertake the steps as we have laid them out in the figure and in the remainder of this paper, beginning with familiar regression techniques. Readers with more experience working within this literature are likely to begin with Stage 2, modeling treatment selection/assignment and constructing weights even before estimating baseline regression results without using the weights.

For readers unfamiliar with the counterfactual model, the rationale for the proposed routine will be presented in Section 5 of the article and is suppressed now for expediency. For readers already familiar with the literature on counterfactual causality, the motivation for the routine will be clear as the steps unfold. However, for these readers, we should note now that we intend for this routine to be undertaken even when treatment assignment is nonignorable because the results of the routine can be (and, we would argue, should be) interpreted under alternative assumptions about ignorability.[3]

Step 1: Estimate a Bivariate Regression Model

Estimate a bivariate regression equation by ordinary least squares:

$$Y = \hat{\alpha} + \hat{\delta}_{\text{OLS, bivariate}} D + \varepsilon, \tag{3}$$

where Y is an interval-scaled outcome variable and D is the causal variable of interest (equal to 1 for those exposed to one level of the cause and equal to 0 for those exposed to the other). The estimated coefficient $\hat{\delta}_{\text{OLS, bivariate}}$ is the estimated causal effect of D on Y.

Demonstration of Step 1. The bivariate regression estimates for the Catholic school effect on achievement are presented in the first row and

[3] An implicit goal of this paper is to promote a mode of empirical analysis in which results are interpreted in light of alternative plausible identifying assumptions (not the more traditional mode of analysis where identifying assumptions are asserted, models are estimated, and then conclusions are drawn in light of the maintained assumptions without explicit consideration of the extent to which the conclusions are dependent on the assumptions).

first column of each panel of Table 2.[4] The three panels offer analogous results for three related outcome variables: tenth-grade math test scores, twelfth-grade math test scores, and math gains between the tenth and twelfth grades. The estimates of $\hat{\delta}_{\text{OLS, bivariate}}$ in equation (3) are 7.314, 8.445, and 2.006 for the three different outcome variables. Because $D = 1$ for students who attend Catholic schools but $D = 0$ for students who attend public schools, each of these estimates suggests that Catholic school students have higher levels of achievement on standardized tests.

Step 2: Estimate a Multiple Regression Model by Introducing Adjustment Variables

Estimate a multiple regression model by ordinary least squares:

$$Y = \hat{\alpha} + \hat{\delta}_{\text{OLS, multiple}} D + X\hat{\beta} + \varepsilon, \tag{4}$$

where X represents observed variables thought to determine D and Y, $\hat{\delta}_{\text{OLS, multiple}}$ is the estimated causal effect of D on Y adjusted for X, and $\hat{\beta}$ is a conformable vector of estimated coefficients that correspond to the variables in X.

Demonstration of Step 2. The multiple regression estimates for the Catholic school effect on achievement are presented in the second row and first column of each panel of Table 2. Descriptive statistics for the 23 variables specified as X in equation (4) were presented earlier in Table 1. The variables in X represent the most common family background, demographic, and educational history variables utilized in school effects research. The coefficients for $\hat{\delta}_{\text{OLS, multiple}}$ are 1.479, 2.130, and

[4] As we note later in Step 7 and discuss in more detail in the appendix, the estimates in the first column of Table 2 take into account the complex sample design of the ELS. For all models reported in the first column, a poststratification weight is utilized and estimation is by weighted ordinary least squares. Furthermore, for the models in the first column in the second and third panels of Table 2, which utilize twelfth grade tests both on their own and as gain scores as the outcome variable, the estimates are based on models that also utilize an attrition-adjustment weight. Thus, the column heading in Table 2 of "No Weight" refers only to the weights introduced in subsequent steps.

TABLE 2
Catholic School Coefficients from Regression Models Predicting 10th Grade Math Test Scores, 12th Grade Math Test Scores, and Math Test Gains

Predictor Variables	Outcome Variable: 10th Grade Math Test		
	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	7.314	1.079	2.438
	(.664)	(.773)	(1.172)
Model 2: Model 1 + family background, demographics, and past history	1.479	1.083	2.415
	(.524)	(.556)	(.592)
	Outcome Variable: 12th Grade Math Test		
	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	8.445	1.420	3.288
	(.740)	(.893)	(1.356)
Model 2: Model 1 + family background, demographics, and past history	2.130	1.604	4.012
	(.601)	(.663)	(.714)
	Outcome Variable: Math Gain (12th–10th Grade Math Test)		
	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	2.006	1.018	1.975
	(.225)	(.313)	(.366)
Model 2: Model 1 + family background, demographics, and past history	1.266	1.052	2.060
	(.256)	(.291)	(.453)

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).

Note: The data for all models are weighted by the base-year poststratification weight. The data for the 12th grade math test and math gains models are weighted by a supplemental attrition-adjustment weight. See note for Table 1 for sample size details and the appendix to the article for details on how these various weights are handled in concert with the ATT and ATC weights that differentiate the models.

1.266 in the three panels, each of which is considerably smaller than the corresponding values of $\hat{\delta}_{\text{OLS, bivariate}}$ from the estimation of equation (3). Nonetheless, the values of $\hat{\delta}_{\text{OLS, multiple}}$ suggest that Catholic school students outperform public school students even after adjustments for the variables in X.[5]

[5] For now, ignore the models presented in the second and third columns of Table 2. These models will be explained in subsequent steps of the routine but are presented in this table to facilitate later comparisons.

Step 3: Estimate a Model Predicting Membership in the Treatment Group from the Adjustment Variables Used in the Multiple Regression Model

Estimate a model of treatment selection/assignment by first designating one value of the cause D a treatment state and one value a control state (i.e., designate subjects with $d_i = 1$ as members of the treatment group and subjects with $d_i = 0$ as members of the control group). Then, utilize an appropriate procedure to estimate the probability of being in the treatment state rather than the control state.[6] Finally, calculate the estimated probability of being in the treatment state for each member of the sample.

Demonstration of Step 3. Students attending Catholic school were designated the treatment group. A logit model was then estimated to predict whether individuals attend Catholic school instead of public school (i.e., are members of the treatment group instead of the control group):

$$\text{Logit}(D) = X\hat{\phi}, \tag{5}$$

where the variables specified as X here are the same 23 variables specified as X for the regression model in Step 2, and where $\hat{\phi}$ is a conformable vector of estimated coefficients. Predicted values for the estimated probability $\hat{p}_i$ that D equals 1 for each individual i were then calculated by undoing the logit transformation through the substitution of $x_i\hat{\phi}$ into $\hat{p}_i = \frac{\exp(x_i\hat{\phi})}{1+\exp(x_i\hat{\phi})}$.

The estimated logit model fit the data reasonably well, delivering a chi-squared test statistic of 404.03 with 23 degrees of freedom. The predicted probabilities $\hat{p}_i$ had a mean of .0440 and a standard deviation of .0688. The distribution was heavily skewed with a minimum of .0000182 but a maximum of .857.

Step 4: Form Weights as a Function of the Predicted Probabilities of Membership in the Treatment Group

Having defined the treatment and the control groups in Step 3 and estimated a corresponding set of predicted probabilities $\hat{p}_i$, from two sets of weights $w_{i,\,\text{ATT}}$ and $w_{i,\,\text{ATC}}$ as

[6]The dominant method of estimating these probabilities is logit modeling, although the case for more intensive data mining approaches is strengthening (see Diamond and Sekhon 2005; Hansen 2004; McCaffrey, Ridgeway, and Morral 2004).

$$\begin{aligned} &\text{For } d_i = 1\text{: } w_{i,\,\mathrm{ATT}} = 1, \\ &\text{For } d_i = 0\text{: } w_{i,\,\mathrm{ATT}} = \frac{\hat{p}_i}{1-\hat{p}_i}, \end{aligned} \tag{6}$$

and

$$\begin{aligned} &\text{For } d_i = 1\text{: } w_{i,\,\mathrm{ATC}} = \frac{1-\hat{p}_i}{\hat{p}_i}, \\ &\text{For } d_i = 0\text{: } w_{i,\,\mathrm{ATC}} = 1. \end{aligned} \tag{7}$$

These weights are equivalent in structure to survey weights that must be used to weight complex samples so that they are representative of their respective target populations.[7] When using the weight $w_{i,\,\mathrm{ATT}}$, the population-level treatment group is specified as the target population. The weight leaves the sampled treatment group unaltered (because $w_{i,\,\mathrm{ATT}} = 1$ for those in the treatment group), but it attempts to turn the control group into a representative sample of the population-level treatment group (because $w_{i,\,\mathrm{ATT}} = \frac{\hat{p}_i}{1-\hat{p}_i}$ for those in the control group). The weight $w_{i,\,\mathrm{ATC}}$ works in the opposite direction.

Demonstration of Step 4. The weights were calculated in equations (6) and (7), using the $\hat{p}_i$ from the logit model estimated in Step 3. When applied to the ELS data, the weight $w_{i,\,\mathrm{ATT}}$ leaves the Catholic school sample unaltered but weights the public school sample in an attempt to generate a sample that is representative of Catholic school students with respect to the distribution of X. Likewise, the weight $w_{i,\,\mathrm{ATC}}$ leaves the public school sample unaltered but weights the Catholic school sample in an attempt to generate a sample that is representative of public school students with respect to the distribution of X. The next step assesses the effectiveness of the estimated weights in achieving these goals.

