2021

A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example

John DiNardo
University of Michigan

Jordan Matsudaira
Columbia University

Justin McCrary
Columbia Law School, jmccrary@law.columbia.edu

Lisa Sanbonmatsu
Harvard University

A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example

John DiNardo, *University of Michigan and National Bureau of Economic Research (NBER)*

Jordan Matsudaira, *Columbia University*

Justin McCrary, *NBER*

Lisa Sanbonmatsu, *Harvard University*

Survey nonresponse and attrition undermine the validity of many and possibly most econometric estimates. We propose that survey administrators and evaluators proactively create an instrument for observation, for example, by ex ante randomizing participants to differing intensity of follow-up. We illustrate how to apply our proposed methodology using a carefully conducted randomized controlled trial, the Moving to Opportunity demonstration project, which de facto randomly assigned a subset of subjects to more intensive follow-up. The approach yields treatment effect estimates similar to the unbiased estimator based on complete administrative data and has narrower confidence intervals than alternative bounding approaches.

Thanks to Jane Garrison for excellent research assistance; to Jeff Kling, without whom the empirical example would not have been possible; and to David Lee for ideas and suggestions that substantially improved the paper. Contact the corresponding author, Jordan Matsudaira, at matsudaira@tc.columbia.edu. Information concerning access to the data used in this paper is available as supplemental material online.

[*Journal of Labor Economics*, 2021, vol. 39, no. S2]
 0734-306X/2021/39S2-0014$10.00
Submitted May 21, 2019; Accepted November 9, 2020

I. Introduction

The problems of attrition, survey nonresponse, sample selection, and missing data more generally are the subject of large literatures in econometrics, biostatistics, survey research, and other subjects. It is widely appreciated that these problems can undermine the validity of the simplest inferential problems, such as estimating the rate of unemployment in a population, as well as more difficult problems.

Problems with missing data, as old as surveys themselves, may be worsening. Consider, for example, the case of the Current Population Survey (CPS), the source of much of what is known about trends in employment and wages in the United States. Prior to 1994, when the survey was redesigned in various ways, household nonresponse rates were near 5%. During the 1990s nonresponse rates rose to more than 6% (Bureau of Labor Statistics and Census Bureau 2002), and in recent years they have risen to roughly 18% (Rothbaum and Bee 2020). Household nonresponse understates the extent of missing data in the CPS, however. Even when households are interviewed, key variables of interest may be missing. Income nonresponse rates in the CPS in the 1990s were roughly 15% (Bureau of Labor Statistics and Census Bureau 2002) and in more recent years have risen to around 30% (Hirsch and Schumacher 2004). Nonresponse due to attrition is also common in evaluations. A recent review of 91 recent field experiments published in top economics journals found an average attrition rate of 15% (Ghanem, Hirshleifer, and Ortiz-Becerra 2019).[1]

As is widely appreciated, missing data pose no econometric difficulty if they are randomly missing. However, the characteristics of those who are missing certain information may be different from the characteristics of those who are not. Careful matched records analysis linking household surveys to the population census (Grovers and Couper 1998; Bollinger et al. 2019; Rothbaum and Bee 2020) substantiate such concerns with nonresponse bias. Importantly, the sign of nonresponse bias is highly context specific and hence difficult to predict: the bias is different for different subpopulations and additionally depends on the type of nonresponse (e.g., respondent could not be located, respondent refused to be interviewed, respondent refused to respond to item). In sum, missing data is an increasingly important problem for empirical research and leads to biases of ambiguous sign.

The first-best strategy for missing data is to collect information on all items for all sampled units. Only rarely is this strategy feasible. Consequently, a variety of approaches are considered in the literature, including modeling the process determining which data are missing (Heckman 1976, 1979; Horowitz and Manski 1998), bounding the parameter of interest (Manski 1989, 1990,

[1] This may understate the problem to the extent that top journals may use attrition problems as a screen.

1994, 1995; Lee 2005), or assuming that nonresponse is ignorable (Rubin 1987).

The most extensive of these three literatures pertains to modeling the selection process. One of the central ideas emerging from this literature is that partial randomization of the probability of observation is a key ingredient for correcting sample selection bias (e.g., Das, Newey, and Vella 2003).[2] Economists routinely devise credible instruments for endogenous regressors in the simultaneous equations model. However, devising instruments for the probability of observation is a more challenging problem. To the best of our knowledge, this is the first paper to use an instrument for the probability of observation generated by actual random assignment.[3]

In this paper, we argue that circumventing attrition may best be facilitated by a proactive approach in which partial randomization of the probability of observation is built into data collection procedures. For example, as part of a survey design, one might randomize half of sampled units to be subject to more intensive follow-up than the other half.

That such a procedure would be useful for estimating sample selection bias and estimating population parameters is implicit in much of the sample selection literature. Nonetheless, such data collection procedures are not standard practice. The efforts and insights of the sample selection literature notwithstanding, economists involved in data collection efforts typically do not recommend procedures to generate partial randomization of the probability of observation. Instead, nonresponse is recognized ex post, at which point it may be too late to prevent inferences from being compromised by that nonresponse. Aside from managerial difficulties, procedures involving partial randomization of the probability of observation are not expected to be any more expensive than standard procedures. Against this backdrop of small costs, partial randomization provides the important benefit of information on the nature of the selection problem.[4]

[2] We use the term "partial randomization" throughout the paper. By this, we mean simply that there be a variable known to influence the probability of observation that is excludable from the outcome equation.

[3] In large nationally representative samples, the procedures for dealing with nonresponse include hot decking (assignment of some individual's completed response to nonrespondents) and sample weight adjustments. These procedures may be valid if nonresponse is idiosyncratic, conditional on the variables used for imputation. However, this is rarely credible. Furthermore, as emphasized by Horowitz and Manski (1998), analyses conducted with sample weight adjustments may yield estimates of parameters that are not logically possible. Attrition is particularly an issue in medical randomized controlled trials (Juni, Altman, and Egger 2001). A major, somewhat successful recent initiative has been to encourage researchers to report when attrition or nonresponse has occurred (Altman et al. 2001; Moher, Jones, and Lepage 2001).

[4] Persuading survey administrators of the value of sample selection correction estimators may be an important impediment, and our econometric recommendations

We demonstrate how the availability of a credible instrument can be used to identify causal estimands of interest in the presence of missing data under different assumptions about the nature of selection into observation. We situate this approach to data collection within the traditional econometric sample selection framework (e.g., Heckman 1979; Ahn and Powell 1993; Das, Newey, and Vella 2003). We use a simple graphical interpretation of the Heckman two-step estimator ("Heckit") that clarifies the nature of identification in the sample selection model, suggests a joint test of the functional form and distributional assumptions associated with traditional parametric sample selection correction techniques, and shows how an alteration to the estimand of interest may be of particular interest in this context.

We provide a concrete implementation of the approach described using a real data set on adult outcomes in the Moving to Opportunity (MTO) experiment. We take advantage of a fortuitous follow-up strategy used by the MTO team. As part of follow-up survey procedures, the team de facto randomized 30% of participants to be followed up more intensively than others.[5] Because the data collected by the MTO project are based on both administrative and survey sources, we are able to compare (1) usually infeasible estimators, based on responses from nearly all respondents, to (2) estimators that assume missing data occurs at random and to (3) sample selection correction estimators that take advantage of the MTO survey design, where random assignment to a more intense follow-up group can be used as an instrument for the probability of observation. We also use the data to construct bounds on the effect of MTO treatments on adult outcomes.

In the case of the MTO, our Heckit estimator based on randomization to more intense follow-up yields estimates of treatment effects that are similar to the (unbiased) estimates based on the complete administrative data. Differential attrition in the MTO was very slight, so further work should assess how well our approach mitigates sample selection bias in cases where more is expected. The confidence intervals are narrower than the inferences resulting from the bounding approaches encountered in the literature.

The remainder of the paper is organized as follows. In section II we present a conceptual framework to illustrate the bias created by nonresponse. Section III illustrates how partial randomization of the probability of observation can be used to identify causal effects in the presence of nonrandom nonresponse, describes our proposal, and reviews different approaches to problems caused by attrition. Section IV gives background information on the MTO experiment, and section V presents our results. Finally, section VI concludes and discusses promising areas for future work.

---

are ultimately importantly guided by a pragmatic approach in which constraints of survey administration are taken seriously.

[5] This information has not been explicitly exploited by researchers evaluating MTO.

II. Conceptual Framework and Statement of the Problem

We start by considering a simple model with randomized treatment and nonrandom nonresponse.[6] In such a model, the average treatment effect would be identified by the difference in observed means by treatment status, were it not for nonresponse. Formally,

$$Y_1^* = \mu_1 + U_1, \tag{1}$$

$$Y_0^* = \mu_0 + U_0, \tag{2}$$

$$Y^* = TY_1^* + (1 - T)Y_0^*. \tag{3}$$

As usual, the counterfactual pair (Y_1^*, Y_0^*) is unobserved. Here, we focus on sample selection, as opposed to endogeneity of treatment.[7] That is, we are concerned with the fact that we only sometimes observe Y^*. More precisely, we observe $Y = Y^*$ when $S = 1$, but otherwise Y is missing, where S is an observation indicator

$$S = \mathbf{1}(\alpha + T\delta_0 - V \geq 0). \tag{4}$$

The trio of errors U_1, U_2, and V are assumed independent of T and mean zero, so that $E[Y_1^*] = \mu_1$, $E[Y_0^*] = \mu_0$, and $P(S = 1|T) = F_V(\alpha + T\delta_0)$, where $F_V(\cdot)$ is the distribution function for V.

