Case Selection in Public Management Research: Problems and Solutions

***David M. Konisky*[*], *Christopher Reenock*[†]**
[*]Georgetown University; [†]Florida State University

ABSTRACT

Case selection is ubiquitous in public management research. Rarely do scholars have access to entire populations of interest. Yet, the manner by which scholars select samples to conduct their analyses can have profound consequences on their ability both to draw valid causal inferences and to estimate accurate relationships. In this article, we review the basic threats to inference that are likely to emerge in the presence of non-random case selection, with specific attention to their manifestation in empirical public management research. The article first reviews the threats to causal inference presented by case selection, focusing on their implications for internal and external validity. We then summarize a standard set of solutions to address potential problems for empirical models caused by non-random case selection. As part of this discussion, we review recent articles published in this journal to illustrate the prevalence of selection issues in contemporary public management studies, and then illustrate several techniques that have been developed to overcome specific problems to show their utility for public management research.

Causal inference is a central goal of social science. Scholars often conduct their research with the primary interest of understanding whether and to what extent a variable of interest influences some outcome. Given this central role, the pitfalls of drawing valid causal inferences have garnered much attention. Scholars highlight two standards with which a causal inference can be evaluated: internal and external validity. The former reflects the analyst's ability to draw valid inferences about the relevant causal relationship under study, whereas the latter reflects the extent to which valid inferences (based on a given sample of cases) may extend to the population or to cases not part of the original analysis. Although a variety of research design features can plague both internal and external validity, this article's focus is on the role of case selection, by which we mean either the explicit or implicit inclusion of a subset of cases (e.g., people, organizations, jurisdictions) from a larger population, in a study seeking to make causal claims.

We would like to thank Ryan Welch for his research assistance in processing and coding *JPART* articles for selection issues, and Stéphane Lavertu and David Weimer for sharing their data. We would also like to thank Ken Meier and Andy Whitford for their valuable suggestions, as well as the comments from the three anonymous reviewers. Address correspondence to the author at dmk74@georgetown.edu.

doi:10.1093/jopart/mus051
Advance Access publication November 15, 2012

Whether in the study of personnel management, public budgeting, or policy analysis, public administration scholars generally do not possess the luxury of having access to populations of interest. Finite resources as well as practical limitations constrain most scholars to conducting analyses on samples of populations—samples that may include non-randomly included cases. When using such data, potential threats to both internal and external validity arise. Our purpose here is neither to present a comprehensive overview of research design nor of the case selection literature. Rather, our goal is to review the main threats to valid causal inference presented by non-random case selection, with particular emphasis on empirical research in public management (broadly defined) that is quantitative in orientation. We deem this choice appropriate given the trends in public management scholarship in this and other leading journals.[1] We orient our theoretical discussion in the counterfactual causal inference model and apply most of our empirical treatments within the regression framework, but it is important to note that, at a conceptual level, similar issues arise in qualitative research as well (Collier and Mahoney 1996; King, Keohane, and Verba 1994).

The article proceeds as follows. In the first two sections, we review the main threats to causal inference presented by case selection, focusing on their implications for internal and external validity. We also summarize a standard set of solutions to address potential problems caused by non-random case selection. We subsequently report the results of a review of recently published *JPART* articles to examine the prevalence of case selection issues. We then turn to detailed illustrations of several particular methods to address non-random case selection. Although these solutions are not novel, they tend to be underutilized in public management research, and one of the objectives of this article is to make the methods more accessible to scholars by illustrating their utility. We conclude the article with a few general suggestions.

CASE SELECTION, CAUSAL INFERENCE, AND THREATS TO INTERNAL VALIDITY

There is perhaps no better place to begin an examination of the conditions under which case selection may threaten the internal validity of a study than the *counterfactual causal inference model* of the classic experiment. In the *ideal* experimental setting, the analyst has a theoretically interesting independent variable, D, whose presence is believed to causally alter the value of some outcome of interest, Y. To test this possibility, the analyst identifies subjects who are randomly assigned to one of two groups, treatment $D(1)$ or control $D(0)$. The treatment is then applied to subjects in $D(1)$ and not to those in $D(0)$, and each subject's response on the outcome variable of interest, Y, is recorded.[2] At the individual level, subjects in the treatment group realize a value of d_1 and subjects in the control group realize a value of d_0. The observable outcome variable, Y, is a function of two potential outcome variables, Y^1 and Y^0, such that $Y = DY^1 + (1-D)Y^0$. This means that each subject in the treatment group has an

[1] More specifically, we limit our focus to empirical work with the explicit purpose of theory or hypothesis testing and do not consider case selection issues in the context of theory building.

[2] To be clear, the "treatment" in most social science analysis and, in this case, public management research is nothing more than the main independent variable under investigation. This "treatment" can be dichotomous, ordinal, or continuous in measurement.

Table 1
The Fundamental Problem of Causal Inference

		Y^1 – Treatment outcomes		Y^0 – Control outcomes	
Treatment group		$E[Y^1 \mid D=1]$ **Observable outcome**	ATT ⇔	$E[Y^0 \mid D=1]$ Potential outcome	
	A ⇕		ATE ⤡		⇕ B
Control group		$E[Y^1 \mid D=0]$ Potential outcome	ATC ⇔	$E[Y^0 \mid D=0]$ **Observable outcome**	

observable outcome in the treatment, Y_i^1, and an unobservable counterfactual in the control, Y_i^0. The same is assumed for each subject in the control group; each subject has an observable outcome in the control and an unobservable counterfactual in the treatment.

At the individual level then, the causal effect of the treatment would be $\delta_i = y_i^1 - y_i^0$. Unfortunately, we cannot observe the outcome at the individual level for the counterfactual case—a feature referred to as the fundamental problem of causal inference (Holland 1986; King, Keohane, and Verba 1994; Morgan and Winship 2007). This problem is akin to one of missing data (Winship and Morgan 1999), since each subject can only be assigned to either the treatment $d_i = 1$ or the control $d_i = 0$. Accordingly, for a given subject, we can only ever observe *either* y_i^1 *or* y_i^0, but never both. For this reason, analysts must focus on aggregate-level causal effects. The key insight of this causal inference model is that the counterfactual outcome, which is essential to estimates of causal inference, cannot be observed directly; accordingly, various research methods must be used to approximate it. (See Hidalgo and Sekhon (2011) for a discussion of causality in this counterfactual framework.)

Table 1 displays the aggregate quantities of interest for each possible distribution of the potential outcome variables. The table shows each of the two potential outcome variables Y^1 and Y^0 across the columns and the treatment and control grouping by the rows. For those assigned to the treatment group, we have two outcomes: the observed outcome of treatment and the unobserved outcome for those in treatment had they been assigned to the control group. For those assigned to the control group, we have two outcomes: the observed outcome of control and the unobserved outcome for those in control had they been assigned to treatment. The most relevant quantity of interest is the average treatment (causal) effect (ATE), which represents the difference between the average outcome for the treatment group in the sample and the average outcome for the control group in the sample.[3] More formally, the estimated ATE is $\widehat{D} = E[\delta] = E[Y^1 \mid D=1] - E[Y^0 \mid D=0]^{\,1}$.

In addition to the ATE, there are two conditional treatment effects that are often of interest, both of which are displayed in Table 1. The average treatment effect of the

[3] Our discussion here is limited to identifying and measuring causal effects. Social scientists are also often interested in the underlying causal mechanisms. One approach to uncovering causal mechanisms is through causal mediation analysis, and recent work has developed techniques for conducting this type of analysis in both experimental and observational settings (Imai et al. 2010, 2011).

treated (ATT), $E[\delta | D=1] = E[Y^1 \mid D=1] - E[Y^0 \mid D=1]$, as well as the average treatment effect for the untreated or control (ATC), $E[\delta | D=0] = E[Y^1 \mid D=0] - E[Y^0 \mid D=0]$. The ATT reveals the treatment effect on those subjects who are likely to take the treatment. This quantity is often of particular value to social scientists. For example, consider the case of an analyst studying the impact of a specific program (e.g., the Supplemental Nutrition Assistance Program or SNAP) on some behavior or outcome (e.g., nutrition) among those who are likely to participate in the program (e.g., the poor). For such a question, the ATE would reveal the average change in nutrition that participating in SNAP induces for a randomly selected citizen from the population. But a randomly selected individual from the population is not likely to participate in the program. As a result, the ATE may be less interesting than the ATT, given that some proportion of the population (e.g., those with high income) may never participate. In this case, the analyst would be interested in the specific causal effect of the program among those who are likely to be treated—the poor. The ATC reveals the average effect for those who typically do not take the treatment. The ATC, for example, might be relevant for an analyst interested in the effect of a program on a new population of potential clients.

We can now discuss the key conditions for drawing valid causal inferences within the counterfactual framework and how these relate to case selection. The first condition derives from the fact that the estimate of the average causal effect, $\widehat{D}$, is produced from samples of individual cases assigned to either treatment or control.[4] Whether $\widehat{D}$ offers an unbiased estimate of D is a function of the randomness of the assignment process to the groups (Winship and Morgan 1999). If the treatment assignment process is random, we can generally assume that the treatment status of a case is jointly independent of the two random variables Y^1 and Y^0 associated with the outcome, Y. Deviating from independence risks biasing the estimate of the average causal effect ($\widehat{D}$). When joint independence holds, the assignment mechanism is said to be "ignorable," meaning that analysts can ignore the possibility that the treatment mechanism is interfering with their ability to draw a valid causal inference between the main independent variable and the dependent variable.

Independence Condition 1: The treatment assignment (selection) must be jointly independent of the potential outcome variables, Y^1 and Y^0.

To see how deviations from independence may bias estimates of $\widehat{D}$, let us consider Table 1 again. Sources of potential bias derive from the potential differences represented by arrows A and B. Arrow A refers to the difference in the treatment effect for those subjects in the treatment and control groups (or $Y_{i\in T}^{t} - Y_{i\in C}^{t}$, conditional on assuming no bias in group assignment). When this quantity is non-zero, it suggests that the treatment effect varies over the population, or that the treatment is having a different effect among those who are likely to take the treatment compared to those who are less likely to take the treatment. This bias is referred to as *differential*

[4] Although we use the concept treatment "assignment" in this section, the use of observational data necessitates a change to treatment "selection." This is because with observational data the analyst cannot typically assign cases to treatment, rather they are taken as given in the data.

treatment effect bias.[5] Arrow *B* refers to the difference between the treatment and control groups in the absence of treatment (or $Y_{i\in T}^{c} - Y_{i\in C}^{c}$). When this quantity is non-zero, it suggests a difference in the baseline outcomes between the two groups in the absence of treatment. This difference is often referred to as the *baseline bias* due to observable (and unobservable) characteristics between the two groups.[6]

Condition 1 asserts that treatment assignment is jointly independent of the potential outcomes. If this assumption holds, then the differences represented by arrows *A* and *B* both reduce to zero. When independence is violated, at least one of these quantities will be non-zero and the estimate of the true average treatment effect will be biased (Winship and Morgan 1999, 46–50). It is important to note that in the absence of *both* types of bias, the ATE = ATT = ATC.