Step 5: Check the Balance of the Adjustment Variables Produced by the Weights

In the counterfactual tradition of observational data analysis that is the foundation of this diagnostic routine, the utilization of weights to align

[7] We can also form a weight for the average treatment effect equal to $1/(1-\hat{p}_i)$ for those with $d_i = 0$ and $1/\hat{p}_i$ for those with $d_i = 1$ (see Imbens 2004 and Morgan and Winship 2007). We do not focus on this weight in this article, as it is not helpful in meeting the specific goals of the diagnostic routine.

treatment and control groups on the distribution of X does not often use the sort of intuition just provided based on estimation from survey data. Rather, because of its relationship to experimental methodology, weights such as $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ are represented as tools to balance the data so that the resulting balanced data can be analyzed as if they have been generated by a randomized experiment.

In this regard, the variables in X are said to be balanced with respect to the treatment variable D if

$$\Pr[X \mid D = 1] = \Pr[X \mid D = 0]. \tag{8}$$

Perfect balance requires that all moments of these distributions be exactly the same in the treatment and control groups, with all departures between the two being small enough to be attributable to finite sample bias. If the variables in X are two-valued indicator/dummy variables, then the means alone must be equal for balance to be achieved. But, if X includes many-valued variables, then all features of their full distributions across treatment and control groups must be equal for the data to be considered fully balanced.

Demonstration of Step 5. The raw data are substantially unbalanced, as shown in the means and standard deviations that were reported earlier in Table 1. In general, public school students are less advantaged and are more heterogeneous with respect to the characteristics in X. For example, the mean of mother's education in years is 13.455 for those in public schools but 14.766 for those in Catholic school. The mean of the log of family income is 10.603 for those in public schools but 11.233 for those in Catholic schools. Moreover, the dispersion of the log of family income is substantially different as well; its standard deviation is 1.092 for those in public schools but only .897 for those in Catholic schools.

To assess the degree of balance achieved by the weights formed in Step 4, a metric of balance must be constructed. The first metric we use is an average of standardized mean differences across treatment and control groups (see Rubin 1973), which can be constructed with different weights and then compared across weighting schemes. The standardized difference of the mean for each variable in X is calculated as

$$\frac{|\bar{x}_{i,d_i=1} - \bar{x}_{i,d_i=0}|}{\sqrt{\frac{1}{2}\mathrm{Var}[x_{i,d_i=1}] + \frac{1}{2}\mathrm{Var}[x_{i,d_i=0}]}}, \tag{9}$$

where $\bar{x}_{i,d_i=1}$ is the mean for those in the treatment group, $\bar{x}_{i,d_i=0}$ is the mean for those in the control group, $\text{Var}[x_{i,d_i=1}]$ is the variance for those in the treatment group, and $\text{Var}[x_{i,d_i=0}]$ is the variance for those in the control group. Equation (9) yields a scaled absolute difference in the mean of a variable in X across the treatment and control groups. These values can be combined across all variables in X in order to construct an average standardized difference of means. The average standardized difference of means can be calculated under different weighting schemes in order to compare the relative performance of alternative weights in achieving balance.

Because balance is not just a property of the means of variables but also of higher moments of the distributions, we used a second metric of balance for variables that are not two-valued indicator/dummy variables. For this metric, we change equation (9) slightly, substituting standard deviations in the treatment and control groups for $\bar{x}_{i,d_i=1}$ and $\bar{x}_{i,d_i=0}$. The modified version of equation (9) then yields a scaled absolute difference in the standard deviation of a variable in X across the treatment and control groups. Because these values are standardized, they can also be combined across alternative variables in X in order to construct an estimate of the average standardized difference in standard deviations.[8]

For the ELS data, we first calculated the baseline level of balance by estimating the average standardized difference of means for the variables in X without using the weights formed in Step 4. The means of the variables (as well as the corresponding standard deviations for variables that take on more than two values) are reported in Table 1, separately for those in Catholic and public schools. We then calculated the balance after using the two separate weights $w_{i,\,\text{ATT}}$ and $w_{i,\,\text{ATC}}$. The weights succeeded in producing substantial balance, reducing the average standardized difference of means from .350 to .00634 when using $w_{i,\,\text{ATT}}$ and to .111 when using $w_{i,\,\text{ATC}}$. The average standardized difference of standard deviations also fell substantially from .0715 to .0391 when using $w_{i,\,\text{ATT}}$ and to .0287 when using $w_{i,\,\text{ATC}}$. The increase in balance that results from employing $w_{i,\,\text{ATT}}$ and $w_{i,\,\text{ATC}}$ is substantial, but the balance

[8] In principle, we could move on to higher moments of the distributions, assessing skewness next. We stop at the second moment here. Note, however, that we consider the mean and standard deviation of log family income, its square, and its cube. Thus, for this variable, we attempt to match up far more than just its expectation and variance.

is not perfect. The remaining imbalance suggests that respecifying the model of treatment selection/assignment may be worthwhile.

Step 6: If the Adjustment Variables Remain Unbalanced, Respecify the Model Predicting Treatment Group Membership Until No Further Improvement in Balance Can Be Obtained (and Repeat Steps 3 through 5)

The initial specification of the model of treatment selection/assignment in Step 3 was borrowed from the specification of adjustment variables in X for the multiple regression in Step 2. The goal of the present step is to enrich the parameterization of the treatment selection/assignment model in an attempt to construct weights that further improve the balance on the variables in X when the weights are deployed. Accordingly, interactions between the variables in X not already included in the regression specification, as well as transformations of the original variables, should be considered.

Although various data mining procedures can be wedded to balancing metrics in pursuit of a best possible model (see Diamond and Sekhon 2005; Rosenbaum 2002), much progress is possible with controlled trial-and-error methods. We recommend a forward selection procedure whereby interactions that have some justification in theory and past research are added progressively until improvements in balance cease to arise.

Demonstration of Step 6. The original logit model from Step 3 fit the data reasonably well and also provided good balance. However, a better fit was available that also yielded weights that provided even better balance. We added 75 interaction terms to the initial logit model that predicted Catholic school attendance. The predicted probabilities $\hat{p}_i$ from this new model have a mean of .0440, a standard deviation of .0756, a minimum of 0, and a maximum of .892.[9]

[9]The model was so predictive that a number of cases were completely determined. In particular, 2406 public school students were given predictive values arbitrarily close to 0 by the program (though no Catholic schools students were deemed completely determined and given values arbitrarily close to 1). These public school students were mostly low-SES rural students. Putting forward a specification of this form is tantamount, as in Morgan (2001), to matching perfectly on nonrural student status.

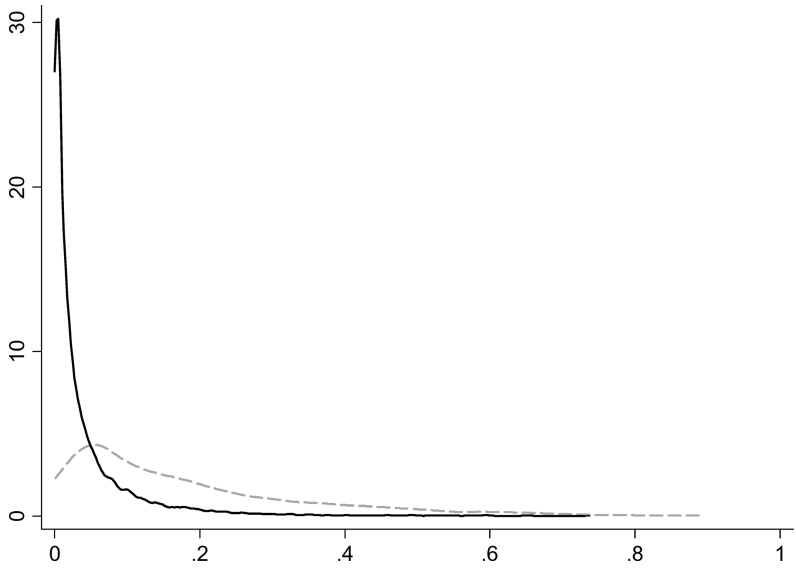

FIGURE 2. Kernel density estimates of the estimated propensity score, calculated separately for public school students (black solid) and Catholic school students (gray dashed).

Figure 2 presents kernel density estimates of these predicted probabilities, separately for those in Catholic schools and public schools. There is substantial overlap in the predicted probabilities, but there are no public school students with $\hat{p}_i$ greater than .738 and no Catholic school students with $\hat{p}_i$ less than .00115. If we could focus in very closely on the tails of the densities, we would be able to see that there are 6 Catholic school students with $.738 < \hat{p}_i \leq .892$ who have no counterparts among public school students as well as 2739 public school students with $0 \leq \hat{p}_i < .00115$ who have no counterparts among Catholic school students.

By the common support standards that prevail in observational data analysis, these data would be regarded as characterized by sufficient overlap for analysis to be worthwhile, since 1912 of the 1918 treatment cases have $\hat{p}_i$ within the range of $\hat{p}_i$ estimated for the control cases. However, there is enough of a lack of overlap that some caution is in order, especially when making inferences about how public school students would fare if they were instead enrolled in Catholic schools. We will discuss these concerns in more detail later when offering estimates restricted to the region of overlap (i.e., the common support) where $.00115 \leq \hat{p}_i \leq .738$.

As just mentioned, the revised weights yielded slightly more balance when applied to the data. In particular, the average standardized difference of means fell further to .00437 when using $w_{i,\,\mathrm{ATT}}$ and to .0899 when using $w_{i,\,\mathrm{ATC}}$. Likewise, the average standardized difference of standard deviations also fell to .0166 when using $w_{i,\,\mathrm{ATT}}$ and to .0229 when using $w_{i,\,\mathrm{ATC}}$.