In this framework, the population average treatment effect is given by

$$E[Y_1^* - Y_0^*] = \mu_1 - \mu_0. \tag{5}$$

However, this estimand will be challenging to identify. To see why, note that average observed outcomes for treatment and control identify the conditional expectations

$$\begin{aligned} E[Y|T = 1, S = 1] &= E[Y_1^*|T = 1, V \leq \alpha + \delta_0] \\ &= \mu_1 + E[U_1|V \leq \alpha + \delta_0] \end{aligned} \tag{6}$$

and

$$\begin{aligned} E[Y|T = 0, S = 1] &= E[Y_0^*|T = 0, V \leq \alpha + \delta_0] \\ &= \mu_0 + E[U_0|V \leq \alpha], \end{aligned} \tag{7}$$

[6] We thank David Lee for suggesting the more general framework we use here compared with that presented in our working paper, DiNardo, McCrary, and Sanbonmatsu (2007).

[7] In the interest of a simplified discussion, we also abstract from covariates. These can be included by modifying the outcome and selection equations in the obvious ways.

respectively, so that the difference in observed means identifies

$$E[Y|T = 1, S = 1] - E[Y|T = 0, S = 1] = \mu_1 - \mu_0 + E[U_1|V \leq \alpha + \delta_0] - E[U_0|V \leq \alpha]. \tag{8}$$

Here, the term $E[U_1|V \leq \alpha + \delta_0] - E[U_0|V \leq \alpha]$ represents the bias due to nonresponse. The bias arises because in general the population with $V \leq \alpha + \delta_0$ will differ from the population with $V \leq \alpha$. In special cases such as $\delta_0 = 0$ (when the two populations coincide) or if V is independent of U_1 and U_0, the nonresponse bias term will be zero, but in general nonresponse bias is expected.

In the literature on the estimation of treatment effects, it is by now commonplace to think of the average treatment effect for subpopulations (e.g., Imbens 2004; Crump et al. 2009). In the context of sample selection, it is helpful to consider average treatment effects for those observed under treatment and for those observed under control. In the framework above, these two estimands would correspond to

$$E[Y_1^* - Y_0^*|V \leq \alpha + \delta_0] = \mu_1 - \mu_0 + E[U_1|V \leq \alpha + \delta_0] - E[U_0|V \leq \alpha + \delta_0]$$

and

$$E[Y_1^* - Y_0^*|V \leq \alpha] = \mu_1 - \mu_0 + E[U_1|V \leq \alpha] - E[U_0|V \leq \alpha],$$

respectively. In general, these will not be identifiable unless $\delta_0 = 0$ because under treatment we observe the population with $V \leq \alpha + \delta_0$ and under control we observe the population with $V \leq \alpha$, but we never observe the same population under treatment and control conditions.

The populations with $V \leq \alpha + \delta_0$ and $V \leq \alpha$ can be nested in a continuum of populations with $V \leq k$. A valid lifestyle choice is thus to generalize the notion of potential estimands of interest beyond just the average treatment effect to

$$E[Y_1^* - Y_0^*|V \leq k] = \mu_1 - \mu_0 + E[U_1|V \leq k] - E[U_0|V \leq k]. \tag{9}$$

As k increases, this estimand converges to the population average treatment effect. For values of k such that $F_V(k)$ is close to the empirical probability of observation, we might think of the estimand as the gettable average treatment effect (GATE) because it focuses on the (sub)population that can be observed. In this context, it is worth emphasizing that even with no differential attrition ($\delta_0 = 0$), the GATE may differ from the population treatment effect in the presence of treatment effect heterogeneity.[8]

[8] As we describe below, going from a GATE estimand to a population average treatment effect estimand likely will require modeling assumptions.

Before proceeding, we want to emphasize that a GATE parameter is already familiar. It is the parameter that, under mild assumptions, is identified from an experiment where there is nontrivial attrition that is, however, similar between treatment and control (Lee 2009).

III. Strategies for Addressing Sample Selection Bias

For researchers unwilling to assume that data are missing completely at random, there are three types of approaches: (i) assume that the data are missing at random conditional on covariates, (ii) assume that there is an instrument for the probability of observation and use sample selection correction techniques, or (iii) resort to bounding treatment effects. We do not discuss the first type of approach. We focus instead on selection correction techniques utilizing a truly randomized instrument for the probability of observation as well as approaches aimed at bounding treatment effects.

A. Selection Correction Approaches

We now extend the model outlined in section II to illustrate how partial randomization of the probability of observation can be used to identify the treatment effects of interest. For reasons of practicality, we limit ourselves to the case of a binary instrument Z that generates two points of support $P(S = 1|T = t, Z = z)$ for $z \in \{0, 1\}$ for each treatment state t. While it is possible to conceive of grander ambitions whereby values of the instrument range widely, inducing many points of support for the probability of observation, an obvious starting point is assessing what can be done with two points of support.[9] We leave extensions to future research. Foreshadowing the discussion below, we think of Z as the level of hassling (i.e., the intensity of follow-up) applied to participants to elicit responses for Y in an outcome survey.[10]

Let $Z = 1$ denote high intensity of effort at follow-up and $Z = 0$ denote low intensity, and modify equation (4) to reflect the impact of hassling,

$$S = \mathbf{1}(\alpha + T\delta_0 + Z\delta_1 + TZ\delta_2 - V \geq 0), \tag{10}$$

with corresponding probability of observation $P(S = 1|T, Z) = F_V(\alpha + T\delta_0 + Z\delta_1 + TZ\delta_2)$.

Average observed outcomes for treatment and control by hassling identify four conditional expectations:

[9] We note that in practice randomizing follow-up effort will involve a certain amount of persuasion of survey firms and workers. This suggests the wisdom of starting with somewhat limited aspirations.

[10] Our analysis thus parallels recent developments in the instrumental variable literature, such as Brinch, Mogstad, and Wiswall (2017), in particular their sec. III.

$$
\begin{aligned}
E[Y|T = 1, Z = 1, S = 1] &= \mu_1 + E[U_1|V \leq \alpha + \delta_0 + \delta_1 + \delta_2], \\
E[Y|T = 1, Z = 0, S = 1] &= \mu_1 + E[U_1|V \leq \alpha + \delta_0], \\
E[Y|T = 0, Z = 1, S = 1] &= \mu_0 + E[U_0|V \leq \alpha + \delta_1], \\
E[Y|T = 0, Z = 0, S = 1] &= \mu_0 + E[U_0|V \leq \alpha].
\end{aligned} \tag{11}
$$

In general these population moments correspond to different subgroups of the population, each with a different probability of being observed and thus a different degree of selection. When Z is randomized, we can compare groups of the same treatment status but different amounts of hassling to assess whether the observed data are positively or negatively selected in the treatment and control groups, respectively—for example, are average earnings higher or lower when the probability of observation is higher—and by how much.

In special cases, the average treatment effect for subpopulations may be identified from the data. For example, if $\delta_0 = \delta_2 = 0$, then treatment does not affect the probability of observation. In this case, a simple comparison of average outcomes by treatment status for different values of the instrument identifies a GATE for two subpopulations. The subgroups exposed to different levels of hassling identify two different GATE parameters. For $Z = 1$ the difference in means by treatment status identifies $E[Y_1^* - Y_0^*|V \leq \alpha + \delta_1]$, while for $Z = 0$ it identifies $E[Y_1^* - Y_0^*|V \leq \alpha]$. The former is the average treatment effect for the subpopulation that regardless of treatment status would respond to the survey if hassled. The latter is the average treatment effect for the subpopulation that responds to the survey under any configuration of treatment without hassling, or what Lee (2009) refers to as the "always observed." The difference in these GATEs depends on the extent to which treatment effect heterogeneity is related to the probability of observation. We return to this below.

The two GATE parameters just described may also be identified in other special cases. For example, it could occur that $\delta_0 = 0$ but that $\delta_2 \neq 0$, in which case the estimated treatment effect for the subsample $Z = 0$ would identify $E[Y_1^* - Y_0^*|V \leq \alpha]$. It could also occur that the sum $\delta_0 + \delta_2$ is zero, in which case the estimated treatment effect for the subsample with $Z = 1$ would identify $E[Y_1^* - Y_0^*|V \leq \alpha + \delta_1]$.[11]

[11] There are also two other obvious possibilities under which a GATE is identified. These occur when $\delta_1 - \delta_0 = 0$, on the one hand, and when $\delta_0 + \delta_1 + \delta_2 = 0$, on the other. For these situations, average observed outcomes by treatment status are compared with hassling switched on and off and identify $E[Y_1^* - Y_0^*|V \leq \alpha + \delta_1] = E[Y_1^* - Y_0^*|V \leq \alpha + \delta_0]$ and $E[Y_1^* - Y_0^*|V \leq \alpha] = E[Y_1^* - Y_0^*|V \leq \alpha + \delta_0 + \delta_1 + \delta_2]$, respectively.

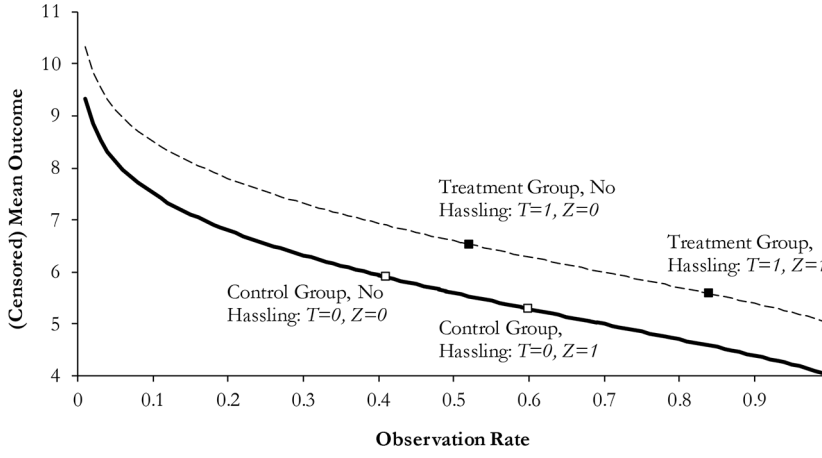

FIG. 1.—Estimating treatment effects in the presence of sample selection with binary hassling. See the discussion in section III.A for details.