Random assignment is the analyst's "protection" against systematic differences in observable (and unobservable) characteristics that can introduce bias into the estimate of the true average causal effect. With random assignment and sufficiently large groups, there is a low probability that the treatment and control groups will be systematically unbalanced on observable (or unobservable) characteristics that may lead to differences in outcome on either arrow *A* or *B*.[7] In the absence of random assignment or large groups, systematic differences between the groups may emerge. What then? If we assume that all of the variables that systematically affect treatment assignment (or selection) are represented by *S*, then we can extend Condition 1. Now we can specify the conditional probability that a case with specific characteristics will be observed in the treatment group, a quantity referred to as a *propensity score*. Across all characteristics *s* in *S*, treatment selection can be characterized by the general conditional probability, $\Pr[D=1\,|\,S]$. This information is useful because if we are aware of and have measures for all of the characteristics in *S*, then we can be confident that whatever variation remains in *D* is random and therefore ignorable (Morgan and Winship 2007, 75). Ignorability holds if, *conditioned on S*, the treatment status or assignment process is jointly independent of the two random variables Y^t and Y^c, or that $(Y^t, Y^c) \perp D\,|\,S$.

Independence Condition 2: The treatment assignment (selection) must be jointly independent of the potential outcome variables, Y^t and Y^c, conditioned on all variables (both observable and unobservable) that systematically determine treatment.

[5] Morgan and Winship (1999) show that the extent of bias registered in the estimate of the average treatment effect due to differential treatment effect bias is a function of the product of 1) the treatment effect between those in treatment versus those in control (ATT−ATC) and 2) the proportion of the population that does not select into treatment (1−π), or [(1−π)(ATT−ATC)]. As a result, the estimated average treatment effect will experience no differential treatment bias if either the difference between the ATT and the ATC is zero or if the proportion of the population that selects into or is assigned to treatment is π = 1. Analysts should take heed to this point, being duly concerned about such bias when the proportion of the population that typically takes the treatment under investigation is relatively small in the population.

[6] It is important to note that the "bias" referred to in this section applies to the counterfactual causal inference model and not to "selection bias" as evidenced in applied statistical models. Although the two are related, they have quite different operational definitions.

[7] Of course, random case selection does not guarantee the absence of selection effects, nor does non-random case selection guarantee the presence of them. Critically important is whether case selection induces dependence between treatment and potential outcome.

Condition 2 suggests that analysts who suspect systematic differences across their treatment and control groups can still extract a valid average treatment effect as long as they condition their analysis on all of these differences.[8] When we are able to identify and condition all variables represented in S, we have the situation referred to as "selection on the observables." In this case, selection is on these observed factors only and treatment is ignorable, which means that the analyst can obtain the true average treatment effect, so long as the outcomes are conditioned on *all variables* in S. Practically, analysts often do not know or cannot observe all of the variables in S. Rather they may only be aware or able to observe a subset of those variables, s. When this situation occurs, the analyst faces "selection on the unobservables," and assignment to treatment is said to be "non-ignorable." This latter situation requires greater effort to guarantee that the true average causal effect is not biased (i.e., conditioning on s is insufficient).

An important implication of the above conditions is that the outcome variable, Y, ought to possess the ex ante possibility of variation over outcomes. Causality is difficult to examine if the analyst *chooses* to not allow the outcome, Y, to take on different values.[9] Such artificial case restrictions are likely to induce dependence between treatment assignment and the potential outcome variables, Y^1 and Y^0. In addition, there are numerous situations where an analyst only partially observes the true value of an outcome, which also can affect the quality of causal claims. In the social sciences, violations of these conditions are ubiquitous. Analysts often incorrectly assume either the absence of systematic differences between their treatment and control cases, or that any such differences are unrelated to the distribution of the potential outcome variables. Occasionally these conditions are unmet in experimental settings, but this is nearly always the case with observational data.

Prior to discussing the specific ways in which case selection can threaten internal validity, it is useful to consider the empirical implications of the theoretical conditions covered above. Most analysts employ regression analysis to estimate the magnitude of a causal effect between a treatment variable and an outcome. There are of course several well-known Gauss–Markov assumptions (too many to cover here) that must be met in order for a regression model to provide the best linear unbiased parameters. Violating one or more of these assumptions can threaten an analyst's ability to draw valid inferences from a model (the parameters may be biased, inconsistent, or both). In the regression context, the empirical analog to the independence conditions discussed above is that the regressors cannot be correlated with the error term (a derivative of the Gauss–Markov strict exogeneity assumption). In a practical sense, this assumption is most often violated when the analyst omits a relevant variable, thereby inducing a correlation between the regressors and the error term. Simply put, to avoid violations of the independence conditions outlined above, the analyst would

[8] The average treatment effect is a theoretical concept, whose empirical analog may be thought of as the relevant marginal effect (associated with the main independent variable) from a regression analysis. However, it is useful to note that the latter is an empirical concept, related to the development and assumptions associated with the standard regression model.

[9] An obvious corollary to Condition 1 is that the key causal independent variable ought to vary as well over treatment and control, without which causal inference is hampered.

do well to avoid the difficulties that omitted variables present to the standard regression model. Given this relationship, as we discuss the ways in which case selection can threaten internal validity, we will also highlight the empirical analog along with its implications and solutions.

Case Selection and Threats to Internal Validity: Selection on the Treatment Variable

A critical threat to internal validity results from selection on the treatment variable. Unlike a controlled experiment, where treatment can be randomly assigned, with observational or field data, treatment is either given or a function of the analyst's case selection. In such situations, the analyst cannot safely assume that correlation between the treatment variable and a dependent variable of interest derives solely from the effect of the treatment. This problem is nearly always evident in non-experimental settings because data are unlikely to be randomly balanced across treatment and control groups. Given that most research in public management must rely on observational or field data, this is no small or rare problem.[10]

Take the example of survey research, which public management scholars increasingly use to collect information on public agencies and administrators' behavior and preferences. Much has been written on how various errors in survey design and implementation—often collectively referred to as "total survey error"—can affect a survey's representativeness (see Groves et al. 2004 for a canonical treatment of total survey error). Of particular concern in the context of selection on the treatment variable is non-response error, which occurs when an analyst does not obtain a response from a subject in a sample. Non-response error may lead to biased survey statistics, if the values or responses from the subjects that participate in the survey are systematically different than those that choose not to participate. When the pattern of non-response is correlated with both the "treatment" in the sample and the outcome of interest, and it is unaccounted for in analysis, the resulting bias can significantly affect the quality of inferences.

With most observational data, there may be one or several factors that are correlated with both case selection into treatment and the outcome of interest. As stated earlier, this violates the condition that the selection mechanism is independent from the unobserved outcome variables. This selection effect can introduce bias into the inferences drawn and exists in two forms.

The first type of selection on treatment can be on observable factors, which occurs when known factors are unbalanced between the treatment and control groups, and these factors are also correlated with the outcome variable. For example, consider a survey of program recipients who either do or do not access online interfaces with some government program (e.g., Medicaid, SSI). The analyst wants to know whether online interfacing affects program satisfaction. Without resources to develop a new survey, the analyst selects an instrument from a pre-existing survey of program

[10] Public management scholars rarely have the opportunity to conduct randomized experiments, and selection into treatment is often a critical part of the process or programs under examination.

recipients and uses a variable in the survey that assessed online usage. This variable is coded dichotomously, such that a value of one represents program recipients that elected to use the online interface and zero for those that did not. The analyst then uses this measure as the treatment variable in a regression on the dependent variable, program satisfaction, along with some controls. A potential problem for the analyst is that the survey may not have been administered in a way to ensure non-random assignment of the treatment.

Indeed, in this case, there was no "assignment" at all; participants self-selected into either treatment (have used online) or control (have not used online) group. If there was non-random selection into the treatment and these factors are also related to program satisfaction (which they easily could be), the analyst will have a problem drawing valid inferences. For example, if the selection into treatment was determined by factors such as gender, race, and education, then upon observing a correlation between the treatment and the outcome, the analyst cannot know whether that correlation represents the "true" causal effect or whether it is biased due to the uncontrolled factors. Any observed correlation may be due to unbalanced factors between the groups. More educated respondents may have been both more satisfied with the program and more likely to own a computer, pay for online access, and know how to use online interfaces. In the absence of correcting for the differences between the two groups, the analyst risks drawing incorrect causal inferences.

Selection on Treatment with Observables: Standard Solutions

If the factors contributing to differences between the treatment and control groups are also expected to explain the outcome of interest, and they are all known and can be measured, then there is a relatively simple solution to this problem—the analyst can condition the sample on variables measuring these other factors. Conditioning on observables imposes independence between the potential outcome variables and the treatment within strata defined by these observable characteristics. There are two standard approaches to impose conditional independence: the regression adjustment approach and the matching approach. Each approach is relatively easy to implement, assuming that the analyst knows and can measure the omitted factors that need to be included.

The regression adjustment approach considers non-random selection into treatment as an omitted variables problem in the regression context (Achen 1986; Heckman 1979). As noted above, a standard assumption of the OLS regression model is that the error term is uncorrelated with the regressors. This assumption is violated when a variable, which is correlated with both another independent variable and the outcome variable, is excluded from the regression. This is akin to the selection into treatment where the analyst excludes either an observable or unobservable variable that is correlated with both the treatment and outcome. To avoid violating this assumption, the analyst must include all variables that are correlated with both the treatment and outcome and are believed to account for non-random differences between the treatment and control groups.[11] In the example from above, if gender, race, and education are each related to both the probability that a program participant used an online interface and their satisfaction with it, then variables measuring these factors should be included in a regression equation. This approach is standard practice in regression analysis and seeks to avoid the problems (biased and inconsistent parameters) of omitted variable bias, where regressors are correlated with the error term.

A second approach is to condition the average treatment effect on observable, pretreatment attributes through a process of matching (Rosenbaum and Rubin 1983). In general, matching techniques pair cases in the treatment and control groups based on their similarity in observable characteristics. Assuming that the differences are fully captured by these characteristics and that they are observed prior to the treatment, the process of matching enables the unbiased estimation of the causal effect, assuming that the outcomes are independent of the assignment to the treatment group, conditional on these characteristics. In essence, the purpose of matching is to create a set of untreated cases that resembles the treated cases in all ways, except for having received the treatment. To illustrate this approach, we provide a detailed example of matching later in the article.