To give a better sense of how well these weights have succeeded in producing balance (and for two different weighting schemes), we present the weighted means of each of the variables in X for Catholic and public school students in Table 3. The differences between each of the two panels can be directly compared to the raw differences reported earlier based on Table 1. For example, the unbalanced raw difference in mother's education between Catholic and public school students is 1.311 (i.e., |14.766 − 13.455| from Table 1), whereas the difference is reduced to .006 when using $w_{i,\,\mathrm{ATT}}$ (i.e., |14.766 − 14.772| from Panel A of Table 3) and .121 when using $w_{i,\,\mathrm{ATC}}$ (i.e., |13.576 − 13.455| from Panel B of Table 3).

The only variables that proved difficult to balance were some of the categorical variables, especially urbanicity and region because of the geographic distribution of Catholic schools. Nonetheless, the balance achieved by these weights is impressive.[10] And, as we show in subsequent steps of the routine, the remaining imbalance can be handled by supplemental parametric adjustment within a weighted regression framework.

Step 7: Reestimate the Initial Regression Models Using the Final Weights

Estimate the bivariate regression in equation (3) from Step 1 and the elaborated multiple regression in equation (4) from Step 2 using the

[10]Perfect balance is not needed to warrant causal inference if it can be assumed that the average treatment effect of interest only depends on some particular features of the distribution of X. Even so, perfect balance for the full distribution of X is the standard for which an analyst should strive. Rubin (2006) discusses these issues, and he concludes with with the advice: "Of course, at some point, this sort of [perfect balance] assessment must terminate, because no matter how large the samples, the investigator will almost certainly not be able to achieve this balance for all covariates and their interactions simultaneously, and higher order terms in prognostically minor covariates are clearly less important than prognostically important ones, and so scientific judgment must enter the process, just as it does when designing a randomized experiment" (p. 462).

TABLE 3
Means and Standard Deviations of Primary Predictor Variables, Weighted Separately by the ATT and ATC Weights from the Final Estimation of the Treatment Selection/Assignment Model

A. ATT Weight

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Female	.476		.475	
Race (White is the reference category)				
Black	.062		.061	
Hispanic	.111		.113	
Asian	.043		.043	
Native American	.001		.002	
Multiracial	.042		.040	
Urbanicity (Suburban is the reference category)				
Urban	.583		.584	
Rural	.010		.010	
Region (Midwest is the reference category)				
Northeast	.310		.311	
South	.230		.227	
West	.162		.165	
Family Background				
Mother's education (in years)	14.772	2.202	14.766	2.215
Father's education (in years)	15.254	2.566	15.253	2.567
SEI score of mother's occupation in 2002 (GSS 1989 coding)	50.563	12.773	50.551	12.850
SEI score of father's occupation in 2002 (GSS 1989 coding)	49.673	11.683	49.813	11.709
Family income (natural log)	11.241	.844	11.233	.897
Family income (natural log) squared	127.065	16.502	126.986	17.041
Family income (natural log) cubed	1441.917	261.939	1441.326	267.984
Two-parent family	.834		.837	
Past History (as reported by parent)				
Learning disability	.068		.068	
Ever held back	.051		.053	
Repeated 4th grade	.002		.002	
Years parents lived in current neighborhood	12.937	8.964	12.897	8.210

Continued

weight $w_{i,\,\mathrm{ATT}}$ and then again using the weight $w_{i,\,\mathrm{ATC}}$. No specialized software is required for this step, and the weights are treated exactly as if they are sampling weights. For example, in software such as Stata, they can be specified as *pweights* within the standard regression routine.

TABLE 3
Continued

B. ATC Weight

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Female	.496		.527	
Race (White is the reference category)				
Black	.151		.195	
Hispanic	.165		.126	
Asian	.041		.059	
Native American	.010		.011	
Multiracial	.043		.046	
Urbanicity (Suburban is the reference category)				
Urban	.280		.326	
Rural	.209		.051	
Region (Midwest is the reference category)				
Northeast	.181		.277	
South	.344		.294	
West	.234		.161	
Family Background				
Mother's education (in years)	13.455	2.322	13.576	2.346
Father's education (in years)	13.587	2.587	13.803	2.677
SEI score of mother's occupation in 2002 (GSS 1989 coding)	44.975	12.865	45.875	13.099
SEI score of father's occupation in 2002 (GSS 1989 coding)	44.146	11.696	44.051	11.706
Family income (natural log)	10.603	1.092	10.621	1.030
Family income (natural log) squared	113.605	19.825	113.867	19.360
Family income (natural log) cubed	1225.640	295.463	1228.936	292.933
Two-parent family	.746		.713	
Past History (as reported by parent)				
Learning disability	.126		.135	
Ever held back	.134		.118	
Repeated 4th grade	.005		.004	
Years parents lived in current neighborhood	10.557	8.001	10.720	6.893

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).
Note: See Table 1.

Although using weights to adjust for differences between the treatment and control groups may appear awkward at first exposure, the procedure is entirely straightforward. As noted earlier, the weights $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ are equivalent in structure to survey weights that are

used to weight complex samples so that they are representative of their respective target populations.[11] We showed earlier how weights adjust mean differences between the treatment and control groups (see our comparison of Tables 1 and 3). For models without additional variables entered as covariates, the adjustment is exactly the same. When the variables that are used to generate the weights via logit modeling are also included in the regression model as covariates, the same logic applies but the models also attempt to adjust for any remaining imbalance in X due to data sparseness and misspecification of the model that was used to estimate the weights.[12]

Demonstration of Step 7. Weighted variants of $\hat{\delta}_{\text{OLS, bivariate}}$ and $\hat{\delta}_{\text{OLS, multiple}}$ in equations (3) and (4) are reported in the second and third columns of the three panels of Table 2. As detailed in the appendix, the complex sample design of the ELS necessitated using the weights $w_{i,\text{ ATT}}$ and $w_{i,\text{ ATC}}$ along with a poststratification weight (and, for the models that utilize twelfth grade tests both on their own and as gain scores, also an attrition-adjustment weight). Finally, as for the regression models reported in Table 2, we calculated heteroskedastic-consistent standard errors with a supplemental adjustment for the clustering of students within schools. The estimated standard errors are therefore comparable across columns.

Step 8: Compare Alternative Weighted Regression Estimates and Accept a Preliminary Positive Diagnosis if the Estimates Differ

No automatic procedure is available to determine whether the regression estimates are sufficiently different to conclude that causal effect heterogeneity is present. Scholars must use standard statistical inference

[11]The intuition here goes in both directions. One way to test for the necessity of using survey weights for a regression model is to test whether the weights predict the outcome in an unweighted regression equation where the weights are added as a supplementary variable (perhaps as a polynomial and interacted with the other variables in X; see DuMouchel and Duncan [1983]).

[12]See Imbens (2004), Morgan and Winship (2007, ch. 5), and van der Laan and Robins (2003) for additional explanation of the weighting approach to adjustment. Kish (1987) and Pearl (2000) provide complementary explanations of the basic weighting approach that do not rely directly on the potential outcomes framework.

procedures and consider other research bearing on the same research question. For this reason, we explain this step in the routine by moving immediately to our process of judgment for the demonstration offered.

Demonstration of Step 8. Consider first the weighted multiple regression model for the Catholic school effect on the tenth grade math test. The estimated coefficient is 1.083 with a standard error of .556 when the $w_{i,\,\mathrm{ATT}}$ weight is utilized. In contrast, when the $w_{i,\,\mathrm{ATC}}$ weight is utilized, the coefficient increases to 2.415 with a standard error of .592. For all analogous comparisons of models, the pattern is very similar. The models suggest that the Catholic school effect is larger for the types of students who attend public schools than for the types of the students who attend Catholic schools.

Are these differences large enough to be considered meaningful? A comparison of the 95-percent confidence intervals for each estimate may suggest not. Consider the regression model for the Catholic school effect on the tenth grade math test. Here, the 95-percent confidence interval is $(-.007, 2.173)$ for the estimate using $w_{i,\,\mathrm{ATT}}$ and $(1.255, 3.575)$ for the estimate using $w_{i,\,\mathrm{ATC}}$. Clearly, these intervals overlap. Scientific judgment, however, suggests that this overlap should not lead researchers to conclude that there are no substantive differences of importance, as we now explain.

First, the difference between the two point-estimates is substantively large at 1.332; this difference suggests that the average effect of Catholic schooling is 123 percent larger for those who typically attend public schools than for those who typically attend Catholic schools (i.e., $\frac{2.415-1.083}{1.083} = 1.23$). The 95-percent confidence interval for the difference, based on a standard error of .812, is $(-.260, 2.924)$. This confidence interval is dominated by positive probability mass and suggests that values for the difference ≥ 2.664 are just as likely as values ≤ 0.[13]

Second, the difference of 1.332 is consistent with much past work on this substantive question (see Bryk, Lee, and Holland [1993]; Hoffer, Greeley, and Coleman [1985]; Morgan [2001]; and Neal [1997], all of which will be discussed later in this article). Evolving interpretive

[13] Moreover, the estimated standard error on which this confidence interval is based does not take into account that the two estimates were generated from the same sample. The confidence interval $(-.260, 2.924)$ is, in fact, a bit too wide (but not by much given the size of the available sample). In other applications, a same-sample correction may be more consequential.

standards in statistical inference demand that such prior information be considered. If a full Bayesian posterior were generated, the lower end of the frequentist confidence interval, −.260, would be judged too negative as guidance for further research.

For these two reasons, we judge the difference between the weighted regression estimate of the average treatment effect for the treated and the average treatment effect for the controls to indicate that, on average, students who are more likely to attend Catholic schools are the least likely to benefit from doing so. Accordingly, our preliminary diagnosis is that the average causal effect estimates suggested by the baseline unweighted regression models in Table 2 mask underlying heterogeneity of the causal effect that is both substantial and consequential.