These special cases correspond to identification of subpopulation-specific average treatment effects.[12] In general, however, these special cases may not hold, in which case we cannot identify any GATEs (indexed by k in eq. [9]) in the absence of functional form restrictions.[13] As noted by several authors (e.g., see Brinch, Mogstad, and Wiswall 2017; Kline and Walters 2019), in the absence of many points of support of $P(S = 1|Z = z)$, identification may be conferred by a functional form assumption.

Figure 1 presents some hypothetical data to provide graphical intuition for how functional form assumptions facilitate identification. First, let us describe the figure. For $t, z \in \{0, 1\}$, let $\tilde{Y}_{tz}$ denote the sample analogue of $E[Y|T = t, Z = z, S = 1]$, and let $\bar{S}_{tz}$ denote the sample analogue of $E[S|T = t, Z = z]$. We graph the four pairs $(\tilde{Y}_{tz}, \bar{S}_{tz})$, along with the two curves depicting the censored means of Y_t for $T = t$, as a function of the probability of observation. The figure assumes that hassling increases the probability of being observed and that those readily observed are positively selected. That is, hassling shifts the probability of observation to the right, and the censored mean outcome curves both slope down.

In the figure, the vertical differences between the censored mean functions at any given probability of observation all represent an average treatment

[12] Extrapolating to the population average treatment effect is possible with functional form assumptions, but that extrapolation becomes harder to justify the greater the extent of nonresponse is.

[13] This arises because, just as in the simple case discussed in sec. II, the populations $V \leq \alpha$ and $V \leq \alpha + \delta_0$ differ, as do the populations $V \leq \alpha + \delta_1$ and $V \leq \alpha + \delta_0 + \delta_1 + \delta_2$—and for that matter the populations $V \leq \alpha$ and $V \leq \alpha + \delta_0 + \delta_1 + \delta_2$.

effect for a subpopulation (a GATE as in eq. [9] for some k). Sample selection bias arises in the figure because we do not observe treatment and control observations at the same probability—if we did, as in the special cases outlined above, a GATE would be identified. While outside of the cases we consider here, we note that a continuous Z might be expected to generate a broad region of observation probabilities where GATEs can be identified (e.g., between $F_V(\alpha)$ and 1).

Clearly, if the special cases above do not obtain and with limited observed points of support for the censored treatment and control mean functions, we need to know the shape of the these functions in order to extrapolate and estimate the difference between them at a common observation probability. With only two points of support from a binary Z, identification will require assuming that the censored mean functions are linear in the probability of selection, or some transformation of it, so that the slope of the functions can be identified by comparing observed average outcomes for the same treatment status between $Z = 1$ and $Z = 0$.

The functions depicted in figure 1 are parallel, implicitly assuming that treatment effect heterogeneity is not related to the probability of observation. Under that assumption—implicit in the standard Heckit procedure since it is a constant coefficients model—the question of "which GATE?" or "at which probability of observation treatment effects should be estimated?" is moot. However, the econometrician can allow for treatment effect heterogeneity by allowing the slopes of the censored mean functions to differ, in which case a key question becomes, How far away from the empirical probabilities of observation should one extrapolate in order to estimate a GATE? In figure 1, for example, the econometrician might feel comfortable assuming linearity of the censored mean functions only "in between the dots" and therefore attempt to estimate GATE for some selection probability between $\bar{S}_{10}$ and $\bar{S}_{01}$.

We can of course extrapolate further from the observed probabilities of observation to identify the population average treatment effect, but only if the assumptions made about functional form are in fact correct. To see why such assumptions might be hard to defend in specific applications, let us first review the connection between distributional assumptions and the functional forms that they generate.

In figure 1, we assumed trivariate normal errors on counterfactual outcomes and an observation equation, assuming equal correlation coefficients between the two counterfactual outcome errors and the observation error. In figure 2, we transform the horizontal axis to inverse Mills ratio space. By so doing, the curves from figure 1 are now lines, and their parallel nature is clear visually. However, the potential fragility of conclusions based on extrapolation is highlighted in figure 3. Here, the error in the observation equation is taken to be Laplacian rather than normal, but the unwitting economist continues to use the inverse Mills ratio. "Connecting" the dots

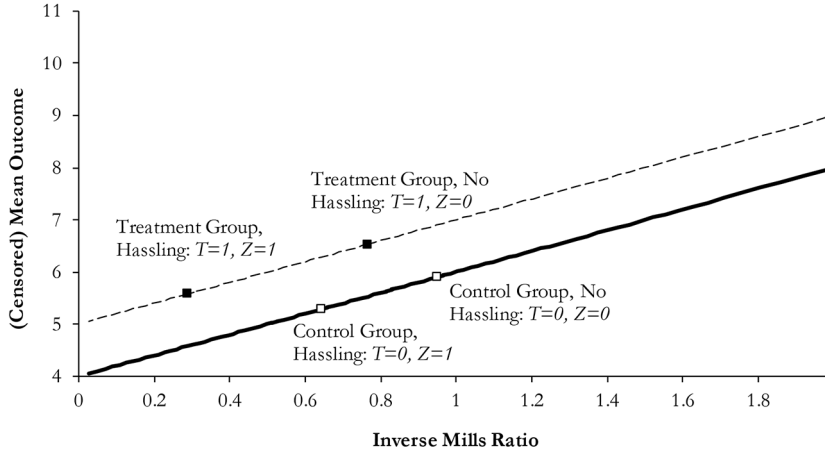

FIG. 2.—Transformations to induce linearity: observation equation error distributed normal. See the discussion in section III.A for details.

and extrapolating to estimate the difference between the two functions at an inverse Mills ratio of 0 (i.e., an observation probability of 1) will clearly result in a biased estimate of the population treatment effect. In general this will be true regardless of whether we impose equal slopes, but in this case the bias will be greater if we allow the slopes to differ.[14] The figure makes plain that assuming a specific functional form is somewhat fraught: even if it were true that there was some space in which the censored means by treatment status were linear and parallel, it does not follow that they will be linear and parallel in the space that the economist is using. On the other hand, the figure also makes plain that if the probabilities of observation are not too different between treatment and control, (*a*) the traditional parallel lines assumption can be relaxed, (*b*) linear approximation may be a good approximation for a specific functional form, and (*c*) linear approximation may be robust across a wide variety of candidate functional forms.

The challenge posed by nonresponse could theoretically be solved by having many points of support for Z, but that may not be feasible. We again proceed with limited ambitions and assume that only two points of support are possible. We assume a parsimonious relationship between the counterfactual errors (U_1, U_0) and the observation equation error V, given by

$$U_t = \rho_t V + \breve{U}_t, \tag{A1}$$

[14] See Heckman, Tobias, and Vytlacil (2000). For generating the data in the display, we continue to draw the counterfactual outcome residuals from the bivariate normal distribution. This leads to a control function of the form $\phi(\Phi^{-1}(F(\Phi^{-1}(s))))/F(\Phi^{-1}(s))$, where $F(\cdot)$ is the Laplacian distribution function.

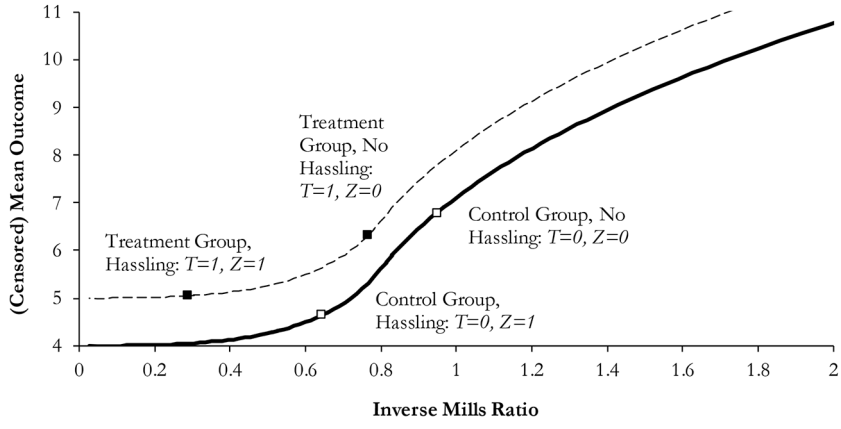

FIG. 3.—Transformations to induce linearity: observation equation error assumed normal but actually distributed as double exponential (Laplacian). See the discussion in section III.A for details.

where $E[\breve{U}_1|V] = E[\breve{U}_0|V] = 0$.[15] We label this equation (A1) because it is an assumption that plays a particularly important role in the context of sample selection corrections, and we will invoke it throughout the rest of the paper. Note in particular that this assumption means that the only scope for the V to be correlated with the gains to treatment—namely, $U_1 - U_0$—is through differences between ρ_1 and ρ_0.[16] This is a strong assumption and may not be appropriate in every application.

Under assumption (A1), we can express the observed moments from equation (11) as

$$\begin{aligned}
E[Y|T = 1, Z = 1, S = 1] &= \mu_1 + \rho_1 E[V|V \leq \alpha + \delta_0 + \delta_1 + \delta_2],\\
E[Y|T = 1, Z = 0, S = 1] &= \mu_1 + \rho_1 E[V|V \leq \alpha + \delta_0],\\
E[Y|T = 0, Z = 1, S = 1] &= \mu_0 + \rho_0 E[V|V \leq \alpha + \delta_1],\\
E[Y|T = 0, Z = 0, S = 1] &= \mu_0 + \rho_0 E[V|V \leq \alpha].
\end{aligned} \tag{12}$$

From these equations we see the power of assumption (A1) in conjunction with a distributional assumption for V. Assumption (A1) reduces the sample selection problem to a linear function of one underlying variable and the distributional assumption furnishes the censored mean function $E[V|V \leq k]$, which in turn provides the appropriate transformation to make the observed counterfactual outcome equations linear.