Selection on Treatment with Unobservables: Standard Solutions

An analyst must turn to a different suite of solutions in situations where the factors contributing to differences between the treatment and control groups are also expected to explain the outcome of interest and they are either unobservable or cannot be measured. This is perhaps the prototypical problem referred to as "selection bias" in the literature. Although the methods differ, each option seeks to eliminate the correlation between unobserved factors that likely explain both treatment and outcome of interest.

One approach is to use an instrument for the treatment variable. If unobserved characteristics are correlated with both the treatment and the error term, proceeding with the analysis would not only violate the independence conditions outlined above but would also violate the standard regression model assumption that regressors are uncorrelated with the error term. Using an instrument for the treatment variable addresses this problem. Consider the two properties of a proper instrument: 1) it must be correlated with the treatment variable; and 2) it must be uncorrelated with the error term of the outcome equation. If these properties hold, an analyst can use the instrument in place of the treatment variable in a standard OLS context. Of course, the difficulty with the instrumental variable approach is finding a strong instrument that meets the above conditions, and there is some evidence that using a

[11] In the case of survey research, if there are observable differences between those receiving the treatment in the sample and those not responding to the survey, the analyst can weight on known subject attributes to adjust for the non-response error. In the case where population weights are not known, the analyst can nevertheless extract unbiased estimates, using sample statistics. For an explanation of this technique, see Heckman and Todd (2009).

weak instrument may be more harmful to one's analysis than to proceed without the instrument (Gawande and Li 2009; Murray 2006). Recent examples of scholars using an instrumental variables approach in public management literature include the Gallo and Lewis (2012) study of the effects of political patronage on the performance of federal agencies and James's (2009) study on citizen satisfaction with public services provided by local governments.

If an analyst has panel data, several other techniques can be employed to address potential bias caused by unobserved characteristics between cases in the treatment and control groups.[12] For example, if the unobserved factors linked to both treatment and outcome do not change with time, the analyst can use fixed-effects or first differencing regression analysis to isolate the unobserved heterogeneity in the system. With fixed-effects analysis, dummy variables representing the subjects can be included to capture unobserved subject-specific heterogeneity.[13] Public management scholars have employed fixed-effects estimation in a variety of contexts to capture unobserved characteristics, including to control for unobserved features of public organizations, such as local governments (Boyne, James, John, and Petrovsky 2010), schools (Grissom and Keiser 2011), and administrative agencies (Moynihan and Landuyt 2008). With first differencing (also referred to as "differences-in-differences"), the analyst regresses a first differenced outcome on all first differenced time-varying variables as well as time period dummy variables. The precise tactic will change depending on the relative size of time to sample size (see Wooldridge 2002 for a complete discussion of these models). Recent examples include Hanushek and Wößmann's (2006) cross-national study of educational tracking and performance, and Gordon's (2009) study of partisan bias in public corruption prosecutions. Although both of these techniques control for non-time varying subject-specific heterogeneity, they do not control for time-varying factors. In other words, if an important (correlated with both treatment and control) time-varying variable is excluded, then the unobservable correlation between treatment and outcome remains.

Last, the analyst can explicitly estimate the selection process through a joint statistical model of both the selection and outcome. The most popular application of

[12] Other research designs that may be available with panel data are regression discontinuity (RD) and interrupted time series. With RD, cases are assigned to treatment and control groups based on where they fall along an observed threshold or cutoff score (generally referred to as the "assignment variable"). Cases with values above this threshold are assigned to treatment and those with values below to control. Assuming further that the outcome variable is a continuous function of the assignment variable, particularly near the threshold, a local treatment effect can be estimated using the regression coefficient on the assignment variable. An obvious advantage of RD designs is that assignment to treatment or control does not require ex ante randomization, but an analyst must be able to justify that the assignment is essentially random (i.e., the case cannot determine assignment status). Interrupted time series can be also used to evaluate the effect of an intervention (e.g., management reform, policy change) in situations where an analyst has a pre- and post-intervention measure of the outcome of interest. The intervention variable is normally measured dichotomously, and a regression coefficient on this variable can be used to estimate the treatment effect, assuming that all other factors accounting for differences between pre- and post-intervention periods affecting the groups are controlled for in the model.

[13] In practice, rather than including unit-specific dummy variables that can consume substantial degrees of freedom, the data can be "de-meaned" by subtracting the unit-specific mean from both the outcome and independent variables. Resulting analysis then allows you to estimate treatment effects "within" the units of interest.

these joint models derives from Heckman (1979). In the Heckman family of selection models, the analyst adjusts OLS estimates of the conditional mean by adding an estimate of the probability of a subject having received the treatment.[14] This is accomplished by first estimating a probit model that predicts the probability of a subject receiving the treatment, conditional on a set of covariates. The parameters from the probit model are then used to include a correction in the OLS outcome equation for the probability that each subject is observed in the sample. These models can be estimated with either a two-step procedure in OLS or a maximum likelihood jointly estimated procedure. There are also a host of models available for selection processes when the outcome of interest is not normally distributed (e.g., a dichotomous outcome). We include a detailed discussion of a Heckman model below.

Case Selection and Threats to Internal Validity: Selection on the Dependent Variable

Case selection on the dependent variable represents another form of potential violation of the independence conditions discussed above because it can artificially limit variance in the dependent variable and, in doing so, induce dependence between the outcomes of interest and the treatment. Case selection on the basis of values of the dependent variable commonly occurs in three scenarios: sampling on a subset of the population (truncation), limited information due to restrictions based on measurement or data availability (censoring or partial observability), or absence of variation in the dependent variable (due to non-random case selection). Each of these scenarios of case selection can have profound consequences for drawing valid causal inferences (see Breen 1996 for a more detailed discussion).

Consider the case of truncation first, which occurs when a case is only observed for a value of the dependent variable that is either above (left truncation) or below (right truncation) a single threshold, or between two thresholds (double truncation). As a result, values on both the dependent and independent variables for truncated cases are not observed. The practical impact of this truncation is that the mean of the truncated distribution is now different from the mean of the original one. To correct for its shifting mean, the analyst must rescale the truncated distribution by including a weighted estimate of the proportion of the distribution that has been truncated (the weighted inverse Mills ratio). If the analyst wishes to draw inferences between the treatment and the outcome for the larger population, then to proceed with OLS without such rescaling will result in biased and inefficient estimates (Achen 1986; King, Keohane, and Verba 1994). To exclude the rescaling term would essentially introduce omitted variable bias into the analysis.

Consider a study of academic performance for students participating in a newly established charter school for the gifted. If we model the performance rates of these students as a function of a set of covariates, we are likely to have bias in our estimates

[14] Multi-equation Heckman selection models are also referred to as the “control function” method to selection, where the outcome equation is adjusted to account for selection bias by including a “control” in the form of an estimated quantity of the probability of being selected into treatment.

if we seek to draw inferences about the larger population of school-age children. It is likely that only students above some minimum academic performance are admitted to the school—those determined by some decision rule to be gifted. As a result, information on both the dependent and independent variables is missing for lower performing children, and observations with sufficiently large errors are eliminated from the sample. As a given independent variable increases, the expected value of the error becomes larger and increasingly correlated. Because this contradicts the standard assumption of OLS that the error and the independent variables are uncorrelated, regression estimates become biased. In particular, estimating an OLS model on truncated data will result in a flattening of the regression line and an underestimation of the effect of the treatment on the outcome of interest.

Dependent variables with truncated distributions are manageable with a statistical estimation strategy that employs an estimator that specifically recognizes and models the presence of the truncation (see Kotchen and Moore 2007 and Naper 2010 for recent examples). The analyst will usually be required to have knowledge about the threshold level of the dependent variable either below or above which (or between) the truncation occurred. Such estimation strategies can yield unbiased estimates of the desired treatment effect for the larger population, and the analyst can also examine differences in the marginal effect of the treatment on the outcome for the truncated sample and the population.

Another type of selection on the dependent variable is one in which values of the dependent variable are artificially restricted. Two cases that are common in public management research are censoring and partial observability. In each of these cases, the analyst has a latent dependent variable or variables of interest (which is/are not fully observable) and an observable dependent variable. In the censoring case, the observable dependent variable is equal to the latent variables when above (below) some value and equal to zero when below (above) it. In the partial observability case (more specifically bivariate probit with partial observability), the observable dependent variable is equal to one when both latent variables are equal to one and is equal to zero for all other cases. In both situations, data on independent variables are available for all cases.

Censoring differs from truncation in the type of information known about the censored cases. Unlike truncation, where all information about truncated cases is unobserved, with censoring, the observable values on the dependent variable are artificially fixed at the threshold level. And, unlike truncated cases, censored cases' values on the independent variables are all observed. In public management research, such censoring can occur in a variety of situations. For example, if we wanted to measure employee or organizational performance, we might use a standardized scoring method on a scale of 0–10, where 10 represents the maximum score possible under the evaluation rubric. It is possible, however, that an employee or organization's performance exceeds the maximum score of 10. Proceeding without correcting for this censoring will lead to inconsistent causal estimates and will yield biased estimates of the parameters associated with the independent variables in the analysis. The tobit (Tobin 1958), and its many varieties, is perhaps the most well-known estimator of censored data that is commonly used to correct for these issues. As with truncation, the analyst will usually be required to have knowledge about the censoring thresholds. Recent applications of censored data in public management research include investigations of student exam results (Boyne and Chen 2007), charitable giving (Brooks 2003), and competitive bidding (John and Ward 2005).[15]

In the case of partial observability, the observable outcome is a dichotomous variable that is generated from two latent dichotomous dependent variables, whose complete distributions cannot be observed and is akin to censoring (Greene 2002, 664). Consider a situation in which two actors must agree in order to generate some policy outcome (an administrative decision, a consent order, etc.). The latent dependent variable for each actor may be dichotomous (with a one assigned to agree and a zero assigned to disagree), yet the observable data only take the form of a one in the presence of both actors agreeing. The observable data take on the value of zero in three cases: one actor agreed although the other disagreed, vice versa, or both actors disagreed. Yet, the analyst only observes the joint outcome. Partial observability is not always thought of as a "case selection" issue, but it is similar to censoring in the sense that information on the dependent variable is limited in a systematic way that can affect the quality of causal inferences.

A standard example of this process in public policy is subject compliance with legal obligations or regulation. Subject data may exist in a compliance dataset with a dichotomous coding, where a value of one reflects non-compliance and zero reflects compliance. But if we imagine this outcome as the joint product of two actors' decisions (the subject's decision to violate and the agency's decision to detect the violation), then we have a situation where we cannot observe whether a zero in this dataset jointly reflects: non-compliance and non-detection, compliance and detection, or compliance and non-detection. Proceeding with a standard estimator (probit or logit) in the presence of such a process will result in biased estimates. A popular estimator of this process is detection-controlled estimation (DCE) (Feinstein 1990), which uses a bivariate probit model with partial observability (Poirier 1980), and we provide a detailed example of its use later in the article.

Last, analysts may elect to only investigate cases that have the same value on the dependent variable. This is generally considered to be the "worst case" of selection on the dependent variable, because an analyst is likely to make one of two inferential errors (Geddes 2003). First, the analyst may conclude that a shared characteristic among the cases must be a cause of the outcome. Second, the analyst may conclude that relationship between the variables in the sample reflects relationships in the entire population. As King, Keohane, and Verba (1994) remind us, it is impossible to learn about causal effect by selecting observations whose outcomes do not vary. This is because when conditioned on the selection of the dependent variable, estimates of the causal relationship are likely to be biased toward zero. As a result, the analyst is likely to accept the null hypothesis in the presence of an actual relationship.

Some scholars (Dion 1998) have suggested that selection on the dependent variable as a research design strategy is a useful way to evaluate hypotheses of necessary

[15] At the time of correcting for the aforementioned biases, the analyst must carefully interpret the coefficients in a tobit model. The coefficients represent the marginal effect of x_i on the latent variable (demand for participation), not on the actual observed variable (actual participation). See Greene (2003) for the relevant equation to calculate the marginal effect of x_i on y. In addition, the analyst can also use the tobit to estimate the effect of a covariate on the probability of a case being censored or uncensored (see Long 1997, 196–210, for a useful discussion).

or sufficient conditions—so called "crucial case" (Eckstein 1975) tests. Consider the two following scenarios. If X is thought to be a necessary condition for $Y(1)$, or rephrased *if* $Y(1)$ *then* X, and the analyst gathers data on all cases with a value of $Y(1)$ and demonstrates at least one case with $\sim X$, then s/he would have falsified the necessary condition hypothesis. Alternatively, if X is thought to be a sufficient condition for $Y(1)$, or rephrased *if* X, *then* $Y(1)$, and the analyst gathers data on all cases with a value of X and demonstrates at least one case with $\sim Y(0)$, then s/he would have falsified the sufficient condition hypothesis. The crucial case argument, however, is based on at least two implicit assumptions: the absence of measurement error and a deterministic rather than probabilistic theoretical framework. As Braumoeller and Goertz (2000) point out, if we allow for the possibility of measurement error of our concepts, then, for example, we may have actually miscoded $Y(0)$ as $Y(1)$, or X as $\sim X$. If this were the case, then the presence of an incongruent case may not be sufficient to declare a necessary condition as falsified. Moreover, if we recognize the probabilistic nature of social science theory, then observing one non-conforming case is likely not sufficient to refute a causal claim (see Braumoeller and Goertz (2000) for a full discussion). In either case, the conditions for selection on the dependent variable seem stark enough to all but rule out its usefulness in examining causal inferences in the social sciences.

In each of the cases described above—truncation, censoring, partial observability, and complete artificial selection of the dependent variable—the internal validity of causal inference may be affected. As we discuss next, case selection can also have implications for the external validity of causal inferences.

CASE SELECTION, CAUSAL INFERENCE, AND THREATS TO EXTERNAL VALIDITY

Scholars often make explicit case selection choices on policy area, geographic area, or time period, and these decisions can raise important external validity problems. In public management research, this problem most often arises in research designs that consider only one bureau or one office within which to conduct an analysis and/or only for a limited time period.

Imagine, for example, we were to conduct an analysis of the effect of public service motivation on job satisfaction in public agencies and had data from a survey conducted in 2008 of civil servants working at the US Environmental Protection Agency (EPA). This type of study would be vulnerable to at least two threats to external validity. First, the opinions of the EPA civil servants may not reflect those serving in other federal (or state or local) administrative agencies. Second, there may be unique features of the time period that we chose to study. For example, civil servants working at the EPA in 2008 would be serving after 8 years of general policy retrenchment in the Bush administration. In general, then the key question is how likely is it that the inferences drawn from the sample would apply to "out of sample" populations? How might we estimate this likelihood?

Although problematic to generalizing findings, selection effects on external validity may potentially be less worrisome than those on internal validity. Analysts are generally aware of the sample to which their inferences apply, and they can explore two specific challenges to the external validity of their study. First, they can explicitly address the extent to which the given context in either policy or time period differs from other contexts and provides readers with a sense of how likely the inferences will "travel." Second, they can explicitly consider what types of bias might occur if inferences drawn from a time- or case-specific study were applied to the larger population.

In a practical sense, the analyst interested in quantifying the likelihood of external validity problems could provide the reader with a sense of how the main variables of interest are likely to vary by the temporal and geographic or organizational focus of the current sample. If there is reason to believe that either the dependent variable or the main independent variable (i.e., the treatment) varies non-randomly over these contexts, then threats to external validity exist. Returning to the EPA example above, the analyst should be able to provide an informed assessment of whether the inferences made about the effect of public service motivation on job satisfaction are biased in a particular direction, given the agency, policy area, and period of study. The analyst could then provide the reader with an assessment of the nature and direction of likely bias of applying the causal inference in the analysis to another population. On the other hand, if these variables exhibit little correlation with the characteristics of the sample (time period, spatial domain, policy area, organization, etc.), then the analyst can be rest assured that the external validity of the analysis is likely preserved.

CASE SELECTION ISSUES IN RECENT PUBLIC MANAGEMENT RESEARCH

To assess the relative occurrence of case selection issues in contemporary public management publications, we reviewed all research articles published in the 2009 and 2010 volumes of *JPART* as well as the first two volumes of *JPART* in 2011.[16] For each of the 93 articles reviewed, we coded whether the authors introduced explicit hypotheses and/or attempted to draw causal inferences from their work. We also coded for the presence of potential case selection issues as well as whether and how authors attempted to address them.

Of course, we note with some irony that choosing these particular articles for review raises its own question about case selection. One issue that is particularly important to highlight is that we only reviewed published *JPART* articles, since we do not have access to submitted but rejected manuscripts over the time period. As such, our assessment of the degree of case selection problems is vulnerable to potential systematic differences between published and rejected pieces and the selection issues that we observe. It is quite possible, in fact, that among the set of reasons that submitted manuscripts were not published in *JPART* is a belief among the anonymous reviewers or the editors that a study may have insufficiently addressed a case selection issue. If this is the case, our estimate of the average number of unresolved selection issues in our sample of published articles is likely a conservative estimate of the extent of the problem in the larger population of submitted manuscripts.

[16] Separate articles written by the authors each appeared in the second issue of the 2008 volume of *JPART*. In different ways, each article also suffers from potential case selection issues; so we in no way excuse our own work from the criticisms of this review.

Table 2 reports the results for the types of potential selection issues that we observed in this sample. In the past few years, scholars publishing in *JPART* have been primarily interested in evaluating causal inferences with the large majority of authors testing or drawing causal inferences from their work. Moreover, a high percentage of publications explicitly stated hypotheses to be empirically analyzed (59% in 2009, 72% in 2010, and 53% in 2011).

The modal research design for *JPART* articles in this time frame was large-n quantitative (76%), with a sizeable minority utilizing a small-n design (16%). Out of 93 articles, only three employed a single observation research design. Although authors of all three of these studies made causal inferences based upon their single observation, they were all careful to acknowledge the limitations of their inferences.

With respect to specific case selection issues, we coded for four potential issues. We first assessed whether the studies relied on non-random case selection. If information was available to determine the specific nature or consequence of the non-random case selection (e.g., selection on the dependent variable, no variance on the dependent variable, or non-random assignment of the treatment), we then coded these sub-categories. If information was insufficient for making this determination, we simply coded the article as appearing to have non-random case selection. As a result, the sub-categories under “Non-random Case Selection” in Table 2 may not be exhaustive. Moreover, a given study could possess one or more selection issues so the sum of the sub-categories is not restricted to the total cases coded.

Many of the large-n and small-n research designs appeared to contain non-random case selection and thus potentially are subject to the types of issues previously highlighted. Our review, however, did not reveal there to be rampant selection on the dependent variable. For both large-n and small-n research designs, these types of problems pertained to just a small minority of the published studies in *JPART* in this two-and-a-half-year period. The more frequent issue regards possible non-random assignment of cases to the treatment group. A fairly standard example of this type of potential selection effect was scholars using a pre-existing survey of a set of respondents and assuming either that the response rate for the survey was randomly determined and/or that a given variable of interest in the survey (which happens to be the main treatment variable) was randomly assigned or selected.

To assess how *JPART* scholars negotiate selection issues, we also coded whether authors explicitly acknowledged potential case selection problems in their work. The second portion of Table 2 shows the results of this analysis. On selection issues relevant to external validity, roughly a third of all the reviewed articles included a discussion of whether their causal inferences based on a single policy area or time period might apply to other contexts. Authors who included this discussion overwhelmingly did so informally, offering brief discussions of their “sense” of whether their inferences could travel to other populations. Very rarely did these discussions precisely evaluate how their sample may differ from the larger population to which they were seeking to apply their inferences. According to our review, roughly half of the researchers using large-n or small-n designs acknowledged selection concerns. To be clear, if the authors did not explicitly discuss a potential case selection problem, we noted this as a potential problem in our accounting, but the authors may very well have addressed the problem through the use of control variables, etc. Of the articles that we believed

Table 2
The Distribution of Selecton Issues across Recent *JPART* Volumes

	JPART 2009		*JPART* 2010		*JPART* 2011	
Selection Issues/Remediation	Number	Percentage	Number	Percentage	Number	Percentage
Total research articles	42		36		15	
Articles with explicit hypotheses	25	59.52	26	72.22	8	53.33
Articles claiming to draw causal inferences	31	73.81	35	97.22	13	86.67
Selection issues						
External validity						
Articles using a single policy area	23	54.76	18	50.00	7	46.67
Articles using a single observation (n = 1)	1	2.38	0	0.00	2	13.33
Internal validity						
Articles using small-n research design (1 > n < 35)	7	16.67	6	16.67	2	13.33
Non-random case selection	6	85.71	4	66.67	2	100.00
Selection on the dependent variable	2	4.76	1	2.78	1	6.67
No variance on the dependent variable	0	0.00	0	0.00	0	0.00
Non-random assignment of treatment on main independent variable	6	100.00	3	75.00	1	50.00
Articles using large-n research design (n>35)	30	71.43	30	83.33	11	73.33
Non-random case selection	27	90.00	23	76.67	6	54.55
Selection on the dependent variable	0	0.00	1	4.35	0	0.00
No variance on the dependent variable	0	0.00	0	0.00	0	0.00
Non-random assignment of treatment on main independent variable	17	62.96	22	95.65	6	100.00
Selection issue remediation						
External validity						
Articles using a single case area that examined external validity	7	30.43	9	50.00	4	57.14
Articles using small-n research design with selection issue	6		4		2	
Acknolwedge selecton process at work	0	0.00	2	50.00	1	50.00
Uses research design to remediate selecton process	0	0.00	1	25.00	1	50.00
Articles using large-n research design with selection issue	27		23		6	
Acknolwedge selecton process at work	15	55.56	13	56.52	4	66.67
Uses research design to remediate selecton process	6	22.22	9	39.13	1	16.67

to exhibit at least one potential selection issue, roughly a quarter attempted to remedy them with explicit changes or treatments in their research design.[17]

DETAILED EXAMPLES OF CASE SELECTION SOLUTIONS

Many techniques have been developed to address the case selection issues that often occur in social science research, particularly in areas where randomized experimental data are not available. In this section, we review three such techniques: propensity score matching, Heckman sample selection, and DCE, which have been developed to address specific types of problems with observable selection on the treatment variable, unobservable selection on the treatment variable, and partially observable information on the dependent variable, respectively. Each of these techniques has been available for some time and applied widely in the social sciences. Yet, from our formal review of recently published articles in *JPART* and a more informal review of other leading public administration journals, they appear to be underutilized in public management scholarship. Our purpose in this section is neither to present formal derivations of these techniques nor to suggest that these tools are the only or even the best solution to the specified case selection problem. Rather, our intention is to illustrate the utility of each approach through a detailed example. We do this through both original data analysis in the case of DCE and summaries of exemplary research from other scholars in the case of propensity score matching and Heckman models.

Propensity Score Matching

A main benefit of randomization for estimating causal effects is that it provides some reasonable assurance that those exposed to the treatment group are statistically equivalent to those in the control group, in terms of both observed and unobserved differences. In a randomized experiment, thus, we can simply calculate the average treatment effect as the average difference in outcomes of cases treated with those not treated. In observational studies, there is no random assignment of the treatment, making it difficult to attribute causality to the specific treatment in question, be it a policy intervention, an institutional reform, or otherwise. In particular, one must worry that the differences in outcomes are attributable to differences in the groups themselves. This is a problem of self-selection, in which individuals who elect to receive the treatment are systematically different than those who do not in ways that may be related to the outcome of interest.

As we summarized above, one approach to this problem is to use a process of matching to condition the average treatment effect on observable, pretreatment attributes. There are several different procedures for pairing cases in a matching analysis (Sekhon 2008). One could match based on a single characteristic or set of

[17] We should note that our estimate of *JPART* articles with selection issues is a conservative one, biased toward coding any potential issue that may have affected causal inferences. We were unable to assess the magnitude of any bias or the importance of the potential issue relative to another econometric issue whose solution may have presented the authors with a tradeoff.

characteristics. That is, an analyst can stratify the data into categories or bins according to their values of the characteristic. This is relatively straightforward in the case of a single, dichotomous characteristic that takes just two values. As the number of characteristics (and potential values) increases, the number of bins increases exponentially, making it difficult to obtain exact matches, causing what is generally referred to as the dimensionality problem. Propensity score matching, as first proposed by Rosenbaum and Rubin (1983), deals with this dimensionality problem. The propensity score is defined as the probability of a case having received the treatment, given the set of observed characteristics. By focusing on this probability, the dimensionality problem is reduced.[18]

The propensity score itself is usually computed using a logit or probit model. Matches are then determined in terms of the probability of having received the treatment. Various procedures have been developed to pair similar cases together based on the closeness of their propensity scores, with the most straightforward being "nearest neighbor" matching where one case is chosen as matched partner for the treated case. For more technical treatments of propensity score matching, we refer readers to fuller discussions elsewhere (e.g., Dehejia and Wahba 2002; Rosenbaum and Rubin 1983, 1984).

Propensity score matching methods have been widely adopted in the social sciences over the past couple of decades but much less so in the public management literature. Our review of recent articles published in *JPART*, for example, identified only Heinrich's (2010) evaluation of the effectiveness of supplemental educational services (SES) for students under the No Child Left Behind (NCLB) law as utilizing this method. In fact, Heinrich's study appears to be the only propensity score matching analysis appearing in print in *JPART* over the last 20 years.[19]

To illustrate the usefulness of propensity score matching, we briefly describe Heinrich's study.[20] The central purpose of the analysis was to evaluate whether students receiving SES from third-party providers (private non-profit and for-profit organizations) achieved better educational outcomes, as measured by improvements in standardized test scores. The study is important because it evaluates a key feature of the NCLB law and because it assesses the role of non-governmental, third parties providing services that might otherwise be delivered by government entities (i.e., public schools). The specific case selection concern stems from possibility of systematic differences between students who received SES (i.e., the treatment group) and students who did not receive SES. If unaccounted for, the average differences in test outcomes between students enrolled in SES programs and those not enrolled in SES programs

[18] Another common approach is multivariate matching using Mahalanobis distance, which is a metric representing the dissimilarity between a vector of characteristics of the treatment group and the control group. Each observation in the treatment group is matched with the closest one in the control group (Cochran and Rubin 1973; Rubin 1979, 1980). Propensity score matching and Mahalanobis metric matching can also be combined. Recent work by Diamond and Sekhon (forthcoming) provides an alternative technique they refer to as genetic matching for situations when the distributional requirements of the other approaches do not hold.

[19] Based on the results of a search of the *JPART* archive with the keywords "propensity score matching." Miller and Nicholson-Crotty (2011) note that their supplementary analysis use matching techniques as well.

[20] The study in *JPART* was based on analysis largely reported in a separate article (Heinrich, Meyer, and Whitten 2010).

Table 3
Propensity Score Matching Results Compared with Unmatched Results from Heinrich's Study of Third-Party Provided, Supplemental Educational Services and Test Scores

Treatment Measure and Method	Change in Math Test Scores	Change in Reading Test Scores	Change in Math Test Scores	Change in Reading Test Scores
	2004–2005 School Year			
SES participation	Middle School		High School	
No matching	−2.486 (4.562)	−3.368 (5.232)	−10.486 (6.243)	−14.420 (7.139)
Matching	2.024 (5.557)	3.038 (5.916)	−5.427 (8.107)	−4.565 (8.860)
Observations	1,562	1,571	1,224	1,262
	2005–2006 School Year			
SES participation	Middle School		High School	
No matching	−0.529 (0.413)	0.708 (1,202)	0.235 (0.297)	2.846 (1.132)
Matching	−0.232(0.427)	0.323 (1.099)	−0.372 (0.357)	1.397 (1.099)
Observations	1,075	1,016	2,215	2,200

Note: Reprinted with permission from Heinrich (2010).

might be attributable to things about these students rather than the educational services themselves.

Heinrich reports results using unmatched and matched samples, where the matched samples were generated using propensity score matching on observable characteristics collected from student transcript and administrative data. Specifically, the propensity scores are the predicted probabilities from a logistic regression model of registration for SES, using a combination of student and school characteristics. The matching is then done between eligible students who registered for SES and those eligible students who did not register for SES. The outcome of interest is student achievement, measured using test scores for reading and math for middle school and high school students in the Milwaukee Public School system from the 2004–5 and 2005–6 school years. The design of the study is such that the outcomes—test scores—are measured both before and after the treatment, so the models are estimating the effect of SES on changes in student achievement.[21]

The core finding of this part of the analysis was that "after matching participants and nonparticipants on their baseline characteristics, there are no statistically significant differences in the changes in test scores for students who attended SES compared with those who did not attend SES" (Heinrich 2010, i68). We summarize the relevant results in Table 3 to illustrate the differences between inferences that would be made with potential selection bias (no matching) with those made with this selection bias accounted for (with matching). Most of the coefficients do not reach statistical significance in either the unmatched or matched analysis. However, in the case of changes in reading test scores among high school students, the coefficients in the analysis without matching suggested that SES participation resulted in a 14-point decline in test scores in 2004–5 and about a 3-point increase in these test scores in 2005–6. In the

[21] The "differences-in-differences" estimator used in the study further allowed the authors to address potential unobserved differences in student characteristics.

matching analysis, neither coefficient is statistically significant. Thus, the results without addressing the potential problem of selection on the treatment variable would have led to a conclusion that SES affected test scores (although in different directions in the two school years), whereas the matching analysis suggests no such differences.

Heckman Selection Model

Matching techniques are an appropriate solution for cases where selection on treatment is due to known, *observable* factors. In many situations with observational data, however, there is also the possibility that there is an imbalance between *unobserved* factors.[22] Stated differently, this is a situation when there are unmeasured factors correlated with both the outcome of interest and the selection mechanism. The classic way to address the selection on unobservables problem is by using the sample selection model developed by Heckman (1979). The Heckman model is a two-step procedure in which, first, a selection equation is estimated via a standard probit specification. This model includes independent variables that are thought to be correlated with a case receiving the treatment. The second part of the two-step procedure is an OLS regression in which the dependent variable is the outcome of interest, a set of independent variables, and the inverse Mill's ratio, which captures the extent of correlation in the errors in the two equations. If the coefficient on the inverse Mill's ratio is statistically significant, this provides evidence that there is selection bias. For more technical treatments as well as important distributional and identification assumptions, we refer readers to a variety of sources (Achen 1986; Greene 2003; Heckman 1979; Sartori 2003; Wooldridge 2002).

The Heckman model to address selection bias is now more than 30 years old, and its application is ubiquitous in the economics literature. These models, however, do not appear with much regularity in the public management literature. A search of the *JPART* archive revealed only a handful of research articles employing a Heckman selection model in either main or supplementary analysis (Daley 2009; Dull 2009; Georgellis, Iossa, and Tabvuman 2011; Lavertu and Weimer 2011; Pandey and Bretschneider 1997; Serra 1995) since 1991.[23]

We illustrate an application of the Heckman model using Lavertu and Weimer's (2011) recent study on drug and medical device approval at the Food and Drug Administration (FDA). Specifically, their analysis investigates the degree of influence of advisory committees on the FDA's approval of pharmaceutical drugs and medical

[22] Between the two popular econometric solutions to selection bias that we review here—matching and Heckman selection models (control functions)—analysts must recognize a tradeoff between them. Standard advantages of matching, with a set of known conditioning variables, include: not requiring conditioning variables to be exogenous, the absence of exclusion restrictions, and no specific functional form of the outcome equations (Heckman and Navarro-Lozano 2004, 33). Yet, matching underperforms in the presence of perfectly predicted treatment by conditioning variables. Control function approaches are robust to such omissions, given their explicit modeling of omitted relevant conditioning variables. Moreover, unlike standard matching, control function approaches do not include an implicit assumption that the average treatment effect is equivalent to the marginal treatment effect (Heckman and Navarro-Lozano 2004). Control functions are, however, sensitive to the precise nature of the exclusion restrictions and the assumption of bivariate normality (when violated, estimates are likely to be inconsistent).

[23] Based on the results of a search of the *JPART* archive with the keywords "Heckman model."

devices over the 10-year period from 1997 to 2006. The study is multifaceted, considering the strategic use of committees by the FDA by modeling the decision of the agency to consult committees as well as the effect of committee consultation on overall review durations. In addition, they study the direct influence of advisory opinions on the probability that a drug is approved for use.

Lavertu and Weimer model the FDA's expected utility for consulting with a federal advisory committee as part of their decision-making process for drug and medical device approval. They argue that the agency's decision to consult an advisory committee is a function of the organizational benefits that come from making a good decision and the political costs stemming from interest groups that may agree or disagree with the outcome. They further argue that the FDA will be more likely to approve a drug or medical device as more members of the advisory committee favor such a decision, and less likely to approve when the political costs increase, and more likely to approve when the political costs decrease. The potential selection problem emerges when estimating their model of approval, because there may be common factors related to the original decision to consult.[24] More specifically, there may be unmeasured or unobserved factors in their consultation and approval models that are correlated. For this reason, they estimate Heckman selection models in their analysis of FDA approval decisions, where the selection model includes the hypothesized factors predicting the agency's decision to seek the opinion of an advisory committee. Because the approval decision is measured dichotomously (1 = approve and 0 = disapprove), they estimate Heckman probit models, which are equivalent to bivariate probit models, and follow the same logic as described above.

Lavertu and Weimer presented the results from three separate Heckman probit models in Table 4 of their article. Their first model is a "parsimonious specification" that models drug approval as a function of three factors: the proportion of the advisory committee that favored approval, whether the consumer representative on the committee voted to approve, and a count of the interest groups that focus on the particular disease in which the drug is meant to treat. The selection equation includes a number of factors, including whether the drug is a "new molecular entity" (NME), whether the review was fast-tracked as a priority review or accelerated review, whether the drug is considered orphan (i.e., a drug that is meant to address a rare condition), and whether the drug application was before or after 2001 to mark the differences in their study period between the Clinton and Bush administrations. Their first model applies to only "primary indications" or cases of single votes of the committee. The second model they estimate includes an additional set of controls, whereas the third model includes these controls and expands the sample to "multiple indications," which includes each vote taken on drugs since some applications are voted on more than once.

Table 4 displays the estimates from Lavertu and Weimer's Heckman probit models, alongside simple probit models of the approval decisions using the same data. Presenting the results side by side enables a comparison of models that account for

[24] Lavertu and Weimer's selection models are limited because of data constraints. Specifically, they did not have cases of non-approved drugs that were not referred to an advisory committee. They clearly note this limitation in the text of their article.

Table 4
Heckman Probit Selection Results Compared with Probit Results from Laveretu and Weimer's Study on FDA Drug and Device Approval

	Heckman Probit	Probit	Heckman Probit	Probit	Heckman Probit	Probit
	(1)		(2)		(3)	
Approval equation	*N* = *129*		*N* = *129*		*N* = *178*	
Proportion of committee voting yes	2.19*** (0.47)	2.36*** (0.43)	2.60*** (0.76)	2.63*** (0.76)	2.04*** (0.49)	2.12*** (0.49)
(Proportion yes) × (Filed before 2001)			−0.36 (0.97)	−0.37 (0.98)	−0.05 (0.62)	−0.6 (0.64)
Filed before 2001			0.78 (0.69)	0.75 (0.68)	0.43 (0.44)	0.34 (0.45)
Consumer committee member voted yes	−0.22 (0.30)	−0.19 (0.32)	−0.18 (0.32)	−0.18 (0.32)	0.03 (0.26)	0.03 (0.27)
New molecular entity			−0.75 (0.61)	−0.86* (0.38)	0.24 (0.24)	0.07 (0.24)
Priority review			1.61*** (0.41)	1.57*** (0.39)	0.77** (0.25)	0.64** (0.25)
Orphan drug			−1.12* (0.45)	−1.16* (0.46)	−0.52 (0.28)	−0.57* (0.29)
Accelerated approval			1.07* (0.47)	1.05* (0.47)	0.31 (0.30)	0.22 (0.31)
Interest group count	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.01 (0.01)	0.01 (0.01)
Washington post			0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Constant	−0.11 (0.59)	−0.78*** (0.25)	−1.11 (1.35)	−0.78 (0.48)	−1.74*** (0.39)	−1.12*** (0.34)
Selection equation	*N=987*		*N=987*		*N=1,036*	
New molecular entity	0.74*** (0.14)		0.79*** (0.12)		0.71*** (0.11)	
Priority review	0.56*** (0.17)		0.48*** (0.14)		0.63*** (0.13)	
Orphan drug	0.25 (0.19)		0.31 (0.18)		0.16 (0.17)	
Accelerated approval	0.30 (0.20)		0.29 (0.21)		0.46* (0.19)	
Filed before 2001	0.30** (0.11)		0.28* (0.11)		0.46*** (0.10)	
Constant	−1.78*** (0.009)		−1.77*** (0.009)		−1.69*** (0.009)	
Model						
Rho	−0.42 (0.29)		0.16 (0.54)		0.32* (0.14)	
Wald chi-square	27.25***	38.37***	43.65***	42.01***	66.46***	59.65***

Note: Two sided z-tests or chi-square tests: **p* < .05; ** *p* < .01; ****p* < .001.
Reprinted with permission from Lavertu and Weimer (2011).

selection bias with those that do not. (For clarity of presentation, we have renamed some of the variables, but we refer readers to the original article for detailed definitions.)

The first thing to note is the Rho coefficients displayed below the table. In the first two models, the coefficient on Rho is null, suggesting that there may not be an issue of sample selection bias present in these data. In the third model, however, the coefficient is statistically significant signaling the presence of selection bias. This suggests that estimates from a standard probit model of FDA approval would have yielded statistically inconsistent estimates.

The variable of central theoretical interest in this part of Lavertu and Weimer's study is the proportion of the advisory committee that recommended that the FDA approve the drug. As hypothesized, the more support in the committee for the drug, the higher the probability of FDA approval. This core result emerges from both the Heckman models accounting for selection bias and the standard probit models ignoring it (even in the model where there is an indication of selection bias). With respect to this particular variable, the authors would not have drawn different inferences if they had simply estimated probit models of the approval decision. It is important to note that this is not always (or even often) the case.

Detection Controlled Estimation

As we noted above, most agree that research designs that select on the dependent variable are not useful for causal analysis. However, this type of investigator-induced selection bias is not the only type of selection issue that emerges with respect to the dependent variable. Here we focus on the problem of partial observability of a binary outcome of interest, which is a type of censoring problem that can occur in a variety of public management research contexts, including in studies of administrative agency enforcement of laws and regulations. In these cases, scholars are interested in evaluating factors that affect compliance decisions of regulated entities (this might include individuals, firms, or non-profit organizations). An observation of compliance in the data used by scholars, however, will reflect two possible data generation processes. First, it may signify actual compliance—that is, a regulated entity that has been found to be behaving as required. Alternatively, an observation of compliance may reflect *non-detection*—that is, a regulated entity that is actually violating a law, regulation, or other obligation, but one that has not been discovered by the administrative agency. This case is coded as compliant in the data, but it is indistinguishable from a case of actual compliance. Failure to account for these two reasons for observing "compliance" can bias inferences in a causal analysis.

To illustrate this problem, consider Feinstein's (1990) example of firm compliance with health and safety regulation. In this study, Feinstein had data on compliance with Occupational Safety and Health Administration (OSHA) regulations for industrial facilities the agency inspected in New England states in 1985, the specific inspector performing the monitoring activity, as well as various firm characteristics such as whether workers at the facility belong to a union, the number of employees on site, the total employees working at the firm, the type of industry, and the state unemployment rate. So, although Feinstein had data on the compliance status of each facility, instances of true compliance were indistinguishable from instances of compliance because of non-detection of a violation with an OSHA requirement.

To account for the non-detection problem, Feinstein developed what he called detection-controlled estimation (DCE), which consists of two binary choice models: one that models the likelihood of a violation (1 = violation, 0 = compliance) and a second that models the likelihood of detection (1 = detection, 0 = non-detection). Because the likelihood of a violation and the likelihood of detection are separately unobservable, these likelihood functions are estimated jointly via maximum likelihood estimation. Details of the estimator are available elsewhere (Feinstein 1990), and the model is the same as a bivariate probit model with partial observability (Abowd and Farber 1982; Poirier 1980). In his OSHA example, Feinstein demonstrates differences in the relationship between various covariates and firm compliance when comparing the results from a standard probit model of firm compliance with the DCE model. This set-up is analogous to a multitude of other compliance contexts, and DCE techniques have been used to study taxpayer compliance (Feinstein 1999) and firm compliance with environmental (Brehm and Hamilton 1996; Helland 1998a, 1998b; Scholz and Wang 2006) and FDA (Olson 1995) regulation.

We further illustrate the utility of the DCE model with an original example from data we have compiled on firm compliance with the federal Clean Air Act (CAA).[25] Table 5 reports two analyses of firm-level non-compliance in the year 2004.[26] The dependent variable in this investigation is a dichotomous indicator of firm non-compliance that is coded one for firms that violated the CAA during the 2004 calendar year and zero for firms that did not violate. The universe of cases is all federally reportable facilities regulated by the CAA.

Model 1 reports the results from a standard probit analysis. This model treats the dependent variable as if it was generated solely from a firm's compliance decision and assumes that the variable's distribution is fully observed. We model the dependent variable as a function of a standard, albeit underspecified for simplicity's sake, suite of covariates found to influence a firm's decision about compliance (Helland 1998b; Scholz and Wang 2006), including firm-level and contextual factors thought to affect the relative costs of complying with regulation. For example, to assess whether a previously inspected firm is likely to be a future violator, we include a lagged inspection variable. In addition, we also include a dummy variable for whether the firm is a "major" emissions source (to capture economies-of-scale pressures on compliance), a dummy variable for whether the facility is a manufacturer (SIC codes 20–39), in the transportation sector (SIC codes 40–48), or a power plant (SIC code 49), county-level non-attainment status (which reflects air pollution severity), county-level unemployment, shares of county income derived from manufacturing, and the complexity of the policy environment (policy entropy).

[25] Studying air pollution control raises obvious questions about external validity. Scholars have shown that compliance issues behave similarly with respect to key theoretical expectations across various policy-specific areas, and although the variables used to assess different policy-specific concepts may change, the key inferences on the relationship between political, economic, and policy task context are quite similar.

[26] Selecting 2004 for this example raises an additional question about external validity. We use this year because we already had the data, and we do not have any specific reason to think that 2004 was unique. We estimated the models with data for 2005 with similar results.

An analyst concerned about how these data are generated might be compelled to also control for features of the environment that shape whether a firm is likely to be detected. To this end, we also include a set of covariates that have been linked to the likelihood of non-compliance being detected, including whether the firm has had prior enforcement actions, the relative vertical location of signature authority within the agency (Reenock and Gerber 2008), community demographics (percent minority) and economic characteristics (a scale consisting of four standardized variables: median household income, percent below poverty line, percent college educated, and percent high school educated), and the partisan control of both the state executive and legislative branches. The level at which each variable is measured is indicated in parentheses in the table.[27]

Model 2 reports the results from the DCE model, which treats the dependent variable as if it were jointly generated from two sources: the firm's decision over compliance and the likelihood of a regulatory agency detecting non-compliance.[28] Unlike the standard probit, the DCE model explicitly recognizes these two distinct processes and models them jointly. Moreover, these decisions are partially observed because undetected/compliant firms, detected/compliant firms, and undetected/non-compliant firms are observationally equivalent in the data—they are all coded as compliant.[29] The DCE model in Table 5 reports the effect of covariates on the joint probability of a firm's compliance and detection, estimating a separate model for each outcome. The first model estimates the effect of covariates on a firm's compliance, and the second model estimates the effect of covariates on the probability of detection. Theory guides the choice of variables to be included in each model. The analyst can include all variables (absent one for identification) in both models or can restrict variables in one or both models as theory allows.[30] For the purposes of demonstration, we have opted for minimal restriction between the models, excluding only the previous enforcement and signature authority variable from the compliance model.

[27] The measure and source for each variable in our model are as follows: Previous enforcement action (lagged total state and EPA punitive actions for a given firm), Previous inspection (lagged state or EPA inspection for a given firm), Major source (dummy variable coding stationary source that emits >10 tons per year), Manufacturing firm (dummy variable coding manufacturing SIC code), EPA's IDEA database, Power plant (dummy variable coding power plant SIC code), Transportation (dummy variable coding transportation plant SIC code) all derive from EPA's Integrated Data for Enforcement Analysis system; Signature authority (Reenock and Gerber 2008), % unemployment (% county unemployment, comes from the Bureau of Labor Statistics), Non-attainment (dummy variable coding non-attainment for at least one CAA criterion pollutant, comes from the EPA Green Book), Median household income, % below poverty, % college educated and % High school educated (2000 Census), % minority (% African American + % Hispanic, comes from the Census Bureau), Regional scale (Total number of firms in a state administrative regional office), Democratic Governor (dummy variable), and % Democrats (% Democrats in state house and senate, comes from the Council of State Governments).

[28] See Konisky and Reenock (forthcoming) for a more detailed application of the DCE model to the case of CAA compliance.

[29] The DCE estimator utilizes "two" dependent variables to model the effect of covariates on the variables' joint distribution. Of course, the analyst only ever observes the single compliance variable from a dataset, so the second dependent variable must be generated by the analyst.

[30] Identification of the bivariate probit with partial observability model suggests an exclusion of at least one exogenous variable to ensure that the parameters being estimated in each model are not identical. Moreover, identification is enhanced when the exogenous variable exhibits sufficient variation over the sample. This condition is likely to be met by ensuring that the exogenous variable is continuous variable (Poirier 1980, 212–5).

Table 5
Compliance Status for Individual Regulated Entities

	Model 1		Model 2	
	Pr(Violation)		Pr(Violation, Detection)	
	Probit		Detection Controlled Estimation	
	b	S.E. b	b	S.E. b
Compliance model				
Previous inspection$_{t-1}$ (firm)	0.1585***	0.0254	0.2304***	0.0551
Previous enforcement action$_{t-1}$ (firm)	0.9540***	0.0303	—	—
Signature authority	0.0000	0.0060	—	—
Major source (firm)	0.7726***	0.0270	0.6082***	0.1584
Manufacturing firm (firm)	0.1123**	0.0312	−0.0875	0.1288
Power plant (firm)	−0.0438	0.0400	−0.0511	0.1554
Transportation (firm)	−0.1331	0.0951	−0.2988	0.3539
Non-attainment (county)	0.0244	0.0317	−0.0723	0.2237
% Unemployment (county)	0.0398***	0.0084	0.1701***	0.0551
% Manufacturing income (county)	0.0034	0.0029	0.0354**	0.0159
Policy entropy (county)	0.0226	0.0224	−0.2244**	0.1108
% Minority (zipcode)	0.0014**	0.0007	0.0045	0.0029
Economic characteristics (zipcode)	−0.0590***	0.0194	−0.1072	0.0961
Regional scale (region)	0.0004***	0.0000	0.0001	0.0002
Democratic governor (state)	−0.0508*	0.0278	−0.2652*	0.1360
% Democrats in state legislature (state)	−0.0008	0.0011	0.0058	0.0128
Constant	−2.6228***	0.0809	−2.0686***	0.5445
Detection model				
Previous inspection$_{t-1}$ (firm)	—	—	—	—
Previous enforcement action$_{t-1}$ (firm)	—	—	1.5876***	0.2779
Signature authority	—	—	−0.0071	0.0088
Major source (firm)	—	—	0.5406**	0.2463
Manufacturing firm (firm)	—	—	0.2323**	0.1075
Power plant (firm)	—	—	−0.0009	0.1245
Transportation (firm)	—	—	0.0828	0.3520
Non-attainment (county)	—	—	0.1097	0.1956
% Unemployment (county)	—	—	−0.0838***	0.0327
% Manufacturing income (county)	—	—	−0.0199**	0.0088
Policy entropy (county)	—	—	0.2077**	0.0897
% Minority (zipcode)	—	—	−0.0017	0.0024
Economic characteristics (zipcode)			0.0238	0.0957
Regional scale (region)			0.0004***	0.0001
Democratic governor (state)			0.1871	0.1689
% Democrats in state legislature (state)	—	—	−0.0058	0.0116
Constant	—	—	−0.8191	0.9959
rho	—	—	−0.1510	0.2920
Log-likelihood	−7,191.43		−7,155.78	
χ^2	(16) 2,359.95***		(29) 634.37***	
Cases	38,407		38,407	

Note: *p < .10, **p <. 05, ***p < .01, two-tailed tests. Standard errors clustered on zipcode.

First, note the difference in the interpretation of coefficients across the two models. With respect to compliance, in each model, positive coefficients suggest that, at higher levels of the independent variable, a firm is more likely to be in non-compliance. According to both the standard probit and the DCE models, firms that have been previously inspected, those that are major sources and manufacturing firms, those located in counties that are struggling with employment, and those in poorer neighborhoods are all more likely to be in violation of CAA requirements. The interpretation of variables linked to detection, however, differs between the models. In the standard probit, where the dependent variable is non-compliance, consistent interpretation of the coefficients relating to detection is more challenging. The results may suggest that firms that are located in poorer and minority neighborhoods within states with Republican governors and larger regional scales are all more likely to be non-compliant. Alternatively, these variables could be driving detection efforts. As a result, firms may appear to be more likely to be violators in the data because they are more likely to have their non-compliance detected under all of the aforementioned conditions. With the standard probit, these two possible interpretations are muddled.

The DCE model by contrast explicitly models compliance and detection separately, enabling a cleaner interpretation of how each set of variables affects detection and non-compliance. A positive coefficient in the compliance portion suggests a greater probability of violation, and a positive coefficient in the detection portion suggests a greater probability of detection. Of course, given that both models are estimated on the same dependent variable, in the absence of strong theory, knowing which equation is modeling compliance and which is modeling detection can also be a challenge (this is referred to as the "labeling problem"). Without strong theoretical guidance on which variables ought to matter for one or the other process, clean interpretation is hindered (Sanford and Smith 2004).

The ability of the DCE estimator to separate these data-generating processes also heightens the prospects of generating quite different inferences from the standard probit. Both the estimated parameters and their associated standard errors vary between the two models. The parameters estimated from a single probit model will be downwardly biased for variables thought to have a positive impact on non-compliance (Feinstein 1990).[31] The origin of this bias lies in the fact that cases coded as compliant in the data used for the standard probit were overpopulated because they include undetected, non-compliant firms. Comparing the coefficients in each model, we see that this is generally the case. Relative to DCE, the standard probit produces downwardly biased estimates of variables linked to non-compliance. The larger coefficients in the DCE model suggest, for example, a less rigorous detection process for firms located in counties with higher unemployment. The factors influencing the variability in detection are reflected by the statistically significant coefficients on the variables in the detection portion of the model (lower half).[32] Relative to the probit, in the DCE model, none of the political variables are statistically significant and, therefore, unlike

[31] Due to the partial observability problem, the DCE estimates are also likely to be less efficient than those assuming full observability (probit) (Poirier 1980).

[32] Depending upon the analyst's interest, the DCE model can extract marginal effects for a given variable on either the marginal success probability for either outcome variable, the joint probability of a desired combination of the two variable outcomes, or the conditional probability of one outcome given the other.

the standard probit and its estimate of the Democratic governor parameter, we would conclude that this factor does not influence the probability of a firm's detection. Also, note that the estimate of Rho (the correlation of the error terms between the two models comprising DCE) is negative but not statistically significant, suggesting that unobserved factors in our model that tend to increase non-compliance are unrelated to those that increase detection (this of course may vary by model choice).

A final feature of the DCE approach to note is that it allows us to estimate several quantities of interest regarding agencies' detection capabilities. With the results in Table 5, we can estimate the average probabilities of detected and undetected violations for the typical case in the data. Our results suggest that the joint probability of detected violating firms is ~5.9% very near the observed 5.8% in our data. However, our model does not assume that all firms appearing in the compliance data as "compliant" are indeed in line with their legal obligations. Rather the model suggests that an estimated 24.3% of the total firms in our data are likely to be undetected violators. Over half of the firms in the data (~53.4%) are expected to be undetected compliant firms, with approximately 16.2% of firms being compliant but detected.

CONCLUSION

Case selection issues are ubiquitous in social science research, particularly in studies making use of observational data. Research in public management is no different in this regard. As public management studies continue to make theoretical and empirical advances, we believe that more attention should be paid to the issues outlined in this article and elsewhere in more formal ways. This is certainly not to suggest that scholars in this field are unaware of how case selection choices can affect causal inferences, and in many respects our review of recent scholarship in *JPART* was encouraging. In many of these studies, scholars not only recognized the myriad concerns raised but also employed a variety of techniques to address these concerns. *JPART* is widely regarded as one of, if not, the leading journal for empirical public management research; so the extent of attention to selection case may reflect an upward bound.

Our purpose in this article was to outline the primary case selection concerns that often arise in public management research and to identify the resulting threats they pose to internal and external validity. The most straightforward resolution to potential case selection problems that arise in a regression context is to condition one's estimates of the treatment effects by all of the observable factors related to treatment. Perhaps due to unobservable factors or perhaps due to the ease with which treatment selection can be overlooked, explicit attempts to control for all observables factors possibly related to treatment and outcome are rare.

In addition to summarizing other standard solutions to instances of selection on either the treatment or outcome variable, we have highlighted three specific methods appropriate to address different types of selection bias in large-n quantitative studies—propensity score matching, Heckman sample selection, and DCE. We focused on these particular methods, even though each has been around for some time, because they have not been widely adopted in public management scholarship. Moreover, each technique is available with standard statistical software, and thus they are user-friendly. This is not to say these methods should be used without caution, as each requires strong distributional and identification assumptions. Moreover, although we illustrated techniques suited for quantitative analysis, case selection concerns are just as strong in qualitative research.

There are many resources available to scholars that provide more technical treatments of case selection concerns, from both a research design (e.g., King, Keohane, and Verba 1994; Morgan and Winship 2007; see Angrist and Pischke 2009 for a particularly accessible treatment) and an econometric perspective (e.g., Achen 1986), and we refer researchers to these and other more detailed presentations of both the problems and fixes to various case selection issues. We conclude by outlining a set of important questions that scholars concerned about these issues can ask themselves when they are interested in drawing causal inferences in their research. First, does the data generation process regarding the dependent variable suffer from any type of selection problem, such that only some values are observed? If the data generation process yielded truncated, censored, partially observed, or no variation in the dependent variable, then causal inferences may be compromised. Second, is there correlation between case selection and treatment condition of interest? This situation often arises in cases where there is non-random selection of cases, and should be acknowledged and if possible addressed to improve the quality of causal inferences. Third, are there any differences in the attributes of cases assigned to the treatment and control condition? Such differences, whether observable or unobservable, can also threaten a researcher's ability to draw valid causal inferences, and they should be recognized and remedied to the extent possible. Careful consideration of these three general questions when crafting research designs, whether qualitative or quantitative (small *n* or large *n*), can greatly improve causal analysis.

REFERENCES

Abowd, John M., and Henry S. Farber. 1982. Job queues and the union status of workers. *Industrial and Labor Relations Review* 35:354–67.

Achen, Christopher. 1986. *The statistical analysis of quasi-experiments*. Berkeley, CA: University of California Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Berry, William D., Evan J. Ringquist, Richard C. Fording, and Russell L. Hanson. 1998. Measuring citizen and government ideology in the American states, 1960–93. *American Journal of Political Science* 42:327–48.

Boyne, George A., and Alex A. Chen. 2007. Performance targets and public service improvements. *Journal of Public Administration Research and Theory* 17:455–77.

Boyne, George A., Oliver James, Peter John, and Nicolai Petrovsky. 2010. Does public service performance affect top management turnover. *Journal of Public Administration Research and Theory* 20:i261–79.

Braumoeller, Bear F., and Gary Goertz. 2000. The methodology of necessary conditions. *American Journal of Political Science* 44:844–58.

Breen, Richard. 1996. *Regression models: Censored, sample-selected, or truncated data*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-111. Thousand Oaks, CA: Sage.

Brehm, John, and James T. Hamilton. 1996. Noncompliance in environmental reporting: Are violators ignorant, or evasive, of the law? *American Journal of Political Science* 40:444–77.

Brooks, Arthur C. 2003. Public goods and posterity: An empirical test of intergenerational altruism. *Journal of Public Administration Research and Theory* 13:165–75.

Cochran, W., and Donald B. Rubin. 1973. Controlling bias in observational studies: a review. *Sankhya*, Ser.A 35:417–46.

Collier, David, and James Mahoney. 1996. Insights and pitfalls: Selection bias in qualitative research. *World Politics* 49:56–91.

Daley, Dorothy M. 2009. Interdisciplinary problems and agency boundaries: Exploring effective cross-agency collaboration. *Journal of Public Administration Research and Theory* 19:477–93.

Dehejia, Rajeev H., and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental casual studies. *Review of Economics and Statistics* 84:151–61.

Diamond, Alexis, and Jasjeet S. Sekhon. Forthcoming. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30:127–46.

Dull, Matthew. 2009. Results-model reform leadership: Questions of credible commitment. *Journal of Public Administration Research and Theory* 19:255–84.

Eckstein, H. 1975. Case studies and theory in political science. In *Handbook of political science*. Political Science: Scope and Theory, vol. 7, eds. F. I. Greenstein and N. W. Polsby, 94–137. Reading, MA: Addison-Wesley.

Feinstein, Jonathan S. 1990. Detection controlled estimation. *Journal of Law and Economics* 33:233–76.

———. 1999. Approaches to estimating noncompliance: Examples from federal taxation in the United States. *Economic Journal* 109:360–9.

Gallo, Nick, and David E. Lewis. 2012. The consequences of presidential patronage for federal agency performance. *Journal of Public Administration Research and Theory* 22:219–43.

Gawande, Kishore, and Hui Li. 2009. Dealing with weak instruments: An application to the protection for sale model. *Political Analysis* 17:236–60.

Geddes, Barbara. 2003. *Paradigms and sand castles: Theory building and research design in comparative politics*. Ann Arbor: University of Michigan Press.

Georgellis, Yannis, Elisabetta Iossa, and Vurain Tabvuma. 2011. Crowding out intrinsic motivation in public sector. *Journal of Public Administration Research and Theory* 21:473–93.

Gordon, Sanford C. 2009. Assessing partisan bias in federal public corruption prosecutions. *American Political Science Review* 103:534–54.

Greene, William H. 2002. *Econometric analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.

Grissom, Jason A., and Lael R. Keiser. 2011. A supervisor like me: Race, representation, and the satisfaction and turnover decisions of public sector employees. *Journal of Policy Analysis and Management* 30:557–80.

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey methodology*. Hoboken, NJ: John Wiley and Sons.

Hanushek, Eric A. and Ludger Wößmann. 2006. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal* 116:C63–76.

Heckman, James. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61.

Heckman, James, and Petra Todd. 2009. A note on adapting propensity score matching and selection models to choice based samples. *Econometric Journal* 12:230–4.

Heckman, James J., and Salvador Navarro-Lozano. 2004. Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *Review of Economics and Statistics* 86:30–57.

Heinrich, Carolyn J., Robert H. Meyer, and Greg Whitten. 2010. Supplemental Education Services Under No Child Left Behind: Who Signs Up, and What Do They Gain? *Educational Evaluation and Policy Analysis* 32:273–98

Heinrich, Carolyn J. 2010. Third-party governance under no child left behind: Accountability and performance management challenges. *Journal of Public Administration Research and Theory* 20:i59–80.

Helland, Eric. 1998a. Environmental protection in the federalist system: The political economy of NPDES inspections. *Economic Inquiry* 36:305–19.

———. 1998b. The enforcement of pollution control laws: Inspections, violation, and self-reporting. *Review of Economics and Statistics* 80:141–53.

Hidalgo, F. Daniel, and Jasjeet Sekhon. 2011. In *International encyclopedia of political science*, eds. Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino, 203–10. Thousand Oaks, CA: Sage Publications.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945–60.

Imai, Kosuke, Luke Keele, Dustin Tingely, and Teppei Yamamoto. 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105:765–89.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* 25:51–71.

James, Oliver. 2009. Evaluating the expectations disconfirmation and expectations anchoring approaches to citizen satisfaction with local public services. *Journal of Public Administration Research and Theory* 19:107–23.

John, Peter, and Hugh Ward. 2005. How competitive is competitive bidding: The case of the single regeneration budget program. *Journal of Public Administration Research and Theory* 15:71–87.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton: Princeton University Press.

Konisky, David M. 2009. Inequities in enforcement? Environmental justice and government performance. *Journal of Policy Analysis and Management* 28:102–21.

Konisky, David M., and Christopher Reenock. Forthcoming. Examining sources of regulatory compliance bias in policy implementation. *Journal of Politics*.

Kotchen, Matthew J., and Michael R. Moore. 2007. Private provision of environmental public goods: Household participation in green-electricity programs. *Journal of Environmental Economics and Management* 53:1–16.

Lavertu, Stéphane, and David L. Weimer. 2011. Federal advisory committees, policy expertise, and the approval of drugs and medical devices at the FDA. *Journal of Public Administration Research and Theory* 21:211–37.

Long, Scott J. 1997. *Regression models for categorical and limited dependent variables*. Advanced Quantitative Techniques in the Social Sciences Number 7. Thousand Oaks, CA: Sage Publications.

Miller, Susan, and Jill Nicholson-Crotty. 2011. Bureaucratic effectiveness and influence in the legislature. *Journal of Public Administration Research and Theory* Advanced Access version, August 22.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

Moynihan, Donald P., and Noel Landuyt. 2008. Explaining turnover intention in state government: Examining the roles of gender, life cycle, and loyalty. *Review of Public Personnel Administration* 28:120–43.

Murray, Michael P. 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20:111–32.

Naper, Linn Renee. 2010. Teaching hiring practices and educational efficiency. *Economics of Education Review* 29:658–68.

Olson, Mary K. 1995. Regulatory agency discretion among competing industries: Inside the FDA. *Journal of Law, Economics, and Organization* 11:379–407.

Pandey, Sanjay, and Stuart I. Brettschneider. 1997. The impact of red tape's administrative delay on public organizations' interest in new information technologies. *Journal of Public Administration Research and Theory* 7:113–30.

Poirier, Dale J. 1980. Partial observability in bivariate probit models. *Journal of Econometrics* 12:209–17.

Reenock, Christopher M., and Brian J. Gerber. 2008. Political Insulation, Information Exchange, and Interest Group Access to the Bureaucracy. *Journal of Public Administration Research and Theory* 18:415–40.

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.

———. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79:516–24.

Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74, 318–28.

———.1980. Bias reduction using mahalanobis metric matching. *Biometrics* 36, 293–98.

Sanford, Gordon C., and Alastair Smith. 2004. Quantitative leverage through qualitative knowledge: Augmenting the statistical analysis of complex causes. *Political Analysis* 12:223–55.

Sartori, Anne E. 2003. An estimator for some binary-outcome selection models without exclusion restrictions. *Political Analysis* 11:111–38.

Scholz, John T., and Cheng-Lung Wang. 2006. Cooptation or transformation? Local policy networks and federal regulatory enforcement. *American Journal of Political Science* 50:81–97.

Sekhon, Jasjeet S. 2008. The Neyman-Rubin model of causal inference and estimation via matching methods. In *The Oxford handbook of political methodology*, eds. Janet Box-Steffensmeier, Henry Brady, and David Collier, 271–99. New York: Oxford University Press.

Serra, George. 1995. Citizen-initiated contact and satisfaction with bureaucracy: A multivariate analysis. *Journal of Public Administration Research and Theory* 2:175–88.

Tobin, James. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25:65–86.

Winship, Christopher, and Stephen L. Morgan. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25:659–707.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.