Step 9: Assess the Stability of the Preliminary Diagnosis to Alternative Decisions

In this step, researchers should reflect on all decisions made in earlier steps, seeking to determine whether plausible alternative decisions would have generated the opposite preliminary diagnosis in Step 8. Two important decisions from prior steps must be examined in applications such as ours: (1) the handling of overlap issues in the estimation of the weighted regression models and (2) the selection of variables included in X for supplemental parametric adjustment in the weighted multiple regression models.[14]

Demonstration of Step 9. For the estimation of the weighted regression models with $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ (columns 2 and 3 of Table 2), we used all sample members, recognizing, however, that with respect to $\hat{p}_i$, there are 6 Catholic school students who have no counterparts among public school students and 2739 public school students who have no counterparts among Catholic school students.

Table 4 presents all of the models from Table 2 again, but this time the estimation sample is restricted to the region of overlap on the

[14]An examination of this sort is often referred to as a sensitivity analysis, although there is disagreement in the literature on whether the phrase "sensitivity analysis" should be restricted only to targeted examinations of the plausibility of alternative assertions of ignorability, as defined later.

TABLE 4
Catholic School Coefficients from Regression Models Predicting 10th Grade Math Test Scores, 12th Grade Math Test Scores, and Math Test Gains, Restricted to the Region of Overlap (i.e., the Common Support)

	Outcome Variable: 10th Grade Math Test		
Predictor Variables	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	7.009	1.022	2.183
	(.686)	(.771)	(1.186)
Model 2: Model 1 + family background, demographics, and past history	1.345	1.054	2.446
	(.524)	(.555)	(.586)
	Outcome Variable: 12th Grade Math Test		
	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	7.973	1.379	2.851
	(.766)	(.894)	(1.370)
Model 2: Model 1 + family background, demographics, and past history	2.000	1.585	3.978
	(.604)	(.663)	(.669)
	Outcome Variable: Math Gain (12th–10th Grade Math Test)		
	No Weight	ATT Weight	ATC Weight
Model 1: Dummy for Catholic school only	1.859	1.025	1.821
	(.229)	(.313)	(.369)
Model 2: Model 1 + family background, demographics, and past history	1.242	1.062	1.954
	(.261)	(.290)	(.345)

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).

Note: The data for this table are analyzed in exactly the same manner as for Table 2, but some Catholic school students and some public school students were dropped before the models were estimated because they are not in the union of the two estimated ranges of the propensity scores for the two groups. For the 10th grade models, 6 Catholic school students and 2739 public schools students were dropped, reducing the *N*s to 1912 and 9286 respectively. For the 12th grade and math gains models, 5 Catholic school students and 2028 public schools students were dropped, reducing the *N*s to 1655 and 6814 respectively.

estimated probability of treatment selection/assignment, $.00115 \leq \hat{p}_i \leq .738$. For the tenth grade math test score models, the sample size is reduced from 13,943 to 11,198. For the twelfth grade math test score and math gains models, the sample size is reduced from 10,502 to 8469. A comparison of the results in Table 4 to those reported earlier in Table 2

shows that the preliminary diagnosis is stable with respect to this decision. The point-estimates of the respective average causal effects change slightly, but, in general, the same pattern holds with estimates of the average treatment effect for students in Catholic schools remaining substantially lower than estimates of the average treatment effect for students in public schools.

As we discuss later, in many applications this step will yield alternative results. If a difference between the estimates that utilize $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ changes substantially when the two weighted regression models are restricted to the region of overlap (i.e., the common support), then heterogeneity in the treatment effect is likely to exist. However, it is unlikely that we can model this heterogeneity effectively because of a lack of overlap in the region where the heterogeneity exists. This would be the case if, for example, the incomparable members of the treatment group (i.e., those whose values for the propensity score lie outside of the range of the propensity score for the control group) differ substantially on the outcome from the members of the treatment group who have counterparts in the control group. In such a case, the preliminary diagnosis that heterogeneity exists would be affirmed, but the prospects for effectively estimating the average treatment effect for either the treated or the controls would diminish substantially. The goals of analysis would then shift toward estimating the average causal effect for a subset of either the treated or the control groups.

In the present application, the relative difference between the estimates that utilize $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ did not change substantially, and thus regression-based extrapolation outside of the region of overlap may be feasible and may allow for the effective estimation of the average treatment effect for the treated and/or for the controls. Thus, these results suggest that heterogeneity exists and that it is amenable to further analysis.

Consider now the consequences of the choice of variables for supplemental adjustment. For the weighted regression models that implement supplemental adjustment for remaining imbalance (i.e., row 2 in each panel of columns 2 and 3 in Table 2), we chose not to include any additional variables beyond the 23 variables in X that were used in the initial unweighted regression model.

The original literature on the Catholic school effect considered slightly different variables for regression adjustment than the variables that we included in X for the original multiple regression (that is,

Model 2 in Table 2). Supplemental adjustments were often performed with variables such as students' educational expectations and parental involvement. Coleman and his colleagues recognized that these variables were not clearly "prior to" Catholic school attendance and thus were likely influenced by the posited causal effect itself. Yet they wanted to show that, even adjusting for these variables, the apparent effects of Catholic schooling persisted in their models.

In our case, with the goal of completing the diagnostic routine, we are not as interested in determining how much the estimated causal effects are reduced when additional adjustment variables are entered into the various regression models (although were they to change in unexpected ways, such as vanishing entirely, one of a number of reasonable interpretations would need to be advanced). Rather, we are interested in determining whether the inclusion of additional adjustment variables in the weighted regressions would change our preliminary diagnosis from Step 8.

With this goal in mind, Table 5 presents the means and standard deviations of variables for educational expectations and parental involvement in school. The first panel of Table 5 presents differences without applying either weight, which shows that students attending Catholic schools are expected to obtain more years of postsecondary schooling (nearly a year in students' own expectations and almost as much in parents' expectations). In addition, more than half of all parents of Catholic school students volunteer at their schools, which is twice as high as the rate for the parents of public school students.

The second and third panels then apply the two weights to the variables to show that both $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ balance these variables to some degree because educational expectations and parental involvement are fairly strongly related to the variables that were used in the prior logit model that generated the weights. Yet, substantial imbalance remains because the probability of treatment assignment is not directly modeled as an outcome of expectations or parental involvement in earlier steps. Thus, regression models that introduce these supplemental variables should reduce the average causal effect estimates.

As shown in Table 6, the variables for expectations and parental involvement reduced the causal effect estimates by a substantial amount. However, the effects remain positive and in the same pattern. Most important for the diagnostic routine, the estimated average effect using the

TABLE 5
Means and Standard Deviations of Additional Predictor Variables, Without Weighting and then Weighted Separately by the ATT Weight and the ATC Weight from the Final Estimation of the Treatment Selection/Assignment Model

A. No Weight

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Educational expectations for student (in years)				
Student	16.471	2.247	17.429	1.774
Mother	16.501	2.236	17.160	1.809
Father	16.410	2.277	17.151	1.859
Parent volunteer at school (as reported by parent)	.251		.518	

B. ATT Weight

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Educational expectations for student (in years)				
Student	17.055	2.006	17.429	1.774
Mother	17.011	1.950	17.160	1.809
Father	16.994	1.963	17.151	1.859
Parent volunteer at school (as reported by parent)	.309		.518	

C. ATC Weight

	Public		Catholic	
Variable	Mean	S.D.	Mean	S.D.
Educational expectations for student (in years)				
Student	16.471	2.247	17.269	1.838
Mother	16.501	2.236	16.926	2.138
Father	16.410	2.277	16.686	2.185
Parent volunteer at school (as reported by parent)	.251		.462	

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).

weight $w_{i,\,\mathrm{ATT}}$ remains smaller than the estimated average effect using the weight $w_{i,\,\mathrm{ATC}}$. Moreover, the relative difference in the estimates is larger for Model 3 in Table 6 than for Model 2 in Table 2.

In sum, our preliminary diagnosis is supported by the supplemental analyses reported in Tables 4 through 6. Alternative decisions about overlap issues and supplemental regression adjustment did not

TABLE 6
Catholic School Coefficients from Regression Models Predicting 10th Grade Math Test Scores, 12th Grade Math Test Scores, and Math Test Gains, Including Additional Covariates

	Outcome Variable: 10th Grade Math Test		
Predictor Variables	No Weight	ATT Weight	ATC Weight
Model 3: Model 2 + expectations and parental involvement	.792 (.521)	.224 (.536)	1.256 (.621)
	Outcome Variable: 12th Grade Math Test		
	No Weight	ATT Weight	ATC Weight
Model 3: Model 2 + expectations and parental involvement	1.581 (.599)	.823 (.625)	2.766 (.738)
	Outcome Variable: Math Gain (12th–10th Grade Math Test)		
	No Weight	ATT Weight	ATC Weight
Model 3: Model 2 + expectations and parental involvement	1.245 (.261)	.977 (.289)	1.937 (.439)

Source: Education Longitudinal Study of 2002 (2002 and 2004 Waves).
Note: See Table 2.

alter the relative sizes of the weighted regression estimates that utilize $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$. As a consequence, the preliminary diagnosis is supported, and we conclude that consequential heterogeneity of the causal effect is present. We explain next why this interpretation is warranted.

5. THE PROPOSED DIAGNOSTIC ROUTINE EXPLAINED

Before using the counterfactual model to justify the routine, a presentation of the primary features of the counterfactual model is required. After offering relevant background material, the foundation of the routine in the principles of the counterfactual model will be explicated. The concluding diagnosis for the demonstration of the last section will then be reiterated and expanded, using concepts that are drawn directly from the counterfactual model.