[15] This condition is satisfied if we assume joint normality of U_1, U_0, and V, but joint normality is not necessary. It is, however, a strong assumption that reduces the dependence between U_t and V to ρ_t.

[16] That is, under eq. (A1), the gains to treatment are given by $U_1 - U_0 = (\rho_1 - \rho_0)V + \breve{U}_1 - \breve{U}_0$.

The most familiar choice for the distribution of V is normality, in which case the conditional mean is the inverse Mills ratio—that is, $E[V|V \leq k] = -\phi(\Phi^{-1}(s))/s$, where $s = \Phi(k)$ is the probability of observation. Other options explored in the literature include uniform (Olsen 1980) and logistic (Mroz 1987) distributions for V (for a recent summary, see Kline and Walters 2019). We leave to future research the question of identifying the class of distributions F_V over which it may be reasonable to assume $E[V|V \leq k]$ is a (nearly linear) function over relevant ranges of observation probabilities.

A common assumption in the sample selection literature is that $\rho_1 = \rho_0$ in the above system. Coupled with normality of V, this leads to a familiar Heckit (see, e.g., Lee 2009, eq. [1]). As noted above, researchers may want to relax this model if they strongly expect treatment gains to be correlated with observation, or they may want to verify whether this assumption is supported by the data. The framework above shows that assuming $\rho_1 = \rho_0$ leads to an overidentified model, allowing an omnibus test for this assumption and the distributional assumption for V. We develop this point below.

B. Binary Treatments and Binary Hassling

In empirical practice, economists often are in the position of discovering an attrition problem after outcome data are collected and making ex post arguments about why some variables already collected might meet the assumptions for Z necessary for identification. Such variables, however, are unlikely to be truly exogenous and may therefore fail to reduce bias in estimated treatment effects. Our proposal instead is for researchers to incorporate proactive strategies into their study protocols aimed at generating partial random assignment of the probability of observation.[17]

To keep the discussion as simple as possible, we continue to assume that the treatment indicator, T, has been unconditionally randomly assigned. In that case, a simple comparison of means—in the absence of sample selection considerations—would identify the treatment effect of interest.

Consider a stylized description of standard data collection practices in this type of evaluation strategy:

1. Randomize individuals at baseline into $T = 1$ or $T = 0$.
2. Collect data on the outcome for as many persons as possible.
3. Compare the differences in means for the observed data.

[17] As noted above, a continuous Z would be ideal. However, we limit ourselves to the case of binary Z. Implicit in this more modest aim is the recognition that (i) generating an appropriate Z requires alteration of data collection procedures, (ii) the increase in administrative burden is rapidly increasing in the support of Z, and (iii) until economists persuade data collection administrators of the value of sample selection correction, obtaining a binary Z will remain an ambitious goal practically, if not econometrically.

In practice, obtaining responses without randomization of the probability of observation already involves differing levels of effort. In the typical survey collection situation, these levels of effort are a black box from the econometric perspective. Given the lack of econometric input into the process, it is unsurprising that survey effort is not measured and thus cannot be modeled econometrically. In a survey without randomization of effort, step 2 might involve different methods of making contact, such as sending an email, calling on the telephone, or making a home visit, possibly multiple times.

The core of our pitch is for economists worried about sample selection to involve themselves more extensively in survey design. This should allow for existing survey efforts to be directed in an econometrically appropriate manner that allows for inferences that are more robust to nonresponse. For example, suppose a potential respondent is randomized into one of two groups: the "intense effort to interview subject" group and the "less intense effort to interview subject" group, corresponding to $Z = 1$ and $Z = 0$, respectively. For those in the former group the survey firm might be instructed to make as many as five phone calls in an attempt at reaching a participant, and for those in the latter group the survey firm might be instructed to make only one phone call.

Our proposal is thus to modify standard practice in the following way:

1. Randomize individuals at baseline into $T = 1$ or $T = 0$.
2. For each group ($T = 1$ and $T = 0$), randomize individuals into two subgroups ($Z = 1$ and $Z = 0$).
3. To collect the necessary outcome data, employ intense effort for the $Z = 1$ subgroup and less intense effort for the $Z = 0$ subgroup.
4. Using the information on T and Z, leverage the sample selection literature to estimate treatment effects addressing nonresponse head-on.

For example, in step 4 one might run a regression of the outcome on T and the inverse Mills ratio based on a first-step probit with covariates T and Z and possibly their interaction. This is the two-step estimator for micro data (Heckman 1976) and is asymptotically equivalent to the maximum likelihood estimator that assumes joint normality.

The standard Heckit approach is likely the most common approach encountered in the literature. The randomization of Z will increase the credibility of the Heckit approach since it requires a covariate that predicts the probability of observation that can be excluded from the outcome equation—and as noted, it is hard to see how such a covariate will fit the bill unless randomization is built into the survey design. However, randomizing Z and applying the Heckit framework is not a panacea for all of the challenges posed by missing data. The Heckit framework, in addition to assuming the existence of an appropriate covariate, such as a randomized Z, makes several other assumptions: it assumes that assumption (A1) holds, that V is distributed normally,

and that the censored means by treatment status are parallel in inverse Mills' ratio space—that is, $\rho_1 = \rho_0$ in the notation above.

Randomizing Z as part of a survey can prove useful, however, even if one does not invoke the standard Heckit framework. For example and as alluded to in section III.A, one could consider estimands such as GATE parameters near observed probabilities of observation. The central point of equation (9) is that any "vertical slice" of figure 1 would be sufficient for identifying a treatment effect. The core challenge is that the probability of observation will not, except in fortunate circumstances, be exactly equal for different treatment arms. If the probabilities of observation differ, it may be possible to linearly interpolate using the variation provided by Z to a common point of support k. Such an approach does not require an assumption of parallel lines but does require adopting a GATE parameter as the estimand of interest.[18] However, extrapolating far beyond the range of the data may lead to fragile estimates.

Another approach to the identification challenge presented by missing data is to impose structure on the problem like in the traditional Heckit framework but to see whether the data reject the modeling assumptions. This can be done using standard overidentification tests and a minimum distance framework. In the Heckit framework, we are imposing the restriction $\rho_1 = \rho_0$. This means that the population model is

$$E[Y|T = t, Z = z, S = 1] = a + bT + cg(E[S|T = t, Z = z]) \quad (13)$$

for parameters (a, b, c), where $g(\cdot)$ is the inverse Mills ratio $g(s) = \phi(\Phi^{-1}(s))/s$.[19] This model has three parameters but the data identify four versions of it, corresponding to $(t, z) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. As before, let $\tilde{Y}_{tz}$ and $\bar{S}_{tz}$ denote sample analogues to the population moments $E[Y|T = t, Z = z]$ and $E[S|T = t, Z = z]$, respectively. Consider the four-vector $\hat{f} = \hat{f}(a, b, c)$ with typical element

$$\hat{f}_{tz}(a, b, c) = \tilde{Y}_{tz} - a - bt - cg(\bar{S}_{tz}). \quad (14)$$

Then a minimum distance approach to estimating (a, b, c) minimizes the quadratic form $\hat{f}' A \hat{f}$ for a general weighting matrix A. Newey's m_T is the appropriate test statistic for overidentification with a general matrix A

[18] Dispensing with parallel lines may be particularly attractive in light of the evidence we summarize below. Results from the MTO provide at least suggestive evidence against this traditional assumption.

[19] More generally, $g(\cdot)$ is a conditional mean function that depends on the distribution specified. As touched on above, if the economist wishes to invoke distributional assumptions that depart from normality, $g(\cdot)$ may be altered accordingly (see Heckman, Tobias, and Vytlacil 2000, 2003; Lee 1982, 1983).

and may differ somewhat from the more familiar minimized value of the quadratic form itself.[20]

For the case of binary treatment and binary hassling, the minimum distance framework outlined has four equations and three unknowns, and Newey's m_T statistic will be distributed x^2 with 1 df. In the case of a trichotomous treatment, as characterizes the MTO demonstration, the same framework suggests that three different lines should be parallel, corresponding to each potential treatment assignment. The overidentification restrictions are in some sense more binding in such a context, and a parallel treatment to that above shows that Newey's m_T statistic is distributed x^2 with 2 df. In this paper, we implement the test using a variance matrix calculated using the delta method (i.e., using second-order Taylor series approximations).[21]

C. Bounding the Treatment Effects and Other Estimators

Some of the potential limitations of modeling the nonresponse process can be avoided if one is willing to settle for bounds on the relevant treatment effects. Note that both bounds we consider in this section allow for treatment effect heterogeneity.

1. Horowitz and Manski "Worst-Case" Bounds

In the case where the outcome of interest is bounded, an appealing way to proceed is to used the bounds discussed in Horowitz and Manski (2000a). Let $\underline{Y}$ denote the lowest possible value of the outcome, and let $\bar{Y}$ denote the greatest possible value. The bounds are constructed by making the worst-case assumptions about the missing data. The upper and lower bounds of the treatment effect are given by

$$\begin{aligned}\bar{\theta}_M &= P[S = 1|T = 1]E[Y|T = 1] + (1 - P[S = 1|T = 1])\bar{Y} \\ &\quad - P[S = 1|T = 0]E[Y|T = 0] + (1 - P[S = 1|T = 1])\underline{Y}, \\ \underline{\theta}_M &= P[S = 1|T = 1]E[Y|D = 1] + (1 - P[S = 1|T = 1])\underline{Y} \\ &\quad - P[S = 1|T = 0]E[Y|T = 0] + (1 - P[S = 1|T = 1])\bar{Y}.\end{aligned}$$

These are the least restrictive of the bounds we consider. In some cases, when the outcome has wide support the bounds can be quite wide—potentially so wide as to be uninformative. Nonetheless, the bounds can be a

[20] See Newey (1985) and discussion surrounding his eq. (9). When A is the inverse of the variance of $\hat{f}$, Newey's m_T statistic reduces to the minimized value of the quadratic form.

[21] Reassuringly, in several Monte Carlo experiments with a normal-normal Heckman data-generating process, the nominal size of the test was very close to the actual size. We leave a detailed study of the power of the test to future work.

useful benchmark since they require no assumptions about the nature of the selection process. As Horowitz and Manski (2000b) and Manski (2016) observe, tighter bounds require additional assumptions about the selection process.

2. *Lee Bounds (2009)*

The next set of bounds we review are those introduced to the literature by Lee (2009). To discuss these bounds it will be helpful to introduce some additional notation. Let S_1 and S_0 denote counterfactual sample selection indicators for the treatment and control groups, respectively. That is, for a given unit S_1 indicates whether they would be observed if they were assigned to treatment, and S_0 indicates whether they would be observed if they were assigned to control.

The size of the treatment effect is allowed to be different for individuals with different values of the pair (S_0, S_1). The potential estimand may be different from when we modeled the attrition. For example, we will be unable to learn about individuals for whom $(S_0 = 0, S_1 = 0)$—that is, for units that would be missing regardless of whether they would have been assigned to treatment or control.

Instead of a selection equation like equation (4) or equation (10), we have

$$S = S_1 T + S_0(1 - T). \tag{15}$$

In words, a person is observed if (*a*) she is assigned to treatment and $S_1 = 1$ or (*b*) she is assigned to control and $S_0 = 1$.

Lee (2009) notes that with two familiar assumptions it is possible to bound the treatment effect for individuals we will always observe. The first assumption is random assignment of the treatment, T. The second assumption parallels the Imbens and Angrist (1994) notion of monotonicity. Specifically, the assumption is that either $S_1 \geq S_0$ for all individuals or $S_0 \geq S_1$ for all individuals. Practically, this assumption rules out the possibility of the treatment causing any individuals not to respond (or vice versa).

The bounds pertain to a specific parameter that differs from the population average treatment effect and that is instead the average treatment effect for the subpopulation of always observed (nonattriters). To illustrate, assume that $S_1 \geq S_0$, and denote by θ the average treatment effect for the always observed. This means that the fraction observed will be higher in the treatment group than in the control group. The observed individuals in the treatment group will be a combination of two types: the "always observed" on the one hand (i.e., those who would have been observed had they been in the control or treatment), and the "compliers" on the other hand (i.e., those who would not have been observed under control but are observed by virtue of their assignment to treatment).

In the case we have discussed, where treatment increases the probability of observation, Lee bounds are given by

$$\underline{\theta}_L = E[Y|T = 1, S = 1, Y \leq G^{-1}(1 - p_0)] - E[Y|T = 0, S = 1],$$
$$\bar{\theta}_L = E[Y|T = 1, S = 1, Y \geq G^{-1}(p_0)] - E[Y|T = 0, S = 1], \text{ where}$$
$$p_0 = \frac{\Pr[S = 1|T = 1] - \Pr[S = 1|T = 0]}{\Pr[S = 1|T = 1]} r,$$

where $G^{-1}(\cdot)$ is the inverse cumulative distribution function of Y given $S = 1$ and $T = 1$ and ρ_0 represents the fraction of observations to be trimmed from the treatment group to construct the bounds.

To illustrate, suppose that 50% of the treatment group is observed but that only 40% of the control group is. We trim observations from the group that is more frequently observed. Thus, in this case we trim observations from the treatment group. The trimming fraction is given by $p_0 = (0.5 - 0.4)/0.5 = 0.2$. The procedure to compute the upper bound for the treatment effect amounts to the following:

1. Compute the mean outcome for the control group.
2. Drop the lowest 20% of outcomes from the treatment group and calculate the mean for the remaining members of the treatment group.
3. Calculate the difference between the trimmed treatment group mean and the control group mean. Label this difference $\widehat{\bar{\theta}}_L$.

The lower bound, denoted $\widehat{\underline{\theta}}_L$, is calculated in an analogous manner: one trims observations in the treatment group where the values of the outcome are above the 80th percentile for the treatment group. Lee (2009) shows that it is also possible to tighten the bounds using covariates, but we do not pursue that here.

IV. The MTO Experiment

The MTO demonstration is a program providing housing vouchers to families living in housing projects located in high-poverty neighborhoods. MTO has been the subject of extensive analysis in economics and elsewhere; see, for example, Katz, Kling, and Liebman (2001), Kling, Ludwig, and Katz (2005), Goering and Feins (2003), and Chetty, Hendren, and Katz (2016). Because of this extensive literature, we do not dwell on substantive issues. Instead, we focus on features of the MTO most salient regarding the implementation of the methodologies described above.

For our purposes, the critical feature of the MTO evaluation effort is that individuals were de facto randomized at baseline into normal- and high-effort follow-up. As discussed above, this feature is useful for assessing the impact of attrition on estimates.

We do not view the empirical analysis that follows as correcting any specific defect of existing MTO evaluation research—because of the overall quality of the MTO evaluation, response rates at follow-up are a high 90% and, moreover, follow-up surveys were augmented by administrative data with negligible attrition problems. Rather, we view the empirical analysis that follows as an opportunity to show clearly the practical impact of correcting for attrition using the methodologies outlined above and to demonstrate the difficulty of carrying out some of these methodologies with a real data set. Importantly, the administrative data collected by the MTO evaluation team provide us with a benchmark for comparing our estimates to those that assume that the data are missing at random.[22]

A. Background

To be considered eligible for an MTO housing voucher, families had to have children and live in an eligible housing project in Baltimore, Boston, Chicago, Los Angeles, or New York.[23] Families who volunteered for the project were randomly selected for one of three treatment groups: an experimental group, in which families were given a Section 8 housing voucher to be used toward housing in a census tract with less than 10% poor, augmented by some counseling; a Section 8 group, in which families were given a Section 8 housing voucher with no strings and no counseling; and a control group. For each subject $i = 1, 2, \ldots, n$, let $T_i \in \{E, S, C\}$ denote whether the subject was assigned to the experimental group, the Section 8 group, or the control group, respectively.

Subjects faced differing probabilities of treatment assignments depending on the location and date of their treatment assignment. This implies that T_i is randomly assigned conditional on R_i but is not randomly assigned unconditionally, where R_i records location by time for each subject (Orr et al. 2003, exhibit B.3). For example, if during the demonstration the local economy in New York had been stronger than in other MTO cities and if New York had assigned subjects to the experimental group at a higher rate than other MTO cities, those in the experimental group would have faced a stronger economy on average than those in the control group. This implies that the effect of the economy on outcomes confounds unconditional contrasts, posing an identification problem. Previous MTO evaluation research

[22] MTO evaluators have consistently used administrative data as a complement to information gathered from follow-up surveys and have been aware of the problems created by nonrandom attrition. Orr et al. (2003, app. F) estimates attrition bias by comparing intention-to-treat parameters for outcomes from administrative data estimated on the entire sample and on the survey sample.

[23] Eligible projects were selected by local public housing authorities from among housing projects in census tracts with at least 40% poor (Goering et al. 1999).

has addressed this problem by reweighting observations on subjects according to $\hat{P}(T_i = t)/\hat{P}(T_i = t|R_i)$ for those assigned to treatment group $t \in E, S, C$ (i.e., the weights depend on treatment assignment). This is analogous to the average treatment effect reweighting from the propensity score literature, adjusted for trichotomous treatment assignments.[24]

To keep the exposition as simple as possible and to focus attention on issues of selection correction, we ignore both of these issues: we restrict our analysis to the subset of MTO subjects faced with identical treatment assignment regimes.[25] By excluding individuals assigned in different treatment regimes we circumvent the need for weighting and covariate adjustment. The individuals we analyze comprise roughly 1,700 of the roughly 3,500 families analyzed in other MTO research.

B. Partially Randomized Follow-Up in the MTO

A central focus of our analysis is the partial randomization of individuals into normal and high-effort groups for follow-up. In the midst of the MTO follow-up survey, seeking to maximize the number of respondents with valid information in the follow-up survey, the MTO evaluation team made a judgment that "continuing to work" all nonrespondents would not be as effective as targeting effort at a subset of nonrespondents. MTO administrators selected three out of 10 nonrespondents for additional follow-up, using the final digit of the family identifier (2, 5, or 8 for Baltimore and Los Angeles and 3, 6, or 9 for Boston, Chicago, and New York; Orr et al. 2003).

Because the family identifier is a baseline characteristic of individuals and the last digit of the ID is effectively random, we interpret this procedure as specifying that three out of 10 individuals were randomly selected by MTO administrators for intensive follow-up.[26] We adopt the notation from above and define $Z_i = 1$ for those with last digit of the family identifier 2, 5, or 8 (3, 6, or 9) for those in Baltimore and Los Angeles (Boston, Chicago, and New York)—regardless of whether the individual was surveyed in the initial attempt at follow-up. Defined in this way, it is reasonable to presume (1) that Z_i is independent of all baseline characteristics of individuals and (2) that those with $Z_i = 1$ will be observed with greater frequency than those with

[24] Additionally, MTO evaluation research has regression adjusted for baseline characteristics. In our own analysis of the MTO data, we have found that these regression adjustments reduce sampling variation only slightly.

[25] These individuals faced $P(T_i = E) = 0.5$, $P(T_i = S) = 0.1875$, and $P(T_i = C) = 0.3125$. See Orr et al. (2003, exhibit B.3).

[26] Orr et al. (2003, app. B, p. B-2) describe the procedure: "Our strategy was to continue to work 3 in 10 of the cases that had not been completed during the main field period. By continuing to work a random subsample of cases, we were able to achieve a higher effective response rate than if we had used the same resources to continue to work the full sample."

$Z_i = 0$. These attributes make Z_i a natural candidate for an instrument for the observation equation.

V. Empirical Results

A. Implementation

How does our proposed approach compare to standard approaches to sample selection in the literature? We use the MTO data described above to present estimates of the intention-to-treat effects, or bounds on that parameter, for key outcomes. For ease of interpretation, we present four sets of figures and four sets of tables corresponding to the four outcomes for which we have MTO administrative data. Figure 4 and table 1 report on our results using fraction of quarters employed; figure 5 and table 2 do the same for annualized earnings; figure 6 and table 3 do the same for Temporary Assistance for Needy Families (TANF) receipt; and figure 7 and table 4 do the same for TANF amount.