5.1. *Background: The Counterfactual Tradition of Causal Modeling*

The most important foundational work on the counterfactual model of observational data analysis was completed in statistics and econometrics (see the citations offered in Heckman 2000, Manski 1995, Rosenbaum 2002, and Rubin 2005 to their own work and that of their respective predecessors). More detail on the following background presentation is available in Morgan and Winship (2007). Other presentations written for sociologists and political scientists are also available (see King, Keohane, and Verba 1994; Sobel 1996, 2000).

Outcomes, Treatment Groups, and the Average Causal Effect. For the Catholic school demonstration offered in the previous section, the outcome of interest Y is a score on a standardized test. Within the counterfactual tradition, an outcome variable such as this one is given a definition that is based on potential outcomes associated with the causal effect of interest. Accordingly, y_i^1 is the potential outcome in the treatment state (Catholic school) for individual i, and y_i^0 is the potential outcome in the control state (public school) for individual i. The individual-level causal effect of the treatment is then defined as

$$\delta_i = y_i^1 - y_i^0, \tag{10}$$

which for the demonstration is the causal effect of Catholic schooling instead of public schooling for individual i. Similarly, Y^1 and Y^0 are population-level potential outcome random variables, and the average treatment effect (ATE) in the population is

$$E[\delta] = E[Y^1 - Y^0], \tag{11}$$

where $E[.]$ is the expectation operator from probability theory.

Similar to the way in which the variable D was utilized for the regression in the demonstration earlier, the treatment and control groups are defined by the random variable D. This variable takes on values of $d_i = 1$ for each individual i who is a member of the treatment group (observed attending a Catholic school) and $d_i = 0$ for each individual i who is a member of the control group (observed attending a public school).

Given these definitions of Y^1, Y^0, and D, the observed outcome variable Y is defined as

$$Y = DY^1 + (1 - D)Y^0. \tag{12}$$

Thus, the observed values for the variable Y are $y_i = y_i^1$ for individuals with $d_i = 1$ and $y_i = y_i^0$ for individuals with $d_i = 0$. Accordingly, the math test variable for the demonstration is equal to the potential outcome under the treatment state for Catholic school students but the potential outcome under the control state for public school students.

Ignorability and Identification of the ATE. In the counterfactual tradition, treatment selection/assignment patterns are represented by the general conditional probability distribution

$$\Pr[D = 1 \mid S], \tag{13}$$

where S denotes *all* variables that systematically determine treatment selection/assignment. The specific values of equation (13) are known as propensity scores. They are the true probability that an individual with characteristics S will be in the treatment group ($d_i = 1$) rather than the control group ($d_i = 0$).

Complete observation of S allows a researcher to assert that treatment selection is "ignorable" and then to estimate consistently the average treatment effect in equation (11). The general idea here is that within strata defined by S the remaining variation in the treatment variable D is completely random and hence that the process generating this remaining variation is ignorable. We now explain the idea of ignorability more formally.

The concept of ignorability is a conditional variant of the independence assumption

$$(Y^0, Y^1) \perp\!\!\!\perp D, \tag{14}$$

where the symbol $\perp\!\!\!\perp$ denotes independence and where the parentheses enclosing Y^0 and Y^1 stipulate that D must be jointly independent of all functions of the potential outcomes (such as δ). If equation (14) holds, then learning whether individuals are in the treatment group or in the control group yields no information whatsoever about the sizes of individual-level treatment effects or average treatment effects across subsets of observed individuals.[15]

[15]All that is learned in this case are the values of d_i, the values of y_i, and whether each individual's y_i is equal to either y_i^1 or y_i^0.

Ignorability of treatment selection/assignment is weaker than independence of potential outcomes from D as represented by equation (14). Ignorability of treatment selection/assignment holds in the case where:

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid S, \tag{15}$$

and where S is fully observed. The treatment assignment mechanism is ignorable when the potential outcomes (and any function of them, such as δ) are independent of the treatment variable within strata defined by all combinations of values on the observed variables in S that determine treatment selection via equation (13). If all of the variables in S in equation (13) are observed, the ATE in equation (11) can be estimated by basic conditioning techniques.[16]

Conditional Average Treatment Effects and Their Identification. The unconditional ATE is not the only average causal effect of interest. The average treatment effect for the treated (ATT) is

$$E[\delta \mid D = 1] = E[Y^1 - Y^0 \mid D = 1], \tag{16}$$

and the average treatment effect for the controls (ATC) is

$$E[\delta \mid D = 0] = E[Y^1 - Y^0 \mid D = 0]. \tag{17}$$

For estimation of the effect of a two-valued cause, meaningful heterogeneity of individual-level causal effects exists when $E[\delta \mid D = 1]$ and $E[\delta \mid D = 0]$ differ from each other to a degree that is deemed substantial by an analyst.[17]

For the Catholic school example, the ATT is the average effect of Catholic schooling on achievement of those who attend Catholic schools. The ATC is the opposite: the average effect of Catholic schooling on achievement of those who attend public schools. Both average

[16]Expectation-based variants of ignorability assumptions are weaker but sufficient to identify the ATE (see Imbens 2004).

[17]Of course, meaningful heterogeneity may still exist even if $E[\delta \mid D = 1] = E[\delta \mid D = 0]$. However, the most important form of heterogeneity to consider is the one where the average effect of the cause differs for those exposed to the different levels of the cause. This type of heterogeneity is the focus of the diagnostic routine.

causal effects are different conceptually from the ATE in equation (11), which is the average effect of Catholic schooling across all students.

The ATT and ATC in equations (16) and (17) can be estimated consistently under weaker assumptions than the unconditional ATE in equation (11). Full ignorability, as specified in equation (15), need not obtain. Suppose instead that only a subset of the variables in S is observed, denoted by X. If partial ignorability holds with respect to X, such that

$$Y^0 \perp\!\!\!\perp D \mid X, \tag{18}$$

then conditioning on X generates a consistent estimate of the ATT. The basic idea is that, on average within strata defined by X, the values of y_i among those in the control group can be used to consistently estimate the counterfactual values of y_i^0 for those in the treatment group. The opposite also obtains. If partial ignorability holds in the other direction, such that

$$Y^1 \perp\!\!\!\perp D \mid X, \tag{19}$$

then the ATC can be estimated consistently because the values of y_i among those in the treatment group can be used to consistently estimate the counterfactual values of y_i^1 for those in the control group, on average within strata defined by X.

5.2. *Why Would the Diagnostic Routine Succeed in Identifying Consequential Causal Effect Heterogeneity?*

To provide a justification for the diagnostic routine, we pose and then answer three questions.

Question 1: *When would potential-outcome-defined, individual-level causal effects generate population-level patterns where the ATT differs from the ATC?* The ATT will differ from the ATC whenever one or more variables predict both treatment group membership D and variation in the individual-level treatment effect δ. If such a variable has a positive relationship with both D and variation in δ, then it is often referred to as "positive selection." The classic case here is ability bias

in the estimate of the causal effect of college education on subsequent earnings in the labor market. High ability individuals are thought to be more likely to enter college and more likely to learn marketable skills while in college. In cases of positive selection, the ATT is larger than the ATC.

The opposite pattern is common as well. For what is often referred to as "negative selection," the variable that generates the heterogeneity has a positive relationship with D but a negative relationship with variation in δ. It generates a larger ATC than ATT. For an example, consider the Catholic schooling effect again.[18] Based on past research, there are at least three possible explanations for why we might expect that the ATC would be larger than the ATT for the effect of Catholic schooling:

1. The *common school* explanation: Catholic schools distribute opportunities for learning, such as advanced course-taking, more equitably than do public schools. This explanation was stressed by Coleman and colleagues in their initial research and was then more comprehensively developed by Bryk, Lee, and Holland (1993). It suggests that variables such as parental education and nonminority status have positive relationships with D but negative relationships with the variation in δ.
2. The *better alternatives* explanation: Catholic schooling is particularly beneficial to those students who have poor public schooling alternatives, in particular those students from families who are not able to afford to live in school districts with the best public schools. This explanation was first fully developed by Neal (1997), and it suggests that variables such as family income and wealth have positive relationships with D but negative relationships with the variation in δ.
3. The *binding constraint* explanation: Differential responsiveness exists to accurate perceptions of students' likely benefits from Catholic schooling. For low-income families for whom tuition at a Catholic school represents a genuine financial sacrifice, the only students who enroll in Catholic schools are those students who are especially likely to benefit from enrolling. In contrast, among high-income families for whom tuition is not a substantial financial sacrifice,

[18] For an alternative example, which lines up nicely with the positive selection example just given, see Brand and Xie (2007).

even students who are not likely to benefit from attending Catholic schooling instead of public schooling may enroll in Catholic schools. This explanation is discussed in Morgan (2001), and it is based on the assumption that there is heterogeneity in δ that students and their parents can forecast. Because tuition is costly, and relatively more so for those from resource-poor families, it takes a larger value of δ to induce low-income students to enroll in Catholic schooling. As a result, variables that capture resource availability have positive relationships with D but negative relationships with the variation in δ among those who enter Catholic schooling.

Although these explanations may appear straightforward, they are not. In this case (which is not unusual in our reading of the applied literature), the true variables that generate the heterogeneity are unobserved, even though the explanation for the heterogeneity is constructed around a surface narrative that can be pegged to variables that are observed.

For the common school explanation, the negative relationship between parental education and variation in δ is produced by differences between Catholic and public schools in their instructional practices, which at least for Coleman had deeper sources in alternative ideological beliefs about the capacities of children. In this case, parental education serves as a proxy for the more fundamental unobserved variables that structure the variation in δ.

For the better alternatives explanation, family income and wealth are proxies for the concrete but unobserved public schooling alternatives that are available to relatively low-income and wealth-constrained Catholic school students and that structure the variation in δ. The complication here is that we need to know the characteristics of the choice sets of all students, not merely the characteristics of the public schools that Catholic school students would have attended according to their residential location. Had Catholic school students not enrolled in Catholic schools, their parents might instead have chosen to move to better public school districts.

Finally, for the binding constraint explanation, the heterogeneity in δ may not be exogenous, as it could be a function of mental ability or taste for a religious educational environment. Moreover, the binding constraint itself is not solely a function of potentially available family income but instead includes other behavioral components, such as the valuation of alternative uses of the same resources and variation across Catholic schools in financial aid programs.

Question 2: *Will the diagnostic routine reveal differences in the ATT and the ATC in a sample of sufficient size?* In answering this question, we first discuss three separate cases based on assumptions about the treatment selection/assignment process, depicting the expected results that we would obtain in each situation. Thereafter, we discuss the possibilities for false positive and negative diagnoses with reference to these three cases.

1. *Full ignorability.* Suppose that full ignorability in equation (15) holds because the variables utilized as X in the routine are equivalent to the variables defined as S in equation (13). In this case, weighted regression using $w_{i,\,\mathrm{ATT}}$ generates a consistent estimate of the ATT, and weighted regression using $w_{i,\,\mathrm{ATC}}$ generates a consistent estimate of the ATC. Thus, if the estimates are judged substantively different, then this result is direct evidence that consequential causal effect heterogeneity is present.
2. *Partial ignorability.* Suppose that full ignorability in equation (15) does not hold because the variables utilized as X in the routine are a subset of the variables defined as S for equation (13). But, suppose that X is sufficiently predictive of treatment selection/assignment that one of the two forms of partial ignorability in equations (18) and (19) holds. Weighted regression using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ will generate a consistent estimate of either the ATT or the ATC, depending on whether equation (18) or (19) holds. In this situation, the estimates yielded by $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ will almost always differ because the unobserved variables that generate the partial nonignorability will be related to the observables that determine treatment selection/assignment. As a result, if the estimates using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ differ, there is good reason to believe that there is a genuine difference between the true ATT and the true ATC. (But see Section 5.3 for a discussion of the prospect of false diagnoses generated by rare patterns of unobserved variables.)
3. *Nonignorability.* Suppose that neither full ignorability nor partial ignorability of any form holds. In this case, weighted regression results using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ that generate two substantively different average causal effect estimates suggest at least one of the two following conclusions, based on how far from full ignorability the conditioning variables in X render the estimation challenge:

a. Even though the variables in X are not rich enough to sustain assumptions of ignorability of any form, enough of the process of treatment selection/assignment has been modeled such that the substantive difference in the weighted regression results using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ is supportive of the conclusion that consequential causal effect heterogeneity is present. In other words, even though consistent estimates of the ATT and the ATC are unavailable, the models are sufficiently rich to conclude that it is sufficiently unlikely that the mismatch of the two estimates is entirely attributable to differential departures from assumptions of partial ignorability.
b. Treatment selection/assignment is so incompletely accounted for by the variables used in X in the original unweighted regression model that there is no basis for concluding anything about the nature or direction of a difference between the ATT and the ATC. As a result, assumptions of causal effect homogeneity (or completely random individual-level heterogeneity) have no empirical support, and the weaker assumption of consequential heterogeneity is more reasonable.

Question 3: *Are there cases in which the diagnostic routine will generate false positive or false negative diagnoses?* Generic sampling error and misleading past research could prompt researchers to mistakenly accept or reject that weighted regression estimates using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ differ. In this case, they would mistakenly conclude that consequential causal effect heterogeneity is present when in fact it is not (and vice versa). Similar possibilities are present in nearly all data analysis situations in which conclusions about features of a population are based on data from samples drawn from the population. We do not have reason to believe that the routine is unusually vulnerable to false diagnoses of this sort.

However, there are some additional considerations beyond these generic inferential issues when full ignorability of treatment selection/assignment cannot be maintained. In the cases of partial ignorability and complete nonignorability, the presupposition that unobserved variables are at play creates additional complications.

Consider first the case of partial ignorability, as sustained by the adoption of the assumption in either equation (18) or (19). It is theoretically possible that the unobserved variables that generate

nonignorability are completely independent of the observed variables that determine treatment selection/assignment and yet are related to the variation in individual-level causal effects. This would be the case if individual-level causal effects were completely random, and yet individuals entered into the treatment group in part based on accurate expectations of their own individual-level treatment effects. In this unlikely case, the estimates yielded by $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ could be nearly the same, even in an infinite sample and even though the true ATT and ATC could be very different. Again, this scenario is unlikely, since it would be rare for the following to be true: (1) the data are rich enough such that partial ignorability obtains, (2) the crucial unobserved variables that prevent full ignorability from being asserted are (a) completely independent of the variables in the conditioning set X and (b) related to the variation in individual-level causal effects.

In the case of complete nonignorability, the data are substantially less rich in the sense that the variables in the conditioning set X may be a small subset of S. In this case, it is more likely that the crucial unobserved variables that relate individual-level causal effects to treatment selection/assignment will have a weak relationship or no relationship to the variables in X.

Nonetheless, the fragility of the routine in these circumstances should not obscure the main point of the routine. If, for example, the regression estimates yielded when using $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ were very similar and yet researchers had reason to believe that treatment selection/assignment is nonignorable, it would be rather audacious to then decide to present their baseline regression estimates and declare that, although they are not consistent estimates of the ATE, ATT, or ATC, it is reasonable to proceed as if homogeneity of individual-level causal effects is present.

5.3. *The Diagnosis for the Demonstration Reconsidered with Reference to the Counterfactual Model*

For two reasons, it is clear that in the Catholic school demonstration full ignorability of treatment selection/assignment cannot be safely assumed. First, sufficient evidence exists to suggest that some Catholic school students attend Catholic schools because they expect to gain from doing so. Because the ELS data do not have measures of these

expectations, the demonstration represents a case in which full ignorability cannot be maintained. In essence, expectations of the causal effect are a variable in S in equation (15), and this variable is not available for inclusion in X.

Second, the results presented in the diagnostic routine indicate that 23 percent of public school students (i.e., 2739 out of 12,025) have estimated propensity scores lower than the lowest estimated propensity score of any Catholic school student. A comparison of the results in Tables 2 and 4 suggests that this lack of overlap is relatively inconsequential for the empirical results of the diagnostic routine, and yet this provides only modest reassurance that the data can inform us of the likely benefit that these 23 percent of public school students would obtain from attending a Catholic school.

At best, it would seem that only partial ignorability holds in the direction where equation (18) is valid for the variables in X available to us (which means that partial ignorability in equation [19] does not hold). As a result, the weighted regression models that utilize $w_{i,\,\mathrm{ATT}}$ may deliver a consistent estimate of the average causal effect of Catholic schooling among Catholic school students (i.e., the ATT), but the weighted regression models that utilize $w_{i,\,\mathrm{ATC}}$ do not deliver a consistent estimate of the average causal effect of Catholic schooling among public school students (i.e., the ATC).

If partial ignorability is present and the data deliver a consistent estimate of the ATT that is substantially different from the estimate of the ATC (even though the estimate of the ATC is biased and inconsistent), we can conclude based on the reasoning in the last section that consequential heterogeneity is present because there is a sufficiently strong likelihood that the true ATT and the true ATC differ meaningfully. In particular, because the unobserved variables that are part of the binding constraint explanation and the better alternatives explanation noted earlier are related to observable variables such as family income, the revealed difference between the estimates that use $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ reflects, to some substantial extent, an underlying pattern of heterogeneity that causes the true ATC to be larger than the true ATT.[19]

[19]In addition, the common support results reported in Table 4 show that the heterogeneity exists in the region of overlap, suggesting that individuals in the treatment group who are most similar to those in the control group have a different average causal effect than those in the treatment group who are least similar to those

If treatment selection/assignment is nonignorable—such that none of the assumptions in equations (15), (18), or (19) is valid—then neither the ATT nor the ATC can be consistently estimated with the data. Without any evidence that differential departures from partial ignorability could have generated the results, we would argue that it is most reasonable to accept the *prima facie* interpretation that differing estimates based on $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ represent substantial evidence that consequential causal effect heterogeneity is present. Again, we would argue that it is reasonable to conclude that the difference between the estimates reflects a true difference between the ATT and the ATC for the same reasons stated for the scenario in which partial ignorability is assumed.[20]

Thus, in this application of the routine, the case for consequential heterogeneity is quite strong. Accordingly, it is reasonable to conclude that the baseline regression results mask consequential heterogeneity.

5.4. *Applicability of the Routine in Other Situations*

The diagnostic routine presented in this paper can be modified for alternative research situations. In most cases, the logic of the routine carries over directly to these scenarios, but the practical implementation of the routine must be changed.

If the outcome variable is not interval scaled, or cannot be treated as such, then the basic steps of the routine can be enacted, substituting a different response model for the linear regression equations. The existing methodological literature provides little guidance on how exactly individual-level causal effects are weighted in these situations, but it seems clear that analogs to conditional-variance weighting are likely to prevail. Much depends on how causal effects are defined with reference to potential outcomes, and clearly the linear-difference definition in equation (10) will be the wrong starting point in most cases.

in the control group. Thus, the difference that is revealed by alternative estimates based on $w_{i,\,\mathrm{ATT}}$ and $w_{i,\,\mathrm{ATC}}$ is not produced by reckless comparison of perhaps fundamentally incomparable cases.

[20]And, even if this judgment were not acceptable to a fair critic, such a critic would presumably also agree that it would be foolhardy to rely only on the baseline regression results for all subsequent interpretation.

For many-valued causes, the challenges are mostly practical. If the number of states of the causal variable is modest, a series of two-way comparisons between the multiple treatment states can be undertaken. The diagnostic routine can be estimated for each two-way comparison, following the basic procedures that are recommended for matching estimators for many-valued causes (see Imbens 2000; Rosenbaum 2002). If the number of states of the causal variable is large, then the causal states can be collapsed in alternative ways for the sake of estimating the diagnostic routine. If a positive diagnosis arises, then selected categories can be subdivided progressively to determine whether or not the diagnosis is an artifact of the initial coarsening.

6. DISCUSSION

How should an analyst proceed if the diagnosis of consequential heterogeneity is positive? Many options are available within the counterfactual modeling framework. All of these options are contingent on the type of ignorability that holds. In some cases, a strategy to fully parameterize the heterogeneity is available, and thus a respecification of the regression model is possible. In other cases, especially when issues of overlap and support come to the foreground in the course of estimating the diagnostic routine, matching estimators can be deployed to enable a more subtle (but perhaps limited) examination of the heterogeneity. Thereafter, a regression model can be reverse-engineered from the matching results if the analyst desires. Finally, it may even be the case that an instrumental variable is lurking within the variables that predict treatment assignment, and if so the instrumental variable can be used to estimate a meaningful component of the average treatment effect for the treated. We discuss these options in this section.

6.1. *Regression Options*

If full ignorability obtains, then the two most obvious forms of consequential heterogeneity have already been modeled by the weighted regression results. It may be possible to offer better estimates of the ATT and the ATC by fitting additional models of treatment assignment, using more intensive data mining techniques (see McCaffrey et al. 2004; Diamond and Sekhon 2005). Thereafter, new weights can be constructed

that may yield slightly different point estimates. However, any improvements are likely to be small if, as in our demonstration, the balance achieved by the final weights constructed in Step 6 of the diagnostic routine is already very good.

In many research scenarios, having the best possible estimates of the ATT and the ATC is all that might be desired. However, if full ignorability can be asserted, then a deeper modeling of the underlying heterogeneity is possible, and such modeling can be carried out by respecifying the original regression model.

The first step for an analyst who seeks such deeper modeling is to consider whether a comprehensive substantive story about the heterogeneity of the causal effect can be effectively parameterized by observed individual-level variables. In the case of full ignorability, this should be possible. Consider, for example, the regression equation presented earlier as equation (2):

$$Y = a + b_d D + b_2 X_2 + \cdots + b_k X_k + e. \tag{20}$$

Suppose that we have theoretical reasons to believe that the full set of variables X_2 through X_k must be conditioned on to guarantee full ignorability, and yet we also have reason to believe that nonrandom individual-level heterogeneity in the causal effect is dependent on X_2 alone. This would be the case if X_2 predicts D as well as individual-level variation in the underlying causal effect δ whereas X_3 through X_k predict D but are independent (individually and in any combination) of individual-level variation in the causal effect δ. In this case, an unweighted regression model can be estimated that includes an interaction term between X_2 and D, as in

$$Y = a + b_d D + b_{d2}(D \times X_2) + b_2 X_2 + \cdots + b_k X_k + e. \tag{21}$$

In this case, the expanded regression model would explicitly account for the variability in the treatment effect across individuals.[21] Estimates of the ATT and ATC are then functions in b_d and b_{d2}, weighted differentially by the distribution of X_2.

[21]Or, if there is no theory that suggests where the heterogeneity lies, we can interact D with all of the variables in X, as in Brand and Halaby (2006) based on the model suggested by Wooldridge (2002).

How can we determine whether an expanded regression model of this sort completely accounts for all consequential heterogeneity across levels of the cause? If the data are extensive enough and the number of categories is small enough in the adjustment variables that are thought to account completely for the individual-level heterogeneity, then separate unweighted regression models can be fit within strata defined by these adjustment variables. For example, if the expanded regression model were the one presented in equation (21), then the regression model in equation (20) would be estimated separately for each value of X_2 (and X_2 would then drop out of each regression model because it would be constant within each of them, resulting in a separate model $Y = a + b_d D + b_3 X_3 + \cdots + b_k X_k + e$ for each value of X_2). The diagnostic routine can then be applied to the models for each of the strata in order to check whether evidence of consequential heterogeneity remains under the new specification. If consequential heterogeneity then reappears within one or more strata, then the expanded regression model that is currently being assessed must be expanded further, and so on. If, on the contrary, the diagnostic routine does not reveal any heterogeneity within any of the strata, then the expanded regression model can be offered as a comprehensive model that accounts for all systematic heterogeneity of the causal effect.[22]

If such deeper modeling of heterogeneity is not possible, perhaps because the individual-level variables that are available can be regarded as nothing other than imperfect proxies for the underlying variables that are truly driving the heterogeneity of the treatment effect, then the situation is different. In this case, there is good reason to doubt that an assumption of full ignorability is valid.

Under these circumstances, and in the more common situation in which theory alone suggests that full ignorability does not hold, analysts should first concede that not all average causal effects can be estimated effectively with the available data. Instead, the prospects of partial ignorability should be assessed.

[22]A complication of this strategy is the implementation of Step 8 in the diagnostic routine. Each time the sample is subdivided, the diagnostic routine loses power to detect consequential heterogeneity. This implies that, if we subdivide the sample enough, we can always reach a level where rigid null hypotheses of homogeneity cannot be rejected. Although we might regard this as a weakness of the diagnostic routine, in our judgment it is a reminder that rigid hypothesis testing can ruin sound scientific judgment.

If partial ignorability obtains, appropriate weighted regression models should be estimated for the identified average effect. Deeper modeling of the heterogeneity beneath the ATT or the ATC is possible via the expanded regression approach just outlined for the case of full ignorability. However, it can easily become conceptually difficult to determine how best to specify a set of interactions designed to pick up heterogeneity of individual-level causal effects within the treatment group or the control group, knowing that in the end we will need to estimate a weighted regression model anyway in order to align the data with the target parameter of interest. In these cases, shifting toward the matching methods that we discuss next may be more natural.

In cases where no form of ignorability is valid, the goals of the analysis should change. Descriptive regression results and analyses of the bounds of causal effects can be very useful for guiding future data collection and for undermining the unwarranted causal claims of other researchers. Providing defendable causal effect estimates will probably not be possible, but much useful analysis can still be offered. It is possible, for example, that progress can be advanced by the matching methods we discuss next, as these can help to describe the estimation challenges that arise from fundamental incompatibilities between some treatment and control cases.

6.2. *Matching Options*

Multivariate matching methods, such as propensity score matching, can also be used to directly model the heterogeneity of the average treatment effect (see Rubin 2006 for a comprehensive review of matching methods and Morgan and Winship [2007, chs. 4 and 5] for a review written for sociologists). Matching methods are a valuable alternative to respecification of the original regression model if concerns about support and overlap are substantial (as would be revealed in the last step of the diagnostic routine).

Although matching and regression methods are not as distinct as is sometimes claimed in the applied literature, matching methods are uniquely valuable in forcing the analyst to consider whether the data can sustain a causal analysis that seeks to identify either the ATT or the ATC. Often, estimation of neither is possible because the support of the adjustment variables differ for the treated and for the controls, as

is the case when the range of the propensity score differs substantially for the treatment and control groups. In such cases, conditional average treatment effects can be estimated for segments of the propensity score that lie within the region of overlap. There are many examples of this strategy, and in sociology Morgan (2001) uses this approach in an analysis of the Catholic school effect. He estimates the average causal effect for quintiles of the propensity score for the treated. In a different application, Xie and Wu (2005) estimate average treatment effects across strata of the entire range of the propensity score.

6.3. *Instrumental Variable Options*

As Heckman and Vytlacil (2005) show, instrumental variables permit direct and comprehensive modeling of the heterogeneity of average treatment effects in idealistic data availability scenarios. They demonstrate how local instrumental variables that qualify for inclusion in an index structure model of treatment selection can be used to estimate marginal treatment effects. These effects can then be used to calculate a wide range of average treatment effects that are functions of the propensity score (see Morgan and Winship [2007, sec. 7.5.1] for an explanation of the argument).

The diagnostic routine proposed in this article would lead us to estimate marginal treatment effects via local instrumental variables if such instrumental variables were available.[23] We know of no application in which local instrumental variables have been available to estimate all marginal treatment effects of interest. In contrast, when convincing instrumental variables have been utilized (see Angrist and Krueger [2001] for a list of such cases and Rosenzweig and Wolpin [2000] for a critical review of these cases), the instrumental variables usually identify very narrow average treatment effects for small subsets of the population. And, in most applications, instrumental variables are simply not available. Thus, although it would be comforting to believe that it is practical

[23] In our setup, local IVs would be available if it was suddenly discovered that it could be assumed that one of the variables in X has no effect on the outcome variable except indirectly through D, has a strictly monotonic effect on D, and predicts the full observed range of the probability of treatment assignment. These conditions would typically arise only when there is a single variable in X that comprehensively accounts for the full pattern of treatment exposure and yet has no effect on the outcome except via D.

to model heterogeneity directly for most applications using instrumental variables, the history of applied research in the social sciences is not encouraging. It seems most reasonable to conclude that instrumental variables will at times provide important insight into causal effect controversies by estimating the average causal effects for potentially important segments of the population. For practical data availability reasons, they cannot be expected to serve as a general solution.

APPENDIX: DETAILS OF THE DEMONSTRATION

Poststratification and Attrition-Adjustment Weights

The complex sampling design of the ELS, where students are sampled from schools that are themselves sampled unequally from alternative population strata, necessitates the use of a poststratification weight developed by the data distributors. Moreover, for models in which twelfth grade achievement is used for the outcome variable, we limit the analysis sample to respondents who did not transfer between schools and who were enrolled in the twelfth grade at the time of the 2004 survey. This restriction results in a twelfth grade analysis sample of 1660 Catholic school respondents and 8842 public school respondents (for a total $N = 10{,}502$). Because these 10,502 respondents are a nonrandom subset of the 13,943 respondents in our base-year analysis sample, the twelfth grade results incorporate a model-based adjustment for attrition patterns.

We chose to adjust the data by estimating a logit model, from which we then extracted the probability of being in the twelfth grade and in the same school in 2004 as when initially sampled in 2002.[24]

[24]We estimated a multinomial logit model for the full sample where there were nine destinations: (1) in school, in grade, nontransfer; (2) in school, in grade, transfer, same sector; (3) in school, in grade, transfer, different sector; (4) in school, out of grade, nontransfer; (5) in school, out of grade, transfer; (6) homeschooled, out of scope; (7) early graduate; (8) dropout; (9) nonrespondent/status unknown. The predictor variables included dummies for gender, race, urbanicity, region, and family structure as well as variables for parents' education, occupational prestige, and family income. The model yielded a chi-squared test statistic of 1388.5 with 152 degrees of freedom, which indicates that these predictor variables account for a substantial portion of the variation in trajectories. However, we then used only the probability of being in category (1) to define the adjustment weight summarized

We then formed a 2004 direct-adjustment weight that is the poststratification weight from the base-year 2002 data multiplied by the inverse probability of being on-track in the twelfth and at the same school in 2004 as in 2002. Our twelfth grade models give disproportionately more weight to individuals who were least likely to remain in our analysis sample between 2002 and 2004. Thus, conditional on the suitability of the underlying logit estimation of the probability of remaining in the analysis sample, our twelfth grade results are interpretable as generalizable estimates of what the patterns would have been in the twelfth grade if all sophomores had stayed in the same school and progressed to the twelfth grade between 2002 and 2004 (and all else remained the same).

Estimation of Weighted OLS Models

Consider how these weights were utilized when estimating the regression models. First, note that the OLS estimate $\hat{\delta}_{\text{OLS, bivariate}}$ in equation (3) is the second element of the vector $\hat{\mathbf{b}}_{\text{OLS}}$ from

$$\hat{\mathbf{b}}_{\text{OLS}} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}, \tag{22}$$

where (1) $\mathbf{Q}$ is an $n \times 2$ matrix that contains a vector of 1s in its first column and a vector of the values of d_i for each individual in its second column and (2) $\mathbf{y}$ is an $n \times 1$ column vector containing values of y_i for each individual.

To generate the OLS estimate $\hat{\delta}_{\text{OLS, multiple}}$ in equation (4), equation (22) still obtains, but the matrix $\mathbf{Q}$ is augmented so that it is $n \times k + 2$, where the first column is a vector of 1s, the second column is a vector of the values of d_i for each individual, and the remaining k columns are vectors of the values of x_i for each of the k variables in X in equation (4).

To estimate weighted OLS regression using weights such as $w_{i,\text{ ATT}}$ and $w_{i,\text{ ATC}}$ from equations (6) and (7), the estimator in equation (22) is augmented to form a weighted ordinary least squares estimator:

in the main text. Thus, patterns of movement between the other categories were examined only to make sure that the predictor variables were sorting the sample in expected ways (i.e., to verify that parental education is more strongly predictive of retention and dropout than of homeschooling, and so on).

$$\hat{\mathbf{b}}_{\text{OLS, weighted}} = (\mathbf{Q}'\mathbf{P}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{P}\mathbf{y}, \tag{23}$$

where $\mathbf{P}$ is an $n \times n$ diagonal matrix with either $w_{i,\,\text{ATT}}$ or $w_{i,\,\text{ATC}}$ along the diagonal.

For the Catholic school effect demonstration, the complex sample design required that the weighted OLS estimator in equation (23) be used throughout. For the models in the first column of Tables 2, 4, and 6, the tenth grade models were estimated with a weight matrix $\mathbf{P}$, where the diagonal is the poststratification weight. For the twelfth grade models and the math gain models, the diagonal of the weight matrix $\mathbf{P}$ is the poststratification weight $\times$ attrition-adjustment weight.

For the estimation of the average treatment effects for the treated and the controls in the second and third columns of Tables 2, 4, and 6, the tenth grade models were estimated with a weight matrix $\mathbf{P}$, where the diagonal is the poststratification weight $\times$ $w_{i,\,\text{ATT}}$ or $w_{i,\,\text{ATC}}$. And, for the twelfth grade models and the math gain models, the diagonal of the weight matrix $\mathbf{P}$ is the poststratification weight $\times$ attrition-adjustment weight $\times$ $w_{i,\,\text{ATT}}$ or $w_{i,\,\text{ATC}}$.

Estimation of Standard Errors

For the standard errors of all regression estimates, we calculated heteroskedastic-consistent standard errors with a supplemental adjustment for the clustering of students within schools (using Stata's *robust* option in its regression routine along with the tenth grade school ID as a cluster identifier). The estimated standard errors are therefore comparable across columns.

We make no adjustment to account for the fact that the weights $w_{i,\,\text{ATT}}$ and $w_{i,\,\text{ATC}}$ are based on an estimated quantity, the propensity score. In the counterfactual causality literature, this complication is often a topic of concern, as it is argued that the estimation error in the propensity score should propagate to the standard error of the estimated causal effect. Although we recognize that we have not accounted for this source of uncertainty in our causal effect estimates, we find comfort in two sources: (1) even though the poststratification weight offered to us by the data distributors is estimated, such weights are routinely treated as known by nearly all researchers who use them, and (2) the sandwich variance estimator that we utilize should be able to implicitly adjust

for any differential clustering on the variables in X that is induced by weighting based on $w_{i,\,\mathrm{ATT}}$ or $w_{i,\,\mathrm{ATC}}$. Nonetheless, work remains to be done within the statistics literature on how to estimate the variances of these types of estimators.

REFERENCES

Alexander, Karl L., and Aaron M. Pallas. 1983. "Private Schools and Public Policy: New Evidence on Cognitive Achievement in Public and Private Schools." *Sociology of Education* 56:170–82.

———. 1985. "School Sector and Cognitive Performance: When Is a Little a Little?" *Sociology of Education* 58:115–28.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66:249–88.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Pp. 1277–366 in *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card. Amsterdam: Elsevier.

———. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15:65–83.

Brand, Jennie E., and Charles N. Halaby. 2006. "Regression and Matching Estimates of the Effect of Elite College Attendance on Educational and Career Achievement." *Social Science Research* 35:749–70.

Brand, Jennie E., and Yu Xie. 2007. "Who Benefits Most From College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." On-Line Working Paper 033-07, California Center for Population Research, Los Angeles.

Bryk, Anthony S., Valerie E. Lee, and Peter B. Holland. 1993. *Catholic Schools and the Common Good*. Cambridge, MA: Harvard University Press.

Coleman, James S., et al. 1966. *Equality of Educational Opportunity*. Washington, DC: US Government Printing Office.

Coleman, James S., and Thomas Hoffer. 1987. *Public and Private Schools: The Impact of Communities*. New York: Basic Books.

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

Diamond, Alexis, and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper, Travers Department of Political Science, University of California, Berkeley.

DuMouchel, William H., and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78:535–43.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. Cambridge, England: Cambridge University Press.

Goldberger, Arthur S., and Glen G. Cain. 1982. "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report." *Sociology of Education* 55:103–22.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99:609–18.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115:45–97.

Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156–245 in *Longitudinal Analysis of Labor Market Data*, edited by J. J. Heckman and B. Singer. Cambridge, England: Cambridge University Press.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." Working Paper, Department of Economics, University of Chicago.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73:669–738.

Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. 1985. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 58:74–97.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87:706–10.

———. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86:4–29.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

———. 1987. *Statistical Design for Research*. New York: Wiley.

Little, Roderick J. A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77:237–50.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9:403–25.

Morgan, Stephen L. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning." *Sociology of Education* 74:341–74.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, England: Cambridge University Press.

Neal, Derek. 1997. "The Effects of Catholic Secondary Schooling on Educational Achievement." *Journal of Labor Economics* 14:98–123.

Noell, Jay. 1982. "Public and Catholic Schools: A Reanalysis of 'Public and Private Schools.'" *Sociology of Education* 55:123–32.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Robins, James M., and Ya'acov Ritov. 1997. "Toward a Curse of Dimensionality Appropriate (Coda) Asymptotic Theory for Semi-Parametric Models." *Statistics in Medicine* 16:285–319.

Rosenbaum, Paul R. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82:387–94.

———. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rosenzweig, Mark R., and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38:827–74.

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159–83.

———. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:322–31.

———. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, Donald B., and Neal Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573–85.

Sobel, Michael E. 1996. "An Introduction to Causal Inference." *Sociological Methods and Research* 24:353–79.

———. 2000. "Causal Inference in the Social Sciences." *Journal of the American Statistical Association* 95:647–51.

Thompson, Steven K. 2002. *Sampling*. New York: Wiley.

van der Laan, M. J., and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Willms, J. Douglas. 1985. "Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-up Study." *Sociology of Education* 58:98–114.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Xie, Yu, and Xiaogang Wu. 2005. "Market Premium, Social Process, and Statisticism: Reply to Jann." *American Sociological Review* 70:865–70.