Following our earlier notation, $\tilde{Y}_{tz}$ denotes the censored sample means (i.e., the sample analogues of $E[Y|T = t, Z = z, S = 1]$) and $\bar{S}_{tz}$ denotes the probability of being observed (i.e., the sample analogues of $E[S|T = t, Z = z]$), where the first subscript denotes whether the observation is from the control, experimental, or Section 8 group ($t = 0, 1, 2$), respectively, and the second subscript ($z = 0, 1$) denotes whether the individual was randomized into nonintensive follow-up ($Z = 0$) or intensive follow-up ($Z = 1$), respectively. Finally, we define $\tilde{M}_{tz} = \phi(\Phi^{-1}(\bar{S}_{tz}))/\bar{S}_{tz}$, the estimated value of the inverse Mills ratio term computed using $\bar{S}_{tz}$. Each of the subsequent figures presents the following:

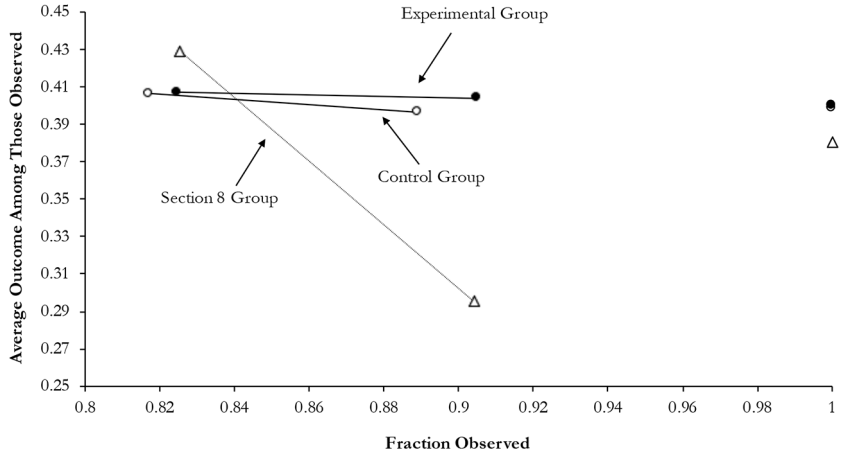

FIG. 4.—Graphical interpretation of sample selection correction: fraction of quarters employed, years 1–4.

Table 1
Estimated Moving to Opportunity Program Impacts: Fraction of Quarters Employed, Years 1–4

Estimator	Intention-to-Treat Parameters		Control Mean	Selection Correction	χ^2 Statistic for Test That Hassling Is Irrelevant	Number of Observations
	Experimental	Section 8				
Full administrative sample, OLS (ideal solution)	.001	−.018	.399			1,248
	(.024)	(.030)	(.019)			
Main sample, OLS	.003	−.018	.404			1,055
	(.026)	(.033)	(.020)			
Worst-case bounds	−.161, .162	−.173, .140	NA			
Lee bounds	−.007, .008	−.027, −.014	NA			
	(.026), (.026)	(.032), (.033)				
	[.0137^{E}]	[.0143^{C}]				
Heckman two-step with randomized Z	.007	−.013	.329	.257	12.320	1,055, 1,248
	(.028)	(.035)	(.066)	(.213)	$p = .006$	

NOTE.—This table presents estimates of the average intention-to-treat effects using the approach noted in the far left column. Standard errors are in parentheses. The first row is estimated using complete administrative data, while other estimators use only the subset of the data with nonmissing responses to the outcome survey. The numbers in brackets under Lee bounds show the fraction of observations that are trimmed to construct Lee's bounds, and the superscript letters indicate which group's outcome distribution is trimmed, as follows: control group (C), experimental group (E), and Section 8 group (S). NA = not applicable; OLS = ordinary least squares.

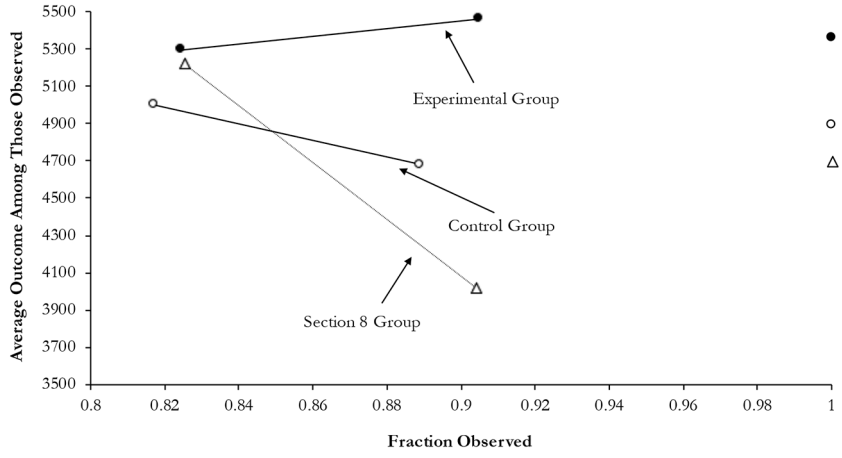

FIG. 5.—Graphical interpretation of sample selection correction: annualized earnings for years 1–4.

1. Separately for $Z = 0$ and $Z = 1$, the (observed) means for the control group, the experimental group, and the Section 8 group, plotted against their respective probabilities of being nonmissing.
2. The administrative (complete data) means for the control group, the experimental group, and the Section 8 group. These are plotted on the right-hand axis, where the probability of observation equals 1.

Note again that estimating treatment effects requires "connecting the dots" $(\bar{S}_{t0}, \tilde{Y}_{t0})$ and $(\bar{S}_{t1}, \tilde{Y}_{t1})$ and extrapolating the censored mean functions to the same estimated probability of observation (see eq. [9]). The appropriate scale for the horizontal axis in each figure depends on the distribution of V: assuming normality, the figures would best be displayed in inverse Mills ratio units, in which case the three slopes of the lines connecting the dots could be read directly from the figure. However, in general we do not know whether normality of V is an appropriate assumption. In light of this ambiguity, we elected to maintain the horizontal axis in terms of the probability of observation.

The figures we present display the key summary statistics needed to compute the overidentifying restrictions test we described above. We interpret this test as a test of the equality of the slopes of the censored means of the outcome in inverse Mills ratio terms. Our intuition for this interpretation is that under $\rho_1 = \rho_0$, we would expect

$$\frac{\tilde{Y}_{01} - \tilde{Y}_{00}}{\tilde{M}_{01} - \tilde{M}_{00}} \approx \frac{\tilde{Y}_{11} - \tilde{Y}_{10}}{\tilde{M}_{11} - \tilde{M}_{10}} \approx \frac{\tilde{Y}_{21} - \tilde{Y}_{20}}{\tilde{M}_{21} - \tilde{M}_{20}}.$$

Table 2
Estimated Moving to Opportunity Program Impacts: Annualized Earnings, Years 1–4

Estimator	Intention-to-Treat Parameters		Control Mean	Selection Correction	χ^2 Statistic for Test That Hassling Is Irrelevant	Number of Observations
	Experimental	Section 8				
Full administrative sample, OLS (ideal solution)	473.32	−190.94	4,888.68			1,248
	(475.58)	(562.69)	(366.74)			
Main sample, OLS	442.54	−75.18	4,904.87			1,055
	(517.76)	(622.94)	(405.62)			
Lee bounds	201.01, 514.39	−431.40, −27.12	NA			
	(437.06), (519.61)	(593.23), (625.88)				
	[.0137^{E}]	[.0143^{C}]				
Heckman two-step with randomized Z	479.55	−36.91	4,333.66	1,961.07	12.320	1,055, 1,248
	(520.57)	(658.02)	(1,230.66)	(3,988.10)	$p = .006$	

NOTE.—This table presents estimates of the average intention-to-treat effects using the approach noted in the far left column. Standard errors are in parentheses. The first row is estimated using complete administrative data, while other estimators use only the subset of the data with nonmissing responses to the outcome survey. The numbers in brackets under Lee bounds show the fraction of observations that are trimmed to construct Lee's bounds, and the superscript letters indicate which group's outcome distribution is trimmed as follows: control group (C), experimental group (E), and Section 8 group (S). NA = not applicable; OLS = ordinary least squares.

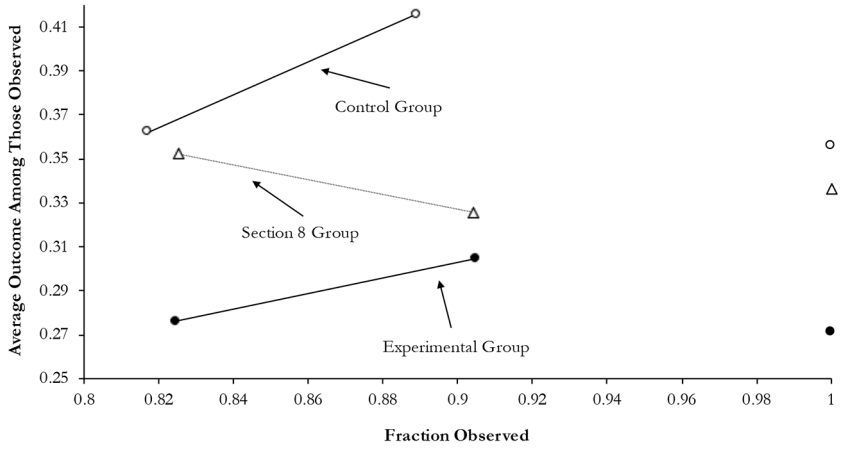

FIG. 6.—Graphical interpretation of sample selection correction: Temporary Assistance for Needy Families recipiency, year 5.

Our proposed test rejects when Newey's m_T is sufficiently far from zero, which occurs when, at the minimizer, at least one of $\hat{f}_{tz}$ is far from zero. When the test rejects, best practice would include potentially adopting a different estimand (e.g., a GATE parameter as discussed) and extrapolating to an interior point of support for the probability of observation. This would not require an assumption of parallel lines. Alternatively, the economist could instead resort to bounding the treatment effect using Manski or Lee bounds.

The table that corresponds to each outcome/figure provides the corresponding numerical estimates as well as information regarding the sensitivity of the estimates to different assumptions about the missing data process. Specifically, in each table we present the following:

1. Estimated program impacts using the complete administrative data.
2. Estimated program impacts using the administrative data but restricted to the subsample of observations with nonmissing survey data. If the survey data are missing at random, then these provide unbiased, consistent estimates of the intention-to-treat parameters. If the survey data are not missing at random, then the estimates would be expected to differ from the population estimate.
3. Estimated worst-case bounds for the treatment effects (Horowitz and Manski 2000a).
4. Estimated (unconditional) bounds for the treatment effects (Lee 2009).

Table 3
Estimated Moving to Opportunity Program Impacts: Temporary Assistance for Needy Families Recipiency, Year 5

Estimator	Intention-to-Treat Parameters		Control Mean	Selection Correction	χ^2 Statistic for Test That Hassling Is Irrelevant	Number of Observations
	Experimental	Section 8				
Full sample, OLS (ideal solution)	−.084	−.019	.356			1,248
	(.027)	(.036)	(.022)			
Main sample, OLS	−.093	−.034	.378			1,055
	(.030)	(.030)	(.024)			
Worst-case bounds	−.165, .157	−.187, .126				
Lee bounds	−.104, −.089	−.044, −.031				
	(.030), (.030)	(.039), (.039)				
	[.0137^{E}]	[.0143^{S}]				
Heckman two-step with randomized Z	−.097	−.038	.437	−.201	12.320	1,055, 1,248
	(.031)	(.039)	(.074)	(.239)	$p = .006$	

NOTE.—This table presents estimates of the average intention-to-treat effects using the approach noted in the far left column. Standard errors are in parentheses. The first row is estimated using complete administrative data, while other estimators use only the subset of the data with nonmissing responses to the outcome survey. The numbers in brackets under Lee bounds show the fraction of observations that are trimmed to construct Lee's bounds, and the superscript letters indicate which group's outcome distribution is trimmed as follows: control group (C), experimental group (E), and Section 8 group (S). OLS = ordinary least squares.

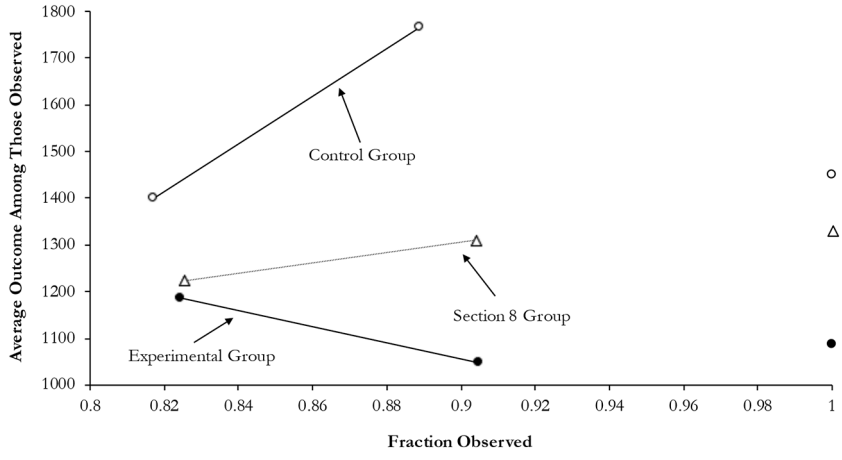

FIG. 7.—Graphical interpretation of sample selection correction: Temporary Assistance for Needy Families amount, Year 5.

5. Estimated program impacts using a Heckit approach based on randomized exposure to hassling. We also provide the overidentifying test described above.[27]

Each row thus represents the results of various approaches to the missing data problem that are routinely used in the literature along with the ideal though generally infeasible (complete data) estimator. Table 5 summarizes results of the overidentifying test statistics.

B. Fraction of Quarters Employed and Annualized Earnings

Beginning first with our graphical analysis, figure 4 presents the key descriptive statistics. The estimates on the right-hand axis, at a probability of observation equal to 1, show estimates of experimental, treatment, and control group means. The other moments plotted in the figure show the observed outcomes and fraction observed among the subset of individuals with nonmissing survey data. It is evident that the more intensive follow-up procedure was effective at increasing the probability of observation. With the less intensive follow-up, the fraction observed is about 0.10 less than the fraction who were subjected to more intensive follow-up. The suggestion from the point estimates is that outcomes were not as good for those who could be obtained only with more extensive follow-up. That is, the evidence suggests

[27] Estimates using the maximum-likelihood version of the Heckman estimator were generally similar to the two-step estimator, although in some cases they failed to converge, which caused one of the coauthors endless delight. See DiNardo, McCrary, and Sanbonmatsu (2007) for these results and further discussion.

Table 4
Estimated Moving to Opportunity Program Impacts: Temporary Assistance for Needy Families Amount, Year 5

Estimator	Intention-to-Treat Parameters		Control Mean	Selection Correction	χ^2 Statistic for Test That Hassling Is Irrelevant	Number of Observations
	Experimental	Section 8				
Full sample, OLS (ideal solution)	−361.85	−118.63	1,448.91			1,248
	(131.17)	(173.51)	(106.77)			
Main sample, OLS	−367.03	−257.33	1,508.586			1,055
	(145.08)	(179.76)	(118.06)			
Lee bounds	−477.02, −351.69	−244.88, −345.46				
	(140.35), (145.59)	(173.96), (180.65)				
	[.0137^{E}]	[.0143^{S}]				
Heckman two-step with randomized Z	−373.85	−264.38	1,613.86	−361.41	12.320	1,055, 1,248
	(142.72)	(180.40)	(337.23)	(1,092.93)	$p = .006$	

NOTE.—This table presents estimates of the average intention-to-treat effects using the approach noted in the far left column. Standard errors are in parentheses. The first row is estimated using complete administrative data, while other estimators use only the subset of the data with nonmissing responses to the outcome survey. The numbers in brackets under Lee bounds shows the fraction of observations that are trimmed to construct Lee's bounds, and the superscript letters indicate which group's outcome distribution is trimmed as follows: control group (C), experimental group (E), and Section 8 group (S). OLS = ordinary least squares.

Table 5
Tests for Equality of Slopes

	Control Slope	Experimental Slope	Section 8 Slope	m_T Test Statistic χ^2	p-Value Under Null of Equal Slopes
Fractions of quarters employed, years 1–4	.090	.026	1.081	1.884	.61
	(.4111)	(.2750)	(.7221)		
Annualized earnings, years 1–4	2,864	−1,314	9,681	.978	.39
	(7,956)	(5,450)	(9,894)		
TANF recipiency, Year 5	−.482	−.227	.214	.882	.36
	(.5481)	(.3157)	(.5280)		
TANF amount, Year 5	−3,299	1,098	−701	1.981	.63
	(2,935)	(1,369)	(2,435)		

NOTE.—The slopes are calculated as the difference in the observed outcome means for $Z = 0$ and 1 (those not subject to intense follow-up and those subject to intense follow up, respectively) divided by the respective difference in the inverse Mills ratio terms, separately for each treatment arm. Assuming the error in the selection equation is normal, different slopes suggest that treatment effect heterogeneity is correlated with the probability of observation. TANF = Temporary Assistance for Needy Families.

positive selection into observation: mean outcomes for those with $Z = 1$ are all less than mean outcomes for those with $Z = 0$.

Table 1 presents the results of our analysis for the outcome "fraction of quarters employed, years 1–4." Although the MTO analysis includes a broader sample than we employ in our analysis, our estimates of the impact of the treatments are qualitatively similar. In the first row of the table, we display the point estimates for the experimental group and the Section 8 treatment group, both relative to the control group, using the full administrative data sample. These are 0.001 and −0.018, respectively (standard errors of 0.024 and 0.030, respectively). In no case can the null hypothesis of no treatment effect be rejected at conventional levels of significance. The substantive significance of these estimates can be gauged relative to the mean values of the outcome. In the case of the control group, the mean fraction of quarters employed is about 0.39.

Proceeding stepwise through various approaches to the selection bias problem, the second row displays estimates using just the survey sample with no adjustments. This approach will yield consistent estimates in the case that data are missing completely at random. In all cases the estimates are well within sampling variability of the full (administrative) data estimates.

In third row and fourth rows of the table, we compute bounds on the treatment effects. Since the fraction of quarters employed is naturally bounded between 0 and 1, the Horowitz and Manski bounds involve no further assumptions but yield bounds that are quite wide: (−0.161, 0.162) for the experimental group and (−0.173, 0.140) for the Section 8 group. The bounds suggested by

Lee (2009) are shown in row 5 and are substantially narrower than the worst-case bounds, especially for the Section 8 group. Still, as might be expected under a one-sided selection model, they provide intervals that are somewhat large and comfortably include the full sample point estimates.

In the fifth row, we present the estimates using the two-step Heckman selection correction procedure based on a randomized Z. The fifth column presents our test statistic evaluating whether our instrument Z indeed induces a higher response. Not surprisingly, the null hypothesis of no effect of the instrument on the probability of being observed is rejected at conventional levels of significance. The point estimates are both slightly larger than those in the second row based on the observed data and are further from the "true" treatment effects shown in row 1. That said, the estimated treatment effects are similar in magnitude and within the sampling error of the full administrative data results. Moreover, the estimate is considerably more precise than the Lee bounds.

In this case, using Z and the Heckman correction yields similar treatment effect estimates, since the degree of selection is relatively low. While this may seem anticlimactic, in the typical case researchers will not have access to complete administrative data to recognize this. In that light, this case demonstrates the value of Z in showing that assuming data are missing at random may be a reasonable assumption, supported by the insignificant coefficient on the selection correction term.

The first row of table 5 shows the estimates of the three slopes depicted in figure 4, along with the m_T test statistic and p-value for the joint test of normality and equal slopes. Despite the larger slope observed for the Section 8 group in figure 4, we fail to reject the null. This likely stems from the fact that the slopes are quite imprecisely estimated. It is also interesting to observe that the point estimates are not consistent with parallel slopes and that the administrative data even suggest that the censored mean curve (at least for the Section 8 group) is nonmonotonic. The complete data sample means are all higher than the means for the "hard-to-get" group but are about the same level for those persons interviewed without need for extensive follow-up. While this could be an artifact of sampling error in this instance, that this would ordinarily be blind to the economist underscores the potential limitations of the selection correction framework and a binary Z. For example, the apparent difference in slopes in figure 4 might lead the economist to estimate a more general selection correction that eschews the parallel slopes (i.e., $\rho_1 = \rho_0$ in [A1]) assumption. But it is clear in this example that would lead to a point estimate of the Section 8 treatment effect for the entire population that is much more negative than that in row 5 of table 1 (assuming parallel slopes) and further from the treatment effect based on the full administrative data.

In figure 5 and table 2, we present the results for annualized earnings. The pattern of results is quite similar to those for fraction of quarters employed.

Again, both the Heckman two-step and Lee bounds provide similar inferences as the complete (administrative) data as well as the potentially biased ordinary least squares (OLS) estimates using only the survey data. The estimates from the Heckman two-step procedure, however, again has much narrower confidence intervals.

C. TANF Results

Figures 6 and 7 and tables 3 and 4 display our results for receipt of TANF and the dollar amount of TANF, respectively. These results depart somewhat from the results for the broader sample used by the MTO investigators: in our smaller analysis sample,[28] the full sample OLS estimates for the experimental group indicate beneficial (i.e., negative) treatment effects on TANF receipt and the dollar amount of assistance. These economically large estimates are statistically distinct from zero at conventional significance levels.[29]

There are several potential explanations for the departure of our full sample estimates from those reported by the MTO investigators, including, of course, sampling error. For example, our analysis sample does not include those randomized into treatment in the second and later rounds. It is consequently more heavily weighted with observations from early periods before state efforts to remove people from TANF eligibility were in full swing.

Turning to the results for different approaches to handling attrition, the results are qualitatively similar to those presented for employment and earnings outcomes. The Heckman two-step estimator based on a randomized Z performed quite well, with standard errors only modestly larger than their full sample counterparts and much smaller than those implied by the worst-case or Lee bounds, although the point estimates are more negative than the full sample results. The point estimates of the selection process suggest that those who are more difficult to follow up are more likely to receive TANF. Again, however, the point estimates are inconsistent with monotonicity—although the "harder to get" appear negatively selected when restricted to those who responded to the survey.

VI. Conclusion

It is widely appreciated that problems of missing data can undermine the validity of even the simplest inferential problems. In this paper, we propose a proactive strategy to deal with this problem that involves partial

[28] Our analysis sample is restricted to include only those individuals randomized with the same randomization ratios. See sec. IV.A for discussion.

[29] The estimates for the Section 8 treatment are imprecise, reflecting the smaller sample size for this group. Perhaps not surprisingly, the Lee bounds fail to include the OLS estimate from the completed data sample.

randomization of nonresponse. The core of our proposal is for economists concerned about the impact of nonresponse on their findings to involve themselves in survey design where possible. Econometric input into the survey process holds out the promise of making inferences more robust to nonresponse.

We propose a simple graphical analysis that may be useful in detecting the potential contamination arising from selection bias. We use this device to develop several points. First, a traditional parametric approach to selection bias involves a series of strong assumptions. The partial randomization we are advocating for will furnish the critical instrument for the probability of observation required by that approach. However, it will not ensure that the additional assumptions of the traditional approach are correct. An economist seeking to adopt a more circumspect approach might be willing to compromise on identifying the population average treatment effect sought by the traditional parametric approach and instead might seek to identify what we have termed a "gettable average treatment effect," or GATE. We show how different GATE parameters may be identified from the observed data under weaker assumptions than in the traditional parametric framework. We also note that both Manski and Lee bounds are available for problems of this type.

We implement these proposals on a well-known and well-conducted experiment, the MTO demonstration, for a subset of outcomes for which complete (administrative) data are available. In this case, the Heckit approach based on randomization to more intense follow-up yields similar estimates to the (unbiased) estimates based on the complete administrative data, but with confidence intervals substantially narrower than the inference permitted by bounding approaches. Further research should assess how well the approach performs in contexts where differential attrition is a more major concern than in the MTO and thus where more sample selection bias is expected.

Our work echoes that of many others in concluding that in the absence of an instrument with many points of support, functional form restrictions will be needed in order to identify meaningful parameters. However, it is as yet unresolved how practical it might be in a variety of settings to engage in randomization. We write this paper with skepticism that more than two points of support for the probability of observation will be practicable. Yet even if survey firms were able to engage in "high, medium, low" approaches as opposed to "high, low" approaches, that could be a major improvement in terms of what could learned about the nonresponse problem.

References

Ahn, Hyungtaik, and James L. Powell. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, no. 1/2:3–29.

Altman, Douglas G., Kenneth F. Schulz, David Moher, Matthias Egger, Frank Davidoff, Diana Elbourne, Peter Getzche, and Thomas Lang. 2001. The revised consort statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine* 134, no. 8:663–94.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy* 127, no. 5:2143–85.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. Beyond LATE with a discrete instrument. *Journal of Political Economy* 125, no. 4:985–1039.

Bureau of Labor Statistics and Census Bureau. 2002. Current Population Survey: Design and methodology. Census Bureau Technical Paper no. 63RV.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review* 106, no. 4:855–902.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, no. 1:187–99.

Das, Mitali, Whitney Newey, and Francis Vella. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* 70, no. 1:33–58.

DiNardo, John E., Justin McCrary, and Lisa Sanbonmatsu. 2007. Constructive proposals for dealing with attrition. Unpublished manuscript, University of Michigan.

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Beccerra. 2019. Testing attrition bias in field experiments. Unpublished manuscript, University of California, Davis.

Goering, John, and Judith D. Feins. 2003. *Choosing a better life? Evaluating the Moving to Opportunity social experiment*. Washington, DC: Urban Institute Press.

Goering, John, Joan Kraft, Judith D. Feins, Debra McInnis, Mary Joel Holin, and Huda Elhassan. 1999. Moving to Opportunity for fair housing demonstration program. Unpublished manuscript, US Department Housing and Urban Development.

Grovers, Robert M., and Mick P. Couper. 1998. *Nonresponse in household interview surveys*. New York: Wiley.

Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, no. 4:475–92.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47, no. 1:153–62.
Heckman, James J., Justin Tobias, and Edward Vytlacil. 2000. Simple estimators for treatment parameters in a latent variable framework with an application to estimating the returns to schooling. Unpublished manuscript, University of Chicago.
———. 2003. Simple estimators for treatment parameters in a latent variable framework. *Review of Economics and Statistics* 85, no. 3:748–55.
Hirsch, Barry T., and Edward J. Schumacher. 2004. Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics* 22, no. 3:689–722.
Horowitz, Joel L., and Charles F. Manski. 1998. Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics* 84, no. 1:37–58.
———. 2000a. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95:77–84.
———. 2000b. Rejoinder: Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95:87.
Imbens, Guido. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86, no. 1:4–29.
Imbens, Guido, and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, no. 2:467–75.
Juni, Peter, Douglas G. Altman, and Matthias Egger. 2001. Assessing the quality of controlled clinical trials. *British Medical Journal* 323:42–46.
Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. Moving to Opportunity in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics* 116, no. 2:607–54.
Kline, Patrick, and Christopher R. Walters. 2019. On Heckits, LATE, and numerical equivalence. *Econometrica* 87, no. 2:677–96.
Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz. 2005. Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *Quarterly Journal of Economics* 120, no. 1:87–130.
Lee, David S. 2005. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Unpublished manuscript, University of California, Berkeley.
———. 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* 76, no. 3:1071–102.
Lee, Lung-Fei. 1982. Some approaches to the correction of selectivity bias. *Review of Economic Studies* 49, no. 3:355–72.

———. 1983. Generalized econometric models with selectivity. *Econometrica* 51, no. 2:507–12.
Manski, Charles F. 1989. Anatomy of the selection problem. *Journal of Human Resources* 24:343–60.
———. 1990. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* 80:319–23.
———. 1994. The selection problem. In *Advances in econometrics Sixth World Congress*, ed. Christopher Sims, no. 23 in Econometric Society Monographs, chap. 4, 143–70. Cambridge: Cambridge University Press.
———. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.
———. 2016. Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics* 191:293–301.
Moher, David, Alison Jones, and Leah Lepage. 2001. Use of the consort statement and the quality of reports of randomized trials: A comparative before-and-after evaluation. *Journal of the American Medical Association* 285, no. 15:1992–95.
Mroz, Thomas A. 1987. The sensitivity of an empirical model of married women's hours off work to economic and statistical assumptions. *Econometrica* 55, no. 4:765–99.
Newey, Whitney. 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29, no. 3:229–56.
Olsen, Randall J. 1980. A least squares correction for selectivity bias. *Econometrica* 48, no. 7:1815–20.
Orr, Larry, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling. 2003. Moving to Opportunity interim impacts evaluation. Final report, US Department of Housing and Urban Development.
Rothbaum, Jonathan, and Adam Bee. 2020. Coronavirus infects surveys, too: Nonresponse bias during the pandemic in the CPS ASEC. SEHSD Working Paper no. 2020-10, US Census Bureau.
Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley.