Constructive Proposals for Dealing with Attrition: An Empirical Example[*]

John DiNardo
University of Michigan, NBER

Justin McCrary
University of Michigan, NBER

Lisa Sanbonmatsu
NBER

July 21, 2006

Abstract

Survey nonresponse and attrition undermine the validity of many and possibly most econometric estimates. In this paper, we discuss simple methods for assessing the nature of the damage caused by such problems. A fundamental component of the methods we discuss is partial randomization of non–response. We evaluate the performance of simple grouped data estimators for sample selection (a là Gronau and Heckman) as well as conservative bounds (a là Manski and Lee) when nonresponse is partially randomized. We develop a graphical interpretation of the conventional sample selection estimator and use it to develop tests of overidentification. To make these ideas concrete we apply the methodology to a carefully conducted randomized controlled trial, the Moving to Opportunity (MTO) demonstration project. Two features of the MTO study are useful for our purposes. First, for a subset of outcomes the MTO study team collected administrative data for *all* persons, allowing for estimation of the bias associated with standard sample selection procedures. Second, thirty percent of subjects were *randomly* assigned to intensive follow–up. In this application the extent of the bias caused by non-response is small. However, our evidence suggests both the importance of addressing nonresponse bias and the real challenges involved, even in this ideal situation.

[*]Thanks to Jane Garrison for excellent research assistance and for help and ideas from Jeff Kling, without whom the empirical example would not have been possible.

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 elsewhere. 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, appear to 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 percent. During the 1990s, nonresponse rates rose to over 6 percent (Bureau of Labor Statistics and Census Bureau 2002) and in recent years have risen to roughly 8 percent (Census Bureau 2005). Similar rates of non-response are associated with the Survey of Income and Program Participation and are further exacerbated by attrition in follow–up waves (Census Bureau 2001). The final wave of the 1996 panel of the SIPP suffers from 35 percent nonresponse. Even when households are interviewed, key variables of interest are often missing. Income non-response rates in the CPS in the 1990s were roughly 15 percent (Bureau of Labor Statistics and Census Bureau 2002) and in more recent years have risen to around 30 percent (Hirsch and Schumacher 2004).

Missing data pose no difficulty if they are randomly missing. However, there is wide agreement that the characteristics of those who are missing certain information are likely to be different from the characteristics of those who are not. This view is corroborated by careful matched records analysis linking household surveys to the population census (Grovers and Couper 1998). 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), bounding the parameter of interest (Manski 1989, 1990, 1994, 1995, Lee 2005, Horowitz and Manski 1998), 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 (Das, Newey and Vella 2003).

Economists routinely devise credible instruments for endogenous regressors in the simultaneous equations model, and evaluate programs where intention-to-treat is randomized. 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.[1,2]

In this paper, we argue that circumventing attrition requires a proactive approach in which the probability of observation is partially randomized. The most obvious way to do so is to alter data collection procedures, such as by randomizing half of sampled units to more intensive follow–up. 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, and despite

[1] In large nationally representative samples, the procedures for dealing with non-response include "hot decking" (assignment of some individual's completed response to non-respondents) and sample weight adjustments. These procedures may be valid if non-response is idiosyncratic, conditional on the variables used for imputation. However, this is rarely credible. Further, as emphasized by Horowitz and Manski (1998), analyses conducted with sample weight adjustments may yield estimates of parameters that are not logically possible.

[2] 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 non-response has occurred (Altman, Schulz, Moher, Egger, Davidoff, Elbourne, Getzche and Lang 2001, Moher, Jones and Lepage 2001).

the insights of the sample selection literature, economists involved in data collection efforts typically do not currently recommend partial randomization of the probability of observation. 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.[3]

We develop a joint test of distributional assumptions and additive separability. Additive separability is the assumption at the heart of nonparametric approaches to sample selection correction (e.g., Ahn and Powell 1993). Under certain configurations of the data, the test becomes primarily a test of additive separability, rather than distributional assumptions. We discuss and provide intuition for these tests using a simple graphical interpretation of the Heckman two-step estimator. This graphical interpretation clarifies the role of additive separability in identification of the sample selection model.

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 percent of participants to be followed up more intensively than others.[4] 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 upon responses from effectively all respondents to (2) estimators that assume missing data occurs at random and (3) sample selection correction estimators that take advantage of the MTO survey design. We also use the data to construct bounds on the effect of MTO treatments on adult outcomes.

The remainder of the paper is organized as follows. In Section II we discuss the conceptual problems with correcting for attrition. Section III gives background information on the MTO

[3] Persuading survey administrators of the value of sample selection correction estimators may be an important impediment.

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

experiment. Section IV presents our results, and Section V concludes.

II. Conceptual Framework

For simplicity, suppose that MTO treatments were binary (the extension to the trichotomous case is straightforward). Using similar notation to Heckman and Vytlacil (2005), define the latent variables

$$\begin{aligned} Y^* &= TY_1^* + (1-T)Y_0^* \\ S^* &= \mu_S(X,T,Z) + U_S \end{aligned}$$

where $Y_1^* = \mu_1(X,U_1)$ and $Y_0^* = \mu_0(X,U_0)$ are counterfactual outcomes. We observe

$$\begin{aligned} Y &= \begin{cases} Y^* & \text{if } \ S=1 \\ \text{missing} & \text{otherwise} \end{cases} \\ S &= \mathbf{1}(S^* \geq 0) \end{aligned}$$

Here U_0, U_1, and U_S are stochastic, mean zero residuals, X is a vector of baseline characteristics, Z is a factor that may affect the probability of observation and is assumed independent of U_0, U_1, and U_S, and μ_0, μ_1, and μ_S are general functions. To maintain a focus on the problem of sample selection, as opposed to the problem of selection into treatment, we assume that T is assigned at random.

The MTO literature focuses on estimation of the intention-to-treat parameter $E[Y_1^* - Y_0^*]$. Because the control group in MTO was self-selected on the basis of interest in the program, but embargoed from participation, this parameter is proportional to what Heckman and Vytlacil (2005) have termed the policy relevant treatment effect.

However, missing data potentially compromises the ability to estimate the intention-to-treat parameter. The (population) average difference between those observed in the treatment and control groups, $E[Y|T = 1, S = 1] - E[Y|T = 0, S = 1]$, does not identify $E[Y_1^* - Y_0^*]$ unless the data are missing-at-random. Researchers unwilling to invoke this assumption have considered two approaches: (i) assume the data are missing at random conditional on X, (ii) assume that Z enters μ_S in a non-trivial way. The first approach leads to propensity score weighting or matching approaches, while the second leads to parametric or nonparametric sample selection correction techniques.

Before proceeding, we would like to acknowledge that there is not necessarily a tight connection between being observed in general missing data settings and the outcome variable under study. This may be an important difference between the study of wages in the presence of heterogeneity in labor market participation, on the one hand, and the study of outcomes in the presence of missing data generally. The one-sided selection model may be overly restrictive in such a setting. The present paper nonetheless operates within the context of the one-sided selection model, focusing on developing intuition for its assumptions and formalizing apparent failures of its assumptions using statistical tests of overidentification. We suspect that the incorporation of partial randomization of follow–up would allow for credible estimation for more general models of selection bias. However, we do not explore those here.

III. Sample Selection in the Grouped Data Case

A. A Simple Illustration

Grouped data approaches to correcting for sample selection bias were proposed in a set of important early papers (Gronau 1973, 1974, Lewis 1974).[5] These papers led to substantial subsequent development in the literature. Heckman (1979) derives statistical properties of the estimators at the micro data level. Heckman (1976) characterizes the relationship of sample selection models to other models. Heckman and Honore (1990) delineate what is identifiable. Heckman and Robb (1984), Newey (1999), Ahn and Powell (1993), and Das

[5] See also Heckman (1974).

et al. (2003) discuss nonparametric estimation of sample selection models. Heckman (1987, 1990) provides surveys of the literature.

In the original Gronau and Lewis application, the motivation was to make economically relevant wage comparisons between different demographic groups. The problem was what to make of the comparison between groups when the fraction actually observed working in each group differed. Consider one of the special cases of the model they studied: the distribution of reservation wages is normal for each of two groups (indexed by T, as in Section II), with common variance σ, and workers take a job if and only if the reservation wage exceeds a threshold value c, assumed to be the same for all workers. Endogeneity concerns are set to the side, so that $E[Y^*|T=1] = E[Y_1^*] \equiv \mu_1$ and $E[Y^*|T=0] = E[Y_0^*] \equiv \mu_0$.[6] Even in this near-ideal case, the comparison of observed mean wages for any two groups is biased for the population difference because

$$\begin{aligned} E[Y|T=1,S=1] - E[Y|T=0,S=1] &= \mu_1 - \mu_0 + \sigma\left(\frac{\phi(\frac{\mu_1-c}{\sigma})}{\Phi(\frac{\mu_1-c}{\sigma})} - \frac{\phi(\frac{\mu_0-c}{\sigma})}{\Phi(\frac{\mu_0-c}{\sigma})}\right) \\ &\equiv \mu_1 - \mu_0 + \sigma\left(g(S_1) - g(S_0)\right) \end{aligned}$$

where $g(s) = \phi(\Phi^{-1}(s))/s$ is a nonlinear function of the participation or survival rate, or the propensity score for being observed, conditional on the observables (Ahn and Powell 1993). In this example, since group status is the conditioning variable, the propensity score is either equal to S_1, the survival rate for group 1, or S_0, the survival rate for group 0. Note that if $S_1 = S_0$, then the population difference is identified, even if there is sample selection. This is a theme of the approaches we discuss and is a consequence of the additive separability assumption discussed in further detail below. The results of Lee (2005) emphasize that this assumption is implied by the structure of the one-sided sample selection model itself.

[6] This is a special case of the model in Section II, with $\mu_0(X,U_0) = \mu_0 + U_0$, $\mu_1(X,U_1) = \mu_1 + U_1$, $\mu_S(X,T,Z) = -c$, and $U_S = Y$. Note that we are implicitly thinking of T as being generated by actual random assignment and the population of interest as being that which has self-selected into being interested in receiving treatments. These implicit assumptions are tied to our empirical example, discussed below, in which these conditions hold. In many other applications, compliance and self-selection likely warrant more than the cursory treatment we give here.

As Gronau (1974) and Lewis (1974) emphasize, built into this framework is an explicit and non-trivial model of behavior—if it were not the case that people acted according to a reservation wage rule of the sort described, the implied econometric model would be different. In this case, the mean wage conditional on being observed is a declining function of the participation rate. In particular, the slope of each conditional mean function is given by the derivative of the inverse Mills ratio term with respect to the participation rate.

This situation is depicted in Example 1, which plots the (censored) mean wage for individuals in the two groups, given a hypothetical participation rate. The true wage differential between the two groups is given by the vertical line at a fixed value of the participation rate. Under the one-sided selection model outlined above, this vertical difference is the same for all values of the participation rate (Das et al. 2003). The bias induced by sample selection arises because in this simple model the participation rate of low wage workers is less than the participation rate of high wage workers. Given the selection rule and the assumptions about the wage distribution, the conditional mean represents a more select set of reservation wages for low wage workers. This leads the researcher's naive estimate of the wage differential to be too small, since wages for low skill workers are, relative to those of high skill workers, artificially high since they are more likely to select out of work.

It is important to note that, in this framework, the wage comparison between the two groups is not *necessarily* biased. Given the assumptions above about the distribution of reservation wages, an unbiased estimate would obtain in the unlikely event that the employment rate of the two groups happened to be the same. This would occur if and only if the difference in reservation wages between the two groups were exactly equal to the difference in population wage rates.

B. A Standard Parametric Threshold Crossing/Index Model

Heckman (1976) provides an articulated data generating process for the model implicit in the above and describes how to calculate the maximum likelihood estimator and conduct inference. Adapted to our setting, the microdata model he considers involves a latent data-generating process

$$Y^\star = \mu_0 + \theta T + \alpha' X + U \tag{1}$$

$$S^\star = \mu_S + \beta T + \gamma' X + \delta' Z + V \tag{2}$$

and observation equations, described below. Our primary parameter of interest is θ, the population difference in outcomes for those with $T = 1$ and $T = 0$.[7] We observe T, X, and Z for all observations, but do not observe Y^* or S^*. Instead, we observe Y and S, where

$$Y = \begin{cases} Y^* & \text{if } \; S = 1 \\ \text{missing} & \text{otherwise} \end{cases} \tag{3}$$

$$S = \mathbf{1}(S^* \geq 0) \tag{4}$$

Here, T is an indicator (for example, indicating MTO treatment), X is a vector of explanatory variables, and Z include factors affecting the probability of observation, but not the outcome under study. It is assumed that T, X and Z are independent of (U, V). In the textbook version of the one-sided selection model, it is also typically assumed that (U, V) are distributed bivariate normal with mean zero and a variance matrix which is (typically) normalized so that the variance of V is 1 (cf., Heckman, Tobias and Vytlacil 2003). However, it is widely understood that normality is not critical to identification, unless $\delta = 0$.[8] The restriction that Z not enter the outcome equation is referred to in the literature as an "exclusion restriction", using an analogy with the simultaneous equations model. Continuing

[7] This is a special case of the model in Section II, with $\mu_1(X, U_1) = \mu_1 + \theta + \alpha' X + U_1$, $\mu_0(X, U_0) = \mu_0 + \alpha' X + U_0$, $U = U_1 - U_0$, $\mu_S(X, T, Z) = \mu_S + \beta T + \gamma' X + \delta' Z$, and $V = U_S$.

[8] The participation rule is merely $V \geq -\mu_S - \beta T - \gamma' X - \delta' Z$. This is equivalent to setting $S = 1$ when $H(V) \geq H(-\mu_S - \beta T - \gamma' X - \delta' Z)$ for any strictly increasing function $H(\cdot)$. In the simplest case if $F(\cdot)$ is any symmetric absolutely continuous cumulative distribution function, then define $\widetilde{V} \equiv \Phi^{-1}(F(V))$ where $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative. One can then rewrite equation (2) as $S^* = \Phi^{-1}(F(\mu_S + \beta T + \gamma' X + \delta' Z)) + \widetilde{V}$ where $\widetilde{V}$ is normal. See Lee (1982, 1983) and Heckman, Tobias, and Vytlacil (2003). Even if $\delta = 0$, parametric knowledge of the distribution of the disturbances and functional forms is enough to identify the parameters of this model.

that analogy, requiring that $\delta \neq 0$ is akin to requiring that there exist a valid "first-stage" relationship, or that Z be "relevant" to observation.

The literature most related to our focus in this paper is the literature on nonparametric estimation of sample selection models (Heckman and Robb 1984, Ahn and Powell 1993, Das et al. 2003). Because this literature does not consider parametric restrictions, a maintained assumption is that Z has at least one element for which δ is not zero. A critical property of Z is that it have no independent effect on the observed outcome *except* through its effect on the probability of participation. If Z did not satisfy this property, then it would be related to U, leading to an identification problem. A truly ideal Z would be continuous and have broad enough support so that one could estimate the participation/selection across the entire range from 0 to 1. In the context of Example 1, for example, a continuous and exogenous Z could in principle allow the economist to estimate the entire (censored) mean function separately for those with $T = 0$ and $T = 1$, allowing characterization of the selection bias as well as estimation of the parameter of interest. Nonetheless, as we describe in Section IV, the binary variable arguably has a minimal necessary property: it is exogenous to the outcome equation residual.

We have outlined a constant coefficient framework with linearity, in which the utility of a valid exclusion restriction is easy to understand. In a slightly more general model, Das et al. (2003) establish the conditions for which it is possible to identify the conditional mean function non–parametrically in the presence of sample selection. The key assumption underlying their approach is additive separability of the selection correction term from the model for the observables. In our context, additive separability implies that for some continuous function $g(\cdot)$ we have

$$E[Y|X,T,Z,S=1] = \mu_0 + \theta T + \alpha' X + \eta g(s(X,T,Z)) \tag{5}$$

where $s(X,T,Z) \equiv P(S=1|X,T,Z)$ and η is the covariance between U and V.[9] Absent

[9] We acknowledge that this is a very strong assumption. See, for example, Little (1985).

such a condition, it would be impossible to separate the effect of the treatment from its effect on the probability of participation without difficult to justify parametric restrictions.[10]

C. Binary Treatments and Binary Hassling

We next discuss proactive strategies aimed at generating partial random assignment of the probability of observation. That is, in the context of equation (2), we discuss strategies for generating a variable Z that meets the assumptions required for identification. 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, and (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.

To keep the discussion as simple as possible initially, we will 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 has possible.
3. Compare the differences in means for the observed data.

In many studies for example, step 2 might involve different types of effort: making a home

[10] Vytlacil (2005) establishes that a large class of selection models take the form of a conditional expectation that can be represented as a function additively separable in the observables and the unobservable. If one interprets the participation condition as a difference between indirect utility functions, the key condition is that in the function determining participation, shifting the observable regressors in a specific direction changes the difference in the indirect utility functions in the same direction regardless of the level of the unobserved regressors (although the magnitude of the difference can vary with unobserved regressors.) See also Newey (1999), Powell (1994), and Das et al. (2003).

visit, calling on the telephone, and so on, possibly multiple times. For simplicity, consider two possible strategies: “intense effort to interview subject”, “less intense effort to interview subject” corresponding to say $Z = 1$ and $Z = 0$ respectively. The former method could involve making two phone calls to try to reach a participant, and the latter to making only one phone call.

Our proposal is 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 based on a binary indicator Z.
3. Based on their values of Z, employ either intense effort ($Z = 1$) or less intense effort ($Z = 0$) to collect the necessary outcome data. Note that this need not be more expensive than the standard practice if some persons require more effort and resources to interview than others. Our suggestion is simply to *randomize* the application of effort and resources.
4. Run the OLS regression suggested by equation (5) substituting estimated values for the inverse Mills ratio (IMR) instead of their unknown values.

This is the two-step estimator for microdata (Heckman 1976) and is asymptotically equivalent to the maximum likelihood estimator that assumes joint normality.

To understand the logic of this procedure, note that under random assignment of T, it is not necessary to model the effect of X on Y in equation (5). Consider, then, conditioning on T and Z alone, noting that this leads to four pairs of salient *observed* means, namely (censored) mean outcomes for those with $T = j$ and $Z = k$ and the probability of observation for those with $T = j$ and $Z = k$, for $j, k \in \{0, 1\}$.

Example 2 provides a hypothetical data configuration under such a setting. The salient treatment effect is the vertical height between the two lines, *at the same survivor probability*. In Example 1, where those with $T = 1$ were compared to those with $T = 0$, it was not possible to disentangle sample selection from treatment—intuitively, there was no way to estimate the slope of the (censored) mean in the probability of observation. Example 2 reflects a different study design. Here, estimation of the slope is possible because (i) there are two subgroups on the same (censored) mean function, and (ii) the two subgroups are generated by random assignment, enabling the economist to estimate a pure sample selection from a simple comparison of the difference in outcomes induced by hassling, among those with the same treatment status ($T = 0$ or $T = 1$).

Under additive separability and joint normality, the slopes of both lines are equal (and proportional to $\partial \mathrm{IMR}/\partial s < 0$, where s is the probability of survival, or non-attrition). This is shown clearly in Example 3, where the units on the horizontal axis have been transformed to inverse Mills ratio space (note that when the probability of observation is one, the Mills ratio is zero, and that the Mills ratio increases as the probability of observation falls). Because the lines in Example 3 are parallel under joint normality, only 3 of the 4 means are needed to identify θ, even if the probability of observation is very different for the four pairs. This is the intuition behind the tests of overidentification that we describe below.

Under additive separability, but with non-normal distributions, the curves will continue to be parallel but will not be linear in inverse Mills ratio space. The curves will be parallel in $g(s)$ space (cf., equation (5)), but this function is unknown. These identification difficulties are depicted in Example 4, where the distribution for the residual in the observation equation is taken to be Laplacian with location parameter 0 and scale parameter 4.[11] The horizontal axis in Example 4 is *not* taken to be the correct control function, but remains the inverse Mills ratio. This is analogous to the expected situation, in which the economist does not know the true data generating process, but imposes a distributional assumption to overcome the fact that the probability of observation is different between the treatment and control groups. (If the correct control function were included, the curves would be both parallel and linear.)

Observing the 4 pairs depicted in Example 4, an economist who believed strongly in joint normality of the errors would conclude that there was a failure of additive separability,

[11] See Heckman, Tobias and Vytlacil (2000). The counterfactual outcome equation residuals continue to be distributed bivariate normal. This leads to a control function of the form $g(s) = \phi(\Phi^{-1}(F(\Phi^{-1}(s))))/F(\Phi^{-1}(s))$, where $F(\cdot)$ is the Laplacian distribution function.

since the lines would not appear parallel. An economist who believed strongly in additive separability would instead be skeptical of the joint normality assumption, since the curves would not appear to be lines.

The figure thus highlights that even in the near-ideal case we have outlined, where the probability of observation is partially randomized by a variable Z, inference is difficult when (i) the distribution of the error terms is unknown, and (ii) there are differential rates of observation for those in the treatment and control groups. A caveat to this pessimism regards the support of the instrument for observation. When Z is binary, departures from linearity in the (censored) mean function given the IMR cannot be detected. However, as the support of Z increases, detecting nonlinearity becomes more feasible. Nonetheless, broad support for Z is not sufficient for identification unless additive separability holds. Without additive separability, the only way to identify the population mean is if, for some z, the probability of observation conditional on $Z = z$ is 1 (Heckman and Honore 1990, Heckman 1990).

We now formalize tests for overidentification, assuming additive separability and joint normality. We begin by setting notation. Let the sample analogue of $E[Y|T = j, Z = k]$ be denoted by $\widetilde{Y}_{jk}$. Let the probability of being observed, $E[S|T = j, Z = k]$, be denoted by S_{jk}, and let the sample analogue be denoted by $\overline{S}_{jk}$. Graphing the 4 pairs $(\widetilde{Y}_{jk}, \overline{S}_{jk})$ is informative for several reasons. First, doing so will clarify whether the probability of observation is equal between $T = 1$ and $T = 0$ groups with the same level of follow–up (i.e., those with either $Z = 0$ or $Z = 1$). Under additive separability, for either $k = 0$ or $k = 1$, if $\overline{S}_{1k} = \overline{S}_{0k}$, then $\widetilde{Y}_{1k} - \widetilde{Y}_{0k}$ is consistent for θ, even without normality. If $\overline{S}_{1k} = \overline{S}_{0k}$ for $k = 0$ and $k = 1$, then there is no sample selection problem (under the one-sided selection model; for discussion, see Lee 2005). In that case, the best estimator is simply the difference in sample means, or a weighted average of $\widetilde{Y}_{11} - \widetilde{Y}_{01}$ and $\widetilde{Y}_{10} - \widetilde{Y}_{00}$. Second, given additive separability, under no distribution can it occur that the slope of the (censored) wage curve is positive for one group and negative for another. Though we do not pursue the idea here, this suggests the possibility of a test for additivity separability with assumptions weaker than those required for the jointly normal Heckman model.

Nonetheless, inspection of the 4 pairs will allow for informal inference. If the slopes appear to be opposite in sign, and the economist is confident that additive separability is violated, it is not straightforward how to proceed except to bound the treatment effect of interest.[12]

In the case of a binary effort level, a simple estimator suggested by our graphical display involves equating 4 moments to 3 parameters (in the case of only one binary treatment):

$$\begin{aligned}\widetilde{Y}_{11} - \mu_0 - \theta - \eta g(\overline{S}_{11}) &= 0\\ \widetilde{Y}_{10} - \mu_0 - \theta - \eta g(\overline{S}_{10}) &= 0\\ \widetilde{Y}_{01} - \mu_0 - \eta g(\overline{S}_{01}) &= 0\\ \widetilde{Y}_{00} - \mu_0 - \eta g(\overline{S}_{00}) &= 0\end{aligned}$$

where we have now specialized $g(s) = \phi(\Phi^{-1}(s))/s$.[13] In the case where we can ignore the sampling error in the various means, this amounts to a regression with four observations

$$\widetilde{Y}_j = \mu_0 + \theta T_j + \eta g(\overline{S}_j) \tag{6}$$

where j now indexes the four groups in Example 2 through 4. This suggests an estimator that solves the minimization problem:

$$\min_{\mu_0,\theta,\eta} \;\; e'\widehat{\Omega}e \tag{7}$$

where $e = (e_1, e_2, e_3, e_4)$, $e_j = \widetilde{Y}_j - \mu_0 - \theta T_j - \eta g(\overline{S}_j)$, and $\widehat{\Omega}$ is an estimate of the variance

[12] As discussed below, Manski bounds are available to bound the intent-to-treat parameter described. These will be unbounded if the support of the outcome is unbounded. Lee bounds are finite, even if the outcome is unbounded, but are only available if the economist is willing to invoke the one-sided selection model.

[13] 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).

matrix of e. The minimized value in (7) suggests itself as a test of over–identification, distributed χ^2 with degrees of freedom equal to the number of moments less the numbers of parameters being estimated. In the case we have been discussing, this is 1. This amounts to a joint test of additive separability and joint normality of (U, V).

In the case of a trichotomous treatment, as characterizes the MTO demonstration, this 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 the analogous overidentification statistic is distributed χ^2 with 2 degrees of freedom. In this paper we implement the test using a variance matrix calculated using the "delta method" (i.e. using simple 2nd order Taylor series approximations). It was comforting to observe that in several monte carlo experiments using a normal–normal Heckman DGP 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.

D. 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. In the bounds we consider in this section, both allow for some treatment effect heterogeneity.

(i) *Horowitz and Manski (2000)*

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 $\overline{Y}$ 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

$$\begin{aligned} \overline{\theta}_M &= P[S=1|T=1]E[y|T=1] + (1 - P[S=1|T=1])\overline{Y} \qquad (8)\\ &- P[S=1|T=0]E[y|T=0] + (1 - P[S=1|T=1])\underline{Y} \\ \underline{\theta}_M &= P[S=1|T=1]E[y|T=1] + (1 - P[S=1|T=1])\underline{Y} \qquad (9)\\ &- P[S=1|T=0]E[y|T=0] + (1 - P[S=1|T=1])\overline{Y} \end{aligned}$$

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

(ii) *Lee (2005)*

The next set of bounds we consider also allow for treatment effect heterogeneity and require no valid exclusion restriction. However, they do involve additional assumptions. 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. To be completely clear, consider the following table:

Type			$T=0$	$T=1$
g_{11}	$S_0=1$	$S_1=1$	Observed	Observed
g_{01}	$S_0=0$	$S_1=1$	Not Observed	Observed
g_{10}	$S_0=1$	$S_1=0$	Observed	Not Observed
g_{00}	$S_0=0$	$S_1=0$	Not Observed	Not Observed

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 will be different than when we modeled the attrition. For example, we will be unable to learn about individuals of type g_{00}—that is, those for whom $(S_0 = 0, S_1 = 0)$, or persons who attrit or generally fail to respond regardless of whether they have been assigned to treatment or control.

Instead of a selection equation like equation (2) above, we have

$$S = S_1\ T + S_0(1 - T) \tag{2a}$$

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

With two familiar assumptions it is possible to bound the treatment effect for individuals of type g_{11}, or the set of individuals who we will always observe. The first assumption is random assignment of the treatment, T. The second—and more important—assumption is what Lee (2005) calls "monotonicity". Specifically, the assumption is that *either* $S_1 \geq S_0$ for all individuals *or* $S_0 \geq S_1$ for all individuals. In terms of our table above, this means ruling out individuals of type g_{10}, or ruling out individuals of type g_{01}. In words, assignment to treatment can only affect attrition in one direction: for example, individuals may be induced to exit by assignment to treatment, but there can not be individuals who would have exited but were induced to stay by assignment to treatment. Interestingly, Vytlacil (2005) has shown that this monotonicity condition is equivalent to assuming that the data generating process is of the latent variable index/threshold crossing model variety that we considered in subsection B, above.

The bounds are for the Average Treatment Effect for the sub-population of always survivors (non–attriters). To illustrate, assume that $S_1 > S_0$, and denote the average treatment effect by θ. This means that the fraction of attriters is lower in the treatment group than in the control. Under the monotonicity assumption, this means that the observed treatment individuals are a combination of two types g_{11} (always survivors) and g_{01} – types we wouldn't have seen if they had been in the control but whom we happen to observe because they were assigned to treatment. By assumption, there are no persons of type g_{10} – types who we would have observed if they had been in the control group, but wouldn't have been observed if they had been assigned to treatment, and hence the controls contain only individuals of type g_{11}.

In this case, Lee's bounds are given by:

$$\begin{aligned}
\underline{\theta}_L &\equiv E[Y|T=1,S=1,Y\leq \mathcal{Y}_{(1-p_0)}]-E[Y|T=0,S=1] \\
\overline{\theta}_L &\equiv E[Y|T=1,S=1,Y\geq \mathcal{Y}_{(p_0)}]-E[Y|T=0,S=1], \quad \text{where} \\
\mathcal{Y}_{(p_0)} &\equiv G^{-1}_{S=1,T=1}(p_0) \\
p_0 &\equiv \frac{P[S=1|T=1]-P[S=1|T=0]}{P[S=1|T=0]}
\end{aligned}$$

and $G^{-1}_{S=1,T=1}(\cdot)$ is the inverse cumulative distribution function given $S=1$ and $T=1$.

To illustrate, suppose that 50 percent of the treatment group has not attrited, but that only 40 percent of the control group remain. 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 $p_0=\frac{0.5-0.4}{0.5}=\frac{1}{5}=0.2$. 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 percent 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 $\widehat{\overline{\theta}}_L$.

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 (2005) shows that it is also possible to tighten the bounds using covariates. The tightening induced by a specific covariate, X, is increasing in the difference in the trimming fractions given differing values x in the support of X. This implies that an ideal covariate for tightening Lee bounds is a variable that predicts observation. An ideal method for producing tight bounds, then, is to utilize data collection procedures that specifically manipulate the probability of observation. Thus, the modifications to standard data collection procedures we advocate in this paper are useful both for point identification and for improving the tightness of Lee bounds.

The procedure that uses a binary variable, Z, to tighten bounds can be described simply. Suppose that there is less attrition in the treatment group. Following the notation used throughout the paper, let $Z = 1$ refer to the case where the respondent has been the subject of additional effort on the part of the survey team. Lee's procedure involves calculating two bounds (one for the observations such that $Z = 0$, and another for the observations such that $Z = 1$) and averaging them.

Using notation analogous to the above, define

$$\begin{aligned}
\overline{\theta}_L^{Z=1} &\equiv E[Y|T=1, S=1, Z=1, Y \geq \mathcal{Y}_{(p_0)}^{Z=1}] - E[Y|T=0, S=1, Z=1] \quad \text{where} \\
\mathcal{Y}_{(p_0)}^{Z=1} &\equiv G_{S=1,T=1,Z=1}^{-1}(p_0) \\
p_0^{Z=1} &\equiv \frac{P[S=1|T=1, Z=1] - P[S=1|T=0, Z=1]}{P[S=1|T=0, Z=1]}
\end{aligned}$$

and define the upper bound for those with $Z = 0$, denoted $\overline{\theta}_L^{Z=0}$, analogously. Let q_1 denote $P(Z = 1|T = 1)$. Then the treatment effect upper bound for the whole sample is given by:

$$q_1\overline{\theta}_L^{Z=1} + (1 - q_1)\overline{\theta}_L^{Z=0}$$

and this upper bound is smaller than the unconditional upper bound $\overline{\theta}_L$ (for proof, see Lee 2005). The average conditional lower bound is defined analogously and is larger than the unconditional lower bound. These bounds may be estimated in the same way as outlined in the 3 step algorithm above.

An interesting point is that the population bounds as well as the sample bounds are tightened by use of covariates. That is, we have $q_1\overline{\theta}_L^{Z=1} + (1-w_1)\overline{\theta}_L^{Z=0} \leq \overline{\theta}_L$, as well as $\widehat{q}_1\widehat{\overline{\theta}}_L^{Z=1} + (1-\widehat{q}_1)\widehat{\overline{\theta}}_L^{Z=0} \leq \widehat{\overline{\theta}}_L$, and analogously with the lower bounds. The inequality becomes strict when the trimming fractions are different for those with $Z = 1$ and $Z = 0$.

Finally, it should be noted that it is straightforward to turn this into a confidence interval. Let z_q denote the qth-quantile for the standard normal distribution and let $\widehat{\sigma}_{\widehat{\overline{\theta}}_L}$ and $\widehat{\sigma}_{\widehat{\underline{\theta}}_L}$ denote the standard errors for the bounds described in Lee (2005). Lee shows shows that the naive confidence interval

$$\widehat{\underline{\theta}}_L + \widehat{\sigma}_{\widehat{\underline{\theta}}_L} z_{\alpha/2} \quad , \quad \widehat{\overline{\theta}}_L + \widehat{\sigma}_{\widehat{\overline{\theta}}_L} z_{1-\alpha/2}$$

has an asymptotic coverage probability of $1-\alpha$ for the *parameter region* $[\underline{\theta}_L, \overline{\theta}_L]$. This is distinct from an asymptotic coverage probability for the *parameter* that lies in the region $[\underline{\theta}_L, \overline{\theta}_L]$. Imbens and Manski (2004, cited in Lee) observe that the interval given above covers the parameter $E[Y|T = 1, S_0 = 1, S_1 = 1] - E[Y|T = 0, S_0 = 1, S_1 = 1]$ (which, under the model given in equations (1)-(4), is equal to θ) with a *greater* probability than $1-\alpha$ (asymptotically). Intuitively, using standard errors for the bound point estimates results in some "double-counting" in forming a confidence region for the parameter. A $1-\alpha$ percent confidence interval for the *parameter* of interest is given by

$$\widehat{\underline{\theta}}_L - \widehat{\sigma}_{\widehat{\underline{\theta}}_L} c_n \quad , \quad \widehat{\overline{\theta}}_L + \widehat{\sigma}_{\widehat{\overline{\theta}}_L} c_n$$

where c_n solves

$$\Phi\left(c_n + \frac{\widehat{\overline{\theta}}_L - \widehat{\underline{\theta}}_L}{\max\{\widehat{\sigma}_{\widehat{\underline{\theta}}_L}, \widehat{\sigma}_{\widehat{\overline{\theta}}_L}\}}\right) - \Phi(-c_n) = 1-\alpha$$

One may shown that c_n is less than $\Phi^{-1}(1-\alpha/2)$ but no smaller than $\Phi^{-1}(1-\alpha)$. When $\alpha = 0.05$, these quantities are equal to the familiar numbers 1.96 and 1.64, respectively,

IV. The Moving To Opportunity Experiment

The Moving to Opportunity (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), and Goering and Feins (2003). 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—due to the overall quality of the MTO evaluation, response rates at follow–up are a high 90 percent 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 which assume that the data are missing at random.[14]

[14] 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 non-random attrition. Orr, Feins, Jacob, Beecroft, Sanbonmatsu, Katz, Liebman and Kling (2003, Appendix 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.

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.[15] 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 percent poor, augmented by some counseling; a second 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 has addressed this problem by re-weighting observations on subjects according to $\widehat{P}(T_i = t)/\widehat{P}(T_i = t|R_i)$ 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 re-weighting from the propensity score literature, adjusted for trichotomous treatment assignments.[16]

[15] Eligible projects were selected by local public housing authorities from among housing projects in census tracts with at least 40 percent poor (Goering, Kraft, Feins, McInnis, Holin and Elhassan 1999).

[16] Assuming binary treatment and no covariates for simplicity, the estimator used throughout MTO evaluation research is (cf., Orr et al., 2003).

$$\widehat{\theta}_{MTO} \equiv \frac{\sum_{i=1}^{n} Y_i T_i \frac{1}{\widehat{p}_i}}{\sum_{i=1}^{n} T_i \frac{1}{\widehat{p}_i}} - \frac{\sum_{i=1}^{n} Y_i (1-T_i) \frac{1}{1-\widehat{p}_i}}{\sum_{i=1}^{n} (1-T_i) \frac{1}{1-\widehat{p}_i}} \equiv \widehat{E}_1 - \widehat{E}_0$$

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.

To maintain the simplest possible treatment 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.[17] By excluding individuals assigned in different treatment regimes we circumvent the need for weighting and our analysis can proceed without covariates. The individuals we analyze comprise roughly 1700 of the roughly 3500 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 non-respondents would not be as effective as targeting effort at a subset of non-respondents. MTO administrators selected 3 out of 10 non-respondents 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,

where $\widehat{p}_i$ is the estimated propensity score. This is different from the average treatment effect estimator from the propensity score literature

$$\widehat{\theta} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i T_i \frac{1}{\widehat{p}_i} - \frac{1}{n}\sum_{i=1}^{n} Y_i (1-T_i)\frac{1}{1-\widehat{p}_i}$$

Both estimators are consistent, but $\widehat{\theta}_{MTO}$ is inefficient, because

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \widehat{\theta}_{MTO}) &= \widehat{E}_1\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_i\frac{1}{\widehat{p}_i} - 1\right) - \widehat{E}_0\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\frac{1}{1-\widehat{p}_i} - 1\right) \\ &\xrightarrow{d} \text{plim } \widehat{E}_1 N\left(0, E\left[\frac{1}{\widehat{p}_i}\right] - 1\right) - \text{plim } \widehat{E}_0 N\left(0, E\left[\frac{1}{1-\widehat{p}_i}\right] - 1\right) \qquad (11)\end{aligned}$$

is a non-degenerate distribution and $\widehat{\theta}$ is efficient (Hirano, Imbens and Ridder 2003).

[17] 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).

and New York (Orr et al. 2003).

Because the family identifier is a baseline characteristic of individuals, we interpret this procedure as specifying that 3 out of 10 individuals were randomly selected by MTO administrators for intensive follow–up. 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 that (1) 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 $Z_i = 0$. These attributes make Z_i a natural candidate for an instrument for the observation equation.

V. Empirical Results

A. Implementation

For ease of interpretation we first present 4 sets of figures and 4 sets of tables corresponding to the four outcomes for which we have administrative data. Figure 1 and Table 1 report on our results using Fraction of Quarters Employed; Figure 2 and Table 2 do the same for Annualized Earnings, Figure 3 and Table 3 for TANF receipt, Figure 4 and Table 4 for TANF amount.

For each outcome we display a figure motivated by our (joint) test for additive separability and normality. Using our earlier notation, let $\widetilde{Y}_{jk}$ denote the censored sample means that correspond to $E[Y|T = j, Z = k]$ and let $\overline{S}_{jk}$ denote the sample probability of being observed, $E[S|T = j, Z = k]$, where the first subscript denotes whether the observation is from the control, experimental, or Section 8 ($j = 0, 1, 2$) respectively) and the second subscript ($k = 0, 1$) denotes whether the individual was randomized into non–intensive follow–up ($Z = 0$) or intensive follow–up ($Z = 1$) respectively. Finally, we define $\widetilde{M}_{jk} = \phi(\Phi^{-1}(\overline{S}_{jk})/\overline{S}_{jk})$, the estimated value of the inverse Mills ratio term computed using $\overline{S}_{jk}$.

Our proposed test rejects when the minimized value given in equation (7) is sufficiently far from zero. The test is equivalent to a test of the equality of the slopes of the conditional outcome means in the respective inverse Mills ratio terms. The intuition for this is that under additive separability we would expect

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

The figures we present display the key summary statistics needed to compute our test for additive separability. Under joint normality, the horizontal axis is appropriately measured in “inverse Mills ratio units”, in which case the 3 slopes could be read directly from the figure. However, given that joint normality is not necessarily affected, and given that over the range observed, the inverse Mills ratio is approximately linear, we choose to measure the horizontal axis in terms of the probability of observation.

The figures present

1. Separately for $Z = 0$ and $Z = 1$, the (observed) means for the control group, the experimental group, and the Section 8 group, or $\widetilde{Y}_{00}$, $\widetilde{Y}_{01}$, $\widetilde{Y}_{10}$, $\widetilde{Y}_{11}$, $\widetilde{Y}_{20}$, and $\widetilde{Y}_{21}$, plotted against their respected probability of being non–missing $\overline{S}_{00} \ldots \overline{S}_{21}$.
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.

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

1. Estimated program impacts using the administrative data, but restricted to the subsample of observations with non-missing 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.
2. Estimated (unconditional) upper and lower bounds for the treatment effects allowing for non–random selection as suggested by Lee (2005). Under monotonicity these estimates are consistent estimates of upper and lower bounds of the program impacts in the presence of non–random selection.

3. Estimated program impacts with the conventional normal Heckman two–step estimator, where in the first step we compute an estimate of the probability of observation with a probit and then use the estimated inverse Mills ratio in the second step. We also provide the test of the hypothesis that randomization into more extensive follow–up had no effect on the probability of being observed in the survey data. The test statistic is labelled "the χ^2 test statistic on the exclusion of instruments."
4. Estimated program impacts using the maximum likelihood version of the Heckman estimator.
5. Estimated program impacts using the (complete) administrative data.

Each step represents various solutions to the missing data problem that are routinely used in the literature along with the ideal (complete data) estimator. We do not display the Manski–Horowitz bounds for our two bounded outcomes since these bounds are substantially larger than the Lee bounds we present and we can not rule out very large negative or positive treatment effects.[18]

Table 5 summarizes the key summary statistics for our two–step test for additive separability under normality.

B. Fraction of Quarters Employed and Annualized Earnings

Table 1 presents the results of our analysis for the outcome "fraction of quarters employed, years 1 to 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.

Beginning first with our graphical analysis, Figure 1 presents the key descriptive statistics. It is evident that the more intensive follow–up producedure 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 only be obtained with more extensive follow–up. That is, the evidence suggests negative selection; mean outcomes for those with $Z = 1$ are all less than for those with $Z = 0$.

[18] The Manski Horowitz bounds for the experimental treatment effect for Fraction of Quarters employed are (-0.161, 0.162) and (-0.165, 0.157) for fraction receiving TANF. The Section 8 treatment effect bounds are (-0.173, 0.140) and (-0.187, 0.126), respectively.

In the final row of the table, we display the point estimates for the experimental group and the treatment group. These are 0.001 and -0.018, respectively, using the full data (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 solutions to the selection bias problem, the first row displays estimates using just the survey sample. In all cases these are well within sampling variability of the full (administrative) data estimates. In the second row of the table, we display the bounds suggested by Lee. As might be expected under a one-sided selection model, they provide intervals that are somewhat larger and comfortably include the full sample point estimates.

In the third row, we present the estimates using the two–step Heckman selection procedure. 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.

Despite strenuous efforts that included profiling the likelihood and trying various transformations, we were unable to successfully maximize the likelihood function that assumes joint normality.[19] We note that this problem occurred despite the availability of a strong instrument for observation. The reasons for failure to achieve convergence are not clear. In extensive Monte Carlo experimentation with generated data from an ideal sample selection model, STATA's `heckman` routine would occasionally fail to converge. A possible source of the problem is that the estimates of the three slopes (which are displayed in the first row of Table 5) are quite imprecisely estimated. It is also interesting to observe that the point estimates are not consistent with monotonicity. The complete data sample means are all *higher* than the means for the "hard to get" group but about the same level for those persons

[19] Practical difficulties with the Heckman MLE are reported in, for example, StataCorp (2003).

interviewed without need for extensive follow–up.

In Figure 2 and Table 2, we present the results for Annualized Earnings. The pattern of results is quite similar to those for fraction of quarters employed. Large standard errors prevent any reliable inference. Again, both the Heckman two-step and Lee bounds provide similar inferences as the complete (administrative) data as well as the potentially biased OLS estimates using only the survey data.

C. TANF Results

Figure 3 and 4 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,[20] 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.[21]

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.

The Heckman two-step and MLE estimator perform quite well, with standard errors only modestly larger than their full sample counterparts 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,

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

[21] 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.

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 randomization of non–response. We propose a simple graphical analysis that may be useful in detecting the potential contamination arising from selection bias as well as a simple test for the key additive separability assumption implied by the one-sided selection model. We implement these proposals on a well–known and well–conducted experiment, the Moving To Opportunity demonstration, for a subset of outcomes for which complete (administrative) data is available.

Our proposal provides a simple, low–cost way of assessing the validity of the missing-at-random assumption frequently invoked in the literature. The proposal is also expected to allow for considerable narrowing of Lee bounds. These bounds will be most useful in settings where the additive separability assumption appears to be corroborated by the data. We have discussed how our graphical approach enables informal inference regarding the plausibility of additive separability.

There are several important questions we did not address in our analysis. For example, we have not yet calculated Lee bounds conditional on Z. Also, we did not consider the implications of the use of partial randomization of follow–up for the design of experiments. For example, if the cost and efficacy of various follow–up technologies are known, it should be possible to organize data collection efforts to maximize the information available given a fixed cost outlay.[22] Finally, we also proposed a fully nonparametric test of additive separability, but have not yet derived the best implementation of this test, or calculated appropriate critical values.

[22] In medical RCTs of chronic pain, for example, attrition often ranges from 40 to 60 percent. Typical practice to incorporate the likely non–response in the power calculation, assuming that the missing data are missing at random. Such procedures could and should be altered to reflect the impact of missing data on inferences as well as the researchers ability to mitigate the harm done by missing data by partially randomizing the probability of observation.

References

Ahn, Hyungtaik and James L. Powell, “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism,” *Journal of Econometrics*, July 1993, *58* (1-2), 3–29.

Altman, Douglas G., Kenneth F. Schulz, David Moher, Matthias Egger, Frank Davidoff, Diana Elbourne, Peter Getzche, and Thomas Lang, “The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration,” *Annals of Internal Medicine*, 17 April 2001, *134* (8), 663–694.

Bureau of Labor Statistics and Census Bureau, “Current Population Survey: Design and Methodology,” *Census Bureau Technical Paper 63RV*, March 2002.

Census Bureau, “Survey of Income and Program Participation Users’ Guide: Supplement to the Technical Documentation,” 2001. Third Edition.

——— , “CPS Basic Monthly Survey: Nonresponse Rates,” 2005. Web page accessed 11/30/2005: `http://www.bls.census.gov/cps/basic/perfmeas/typea.htm`.

Das, Mitali, Whitney Newey, and Francis Vella, “Nonparametric Estimation of Sample Selection Models,” *Review of Economic Studies*, January 2003, *70* (1), 33–58.

Goering, John and Judith D. Feins, *Choosing a Better Life? Evaluating the Moving to Opportunity Social Experiment*, Washington, D.C.: Urban Institute Press, 2003.

——— , Joan Kraft, Judith D. Feins, Debra McInnis, Mary Joel Holin, and Huda Elhassan, “Moving to Opportunity for Fair Housing Demonstration Program,” September 1999. Unpublished manuscript. U.S. Department Housing and Urban Development.

Gronau, Reuben, “The Effect of Children on the Housewife’s Value of Time,” *Journal of Political Economy*, March-April 1973, *81* (2), S168–S199.

——— , “Wage Comparisons—A Selectivity Bias,” *Journal of Political Economy*, November-December 1974, *82* (6), 1119–1143.

Grovers, Robert M. and Mick P. Couper, *Nonresponse in Household Interview Surveys*, New York: John Wiley and Sons, 1998.

Heckman, James J., “Shadow Prices, Market Wages, and Labor Supply,” *Econometrica*, July 1974, *42* (4), 679–694.

——— , “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*, 1976, *5* (4), 475–492.

——— , “Sample Selection Bias as a Specification Error,” *Econometrica*, January 1979, *47* (1), 153–162.

——— , "Selection Bias and Self-Selection," in P. Newman, M. Milgate, and J. Eatwell, eds., *The New Palgrave—A Dictionary of Economics*, New York: Macmillan, 1987, pp. 287–296.

——— , "Varieties of Selection Bias," *American Economic Review*, May 1990, *80* (2), 313–318.

——— and Bo E. Honore, "The Empirical Content of the Roy Model," *Econometrica*, September 1990, *58* (5), 1121–1149.

——— and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

——— and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1984.

——— , Justin Tobias, and Edward Vytlacil, "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling," September 2000. Unpublished manuscript, University of Chicago.

——— , ——— , and ——— , "Simple Estimators for Treatment Parameters in a Latent Variable Framework," *Review of Economics and Statistics*, August 2003, *85* (3), 748–755.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

Hirsch, Barry T. and Edward J. Schumacher, "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, 2004, *22* (3), 689–722.

Horowitz, Joel L. and Charles F. Manski, "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, May 1998, *84* (1), 37–58.

——— and ——— , "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 2000, *95*, 77–84.

——— and ——— , "Rejoinder: Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 2000, *95*, 87.

Imbens, Guido and Charles F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, November 2004, *72* (6), 1845–1857.

Juni, P., D. G. Altman, and M. Egger, "Assessing the Quality of Controlled clinical Trials," *British Medical Journal*, 7 July 2001, *323*, 42–46.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, May 2001, *116* (2), 607–654.

Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *Quarterly Journal of Economics*, February 2005, *120* (1), 87–130.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," October 2005. Unpublished manuscript, University of California, Berkeley.

Lee, Lung-Fei, "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 1982, *49* (3), 355–372.

____ , "Generalized Econometric Models with Selectivity," *Econometrica*, 1983, *51* (2), 507–512.

Lewis, H. Gregg, "Comments on Selectivity Biases in Wage Comparisons," *Journal of Political Economy*, 1974, *82* (6), 1145–1155.

Little, Roderick J.A., "A Note About Models for Selectivity Bias," *Econometrica*, November 1985, *53* (6), 1469–1474.

Manski, Charles F., "Anatomy of the Selection Problem," *Journal of Human Resources*, 1989, *24*, 343–360.

____ , "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 1990, *80*, 319–323.

____ , "The Selection Problem," in Christopher Sims, ed., *Advances in Econometrics Sixth World Congress*, number 23. In 'Econometric Society Monographs.', Cambridge: Cambridge University Press, 1994, chapter 4, pp. 143–170.

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

Moher, David, Alison Jones, and Leah Lepage, "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*, 18 April 2001, *285* (15), 1992–1995.

Newey, Whitney K., "Two Step Series Estimation of Sample Selection Models," 1999. Unpublished manuscript, Massachusetts Institute of Technology.

Orr, Larry, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling, "Moving to Opportunity Interim Impacts Evaluation," *Final Report, U.S. Department of Housing and Urban Development*, 2003.

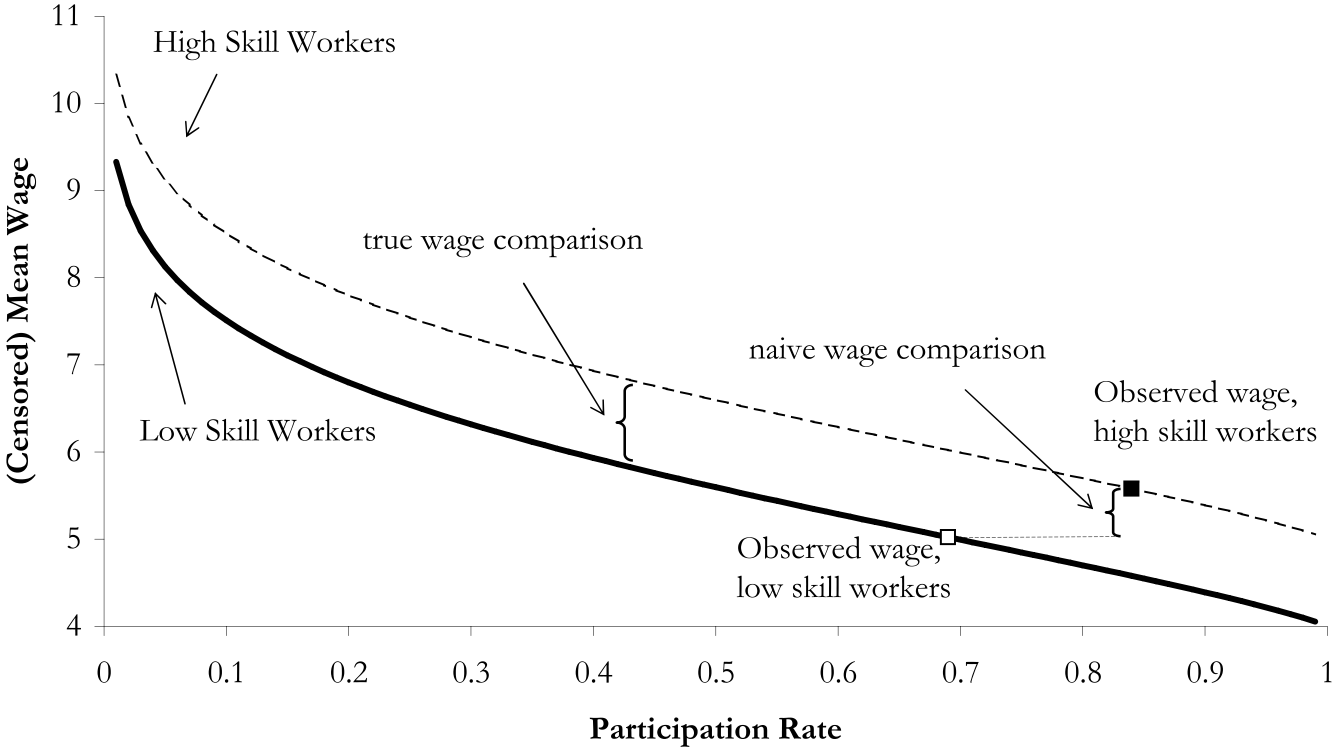
Powell, James L., "Estimation of Semiparametric Models," in Robert F. Engle and Daniel McFadden, eds., *Handbook of Econometrics*, Vol. 4, New York: North–Holland, 1994.

Rubin, Donald B., *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons, 1987.

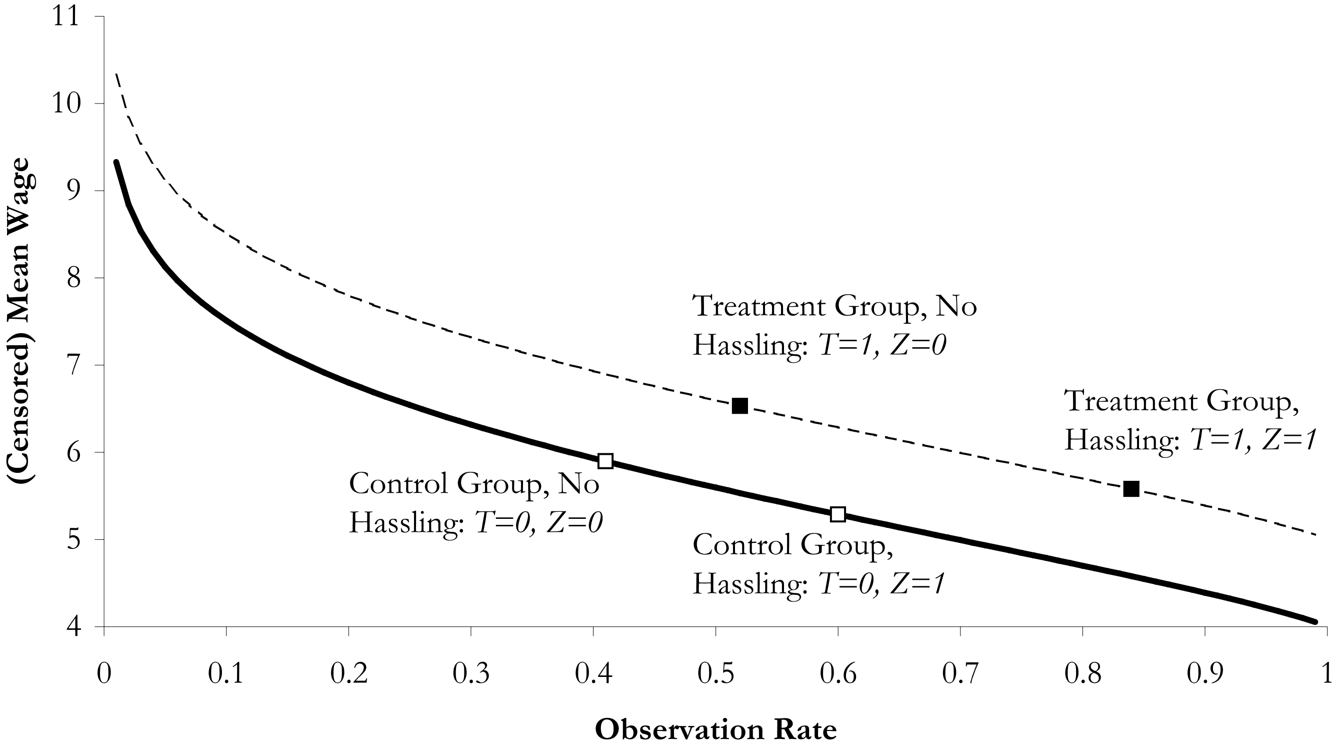
StataCorp, *Stata Statistical Software: Release 8.0: Reference*, College Station: State Corporation, 2003.

Vytlacil, Edward, "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," Unpublished Manuscript, Columbia University June 3 2005.

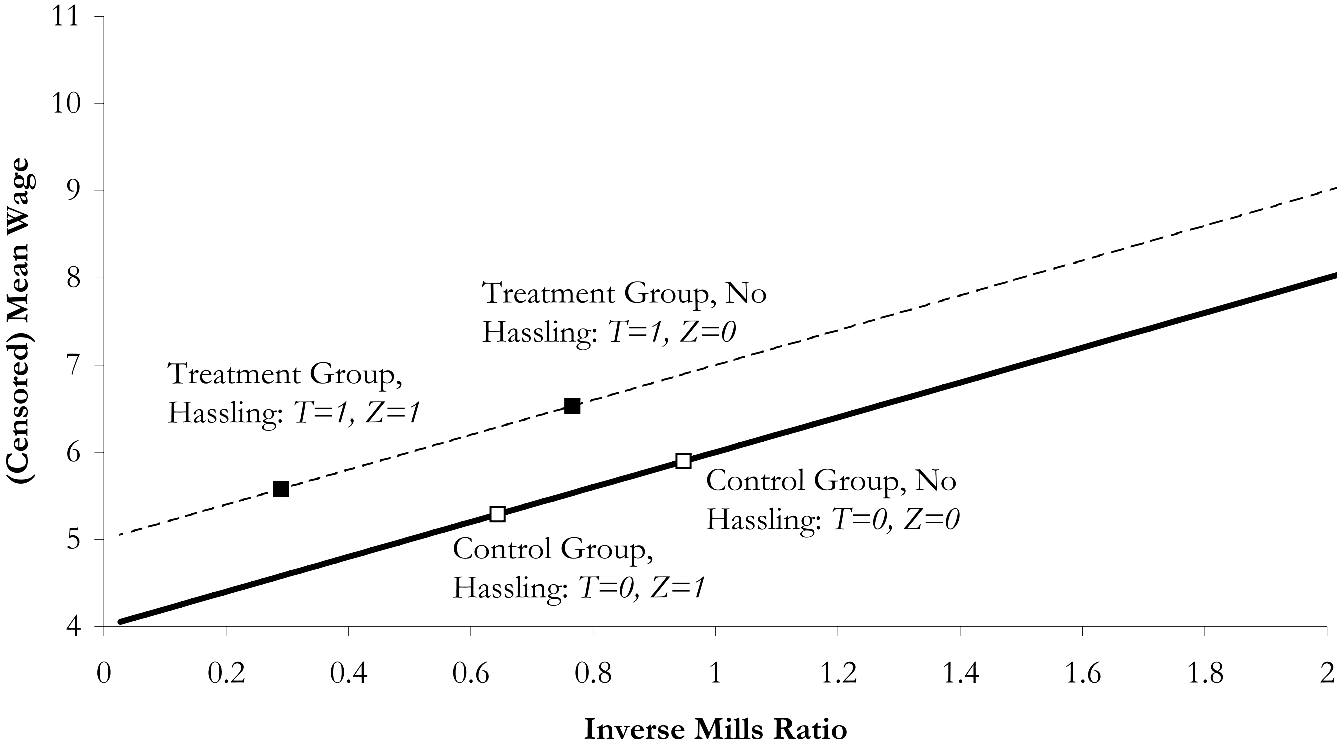
Example 1. Wage Comparisons and Labor Supply in the One-Sided Selection Model



Example 2. Estimating Treatment Effects in the Presence of Sample Selection with Binary Hassling



Example 3. Sample Selection Correction under Additive Separability: Joint Normality of All Residuals



Example 4. Sample Selection Correction under Additive Separability: Observation Equation Error Distributed as Double Exponential (Laplacian)

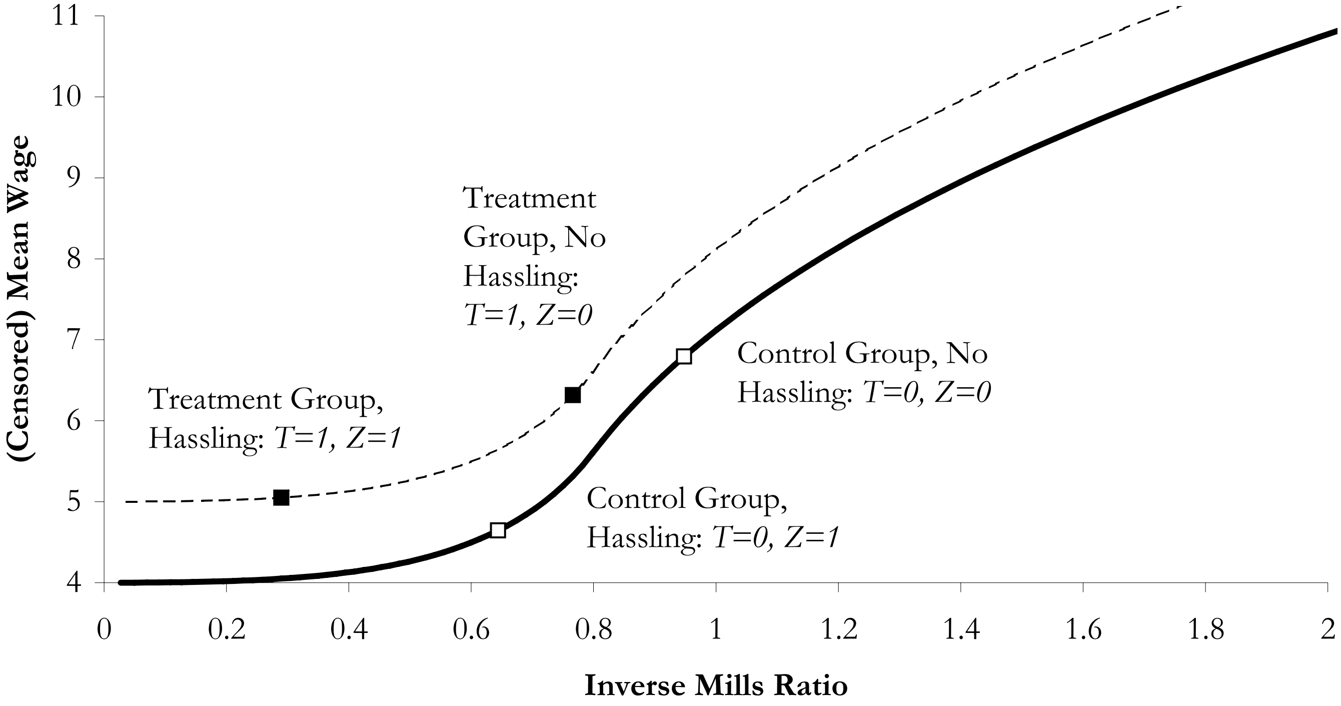


Figure 1. Graphical Interpretation of Sample Selection Correction: Fraction of Quarters Employed, Years 1 to 4

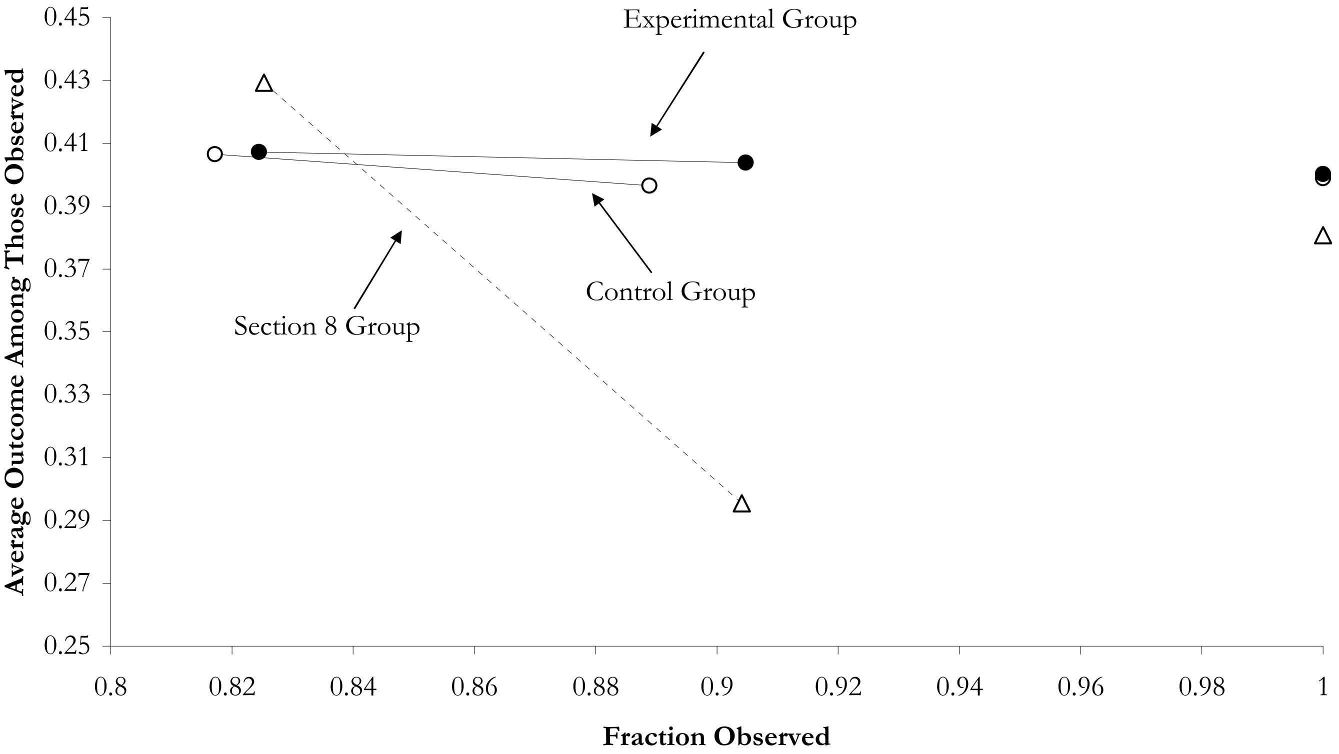


Figure 2. Graphical Interpretation of Sample Selection Correction: Annualized Earnings for Years 1 to 4

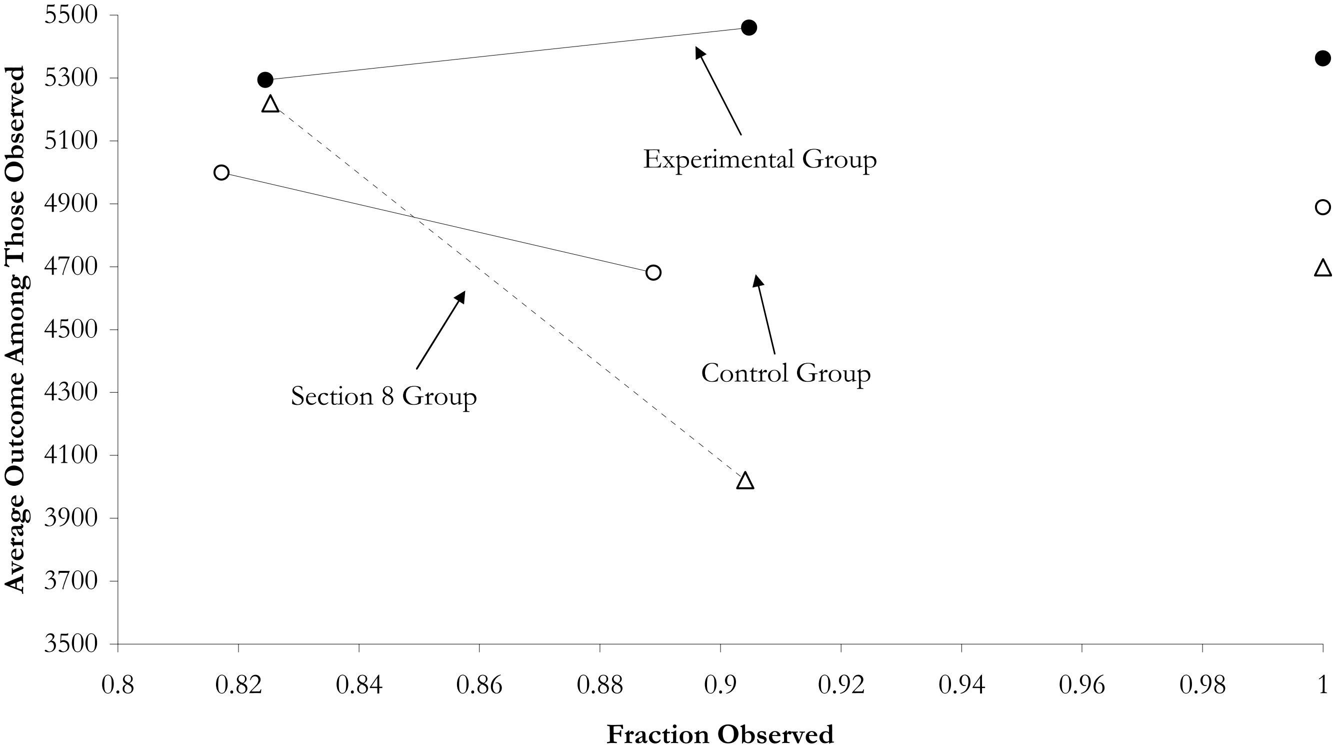


Figure 3. Graphical Interpretation of Sample Selection Correction: TANF Recipiency, Year 5

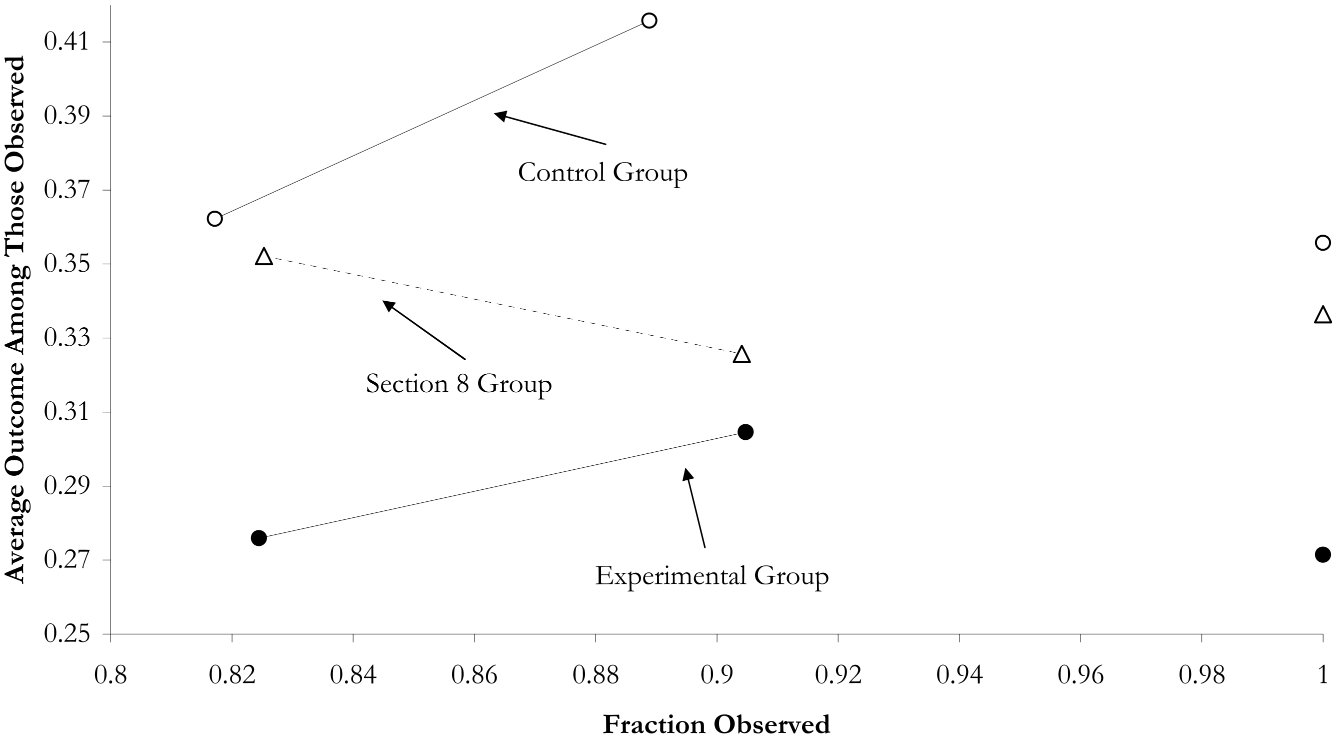


Figure 4. Graphical Interpretation of Sample Selection Correction: TANF Amount, Year 5

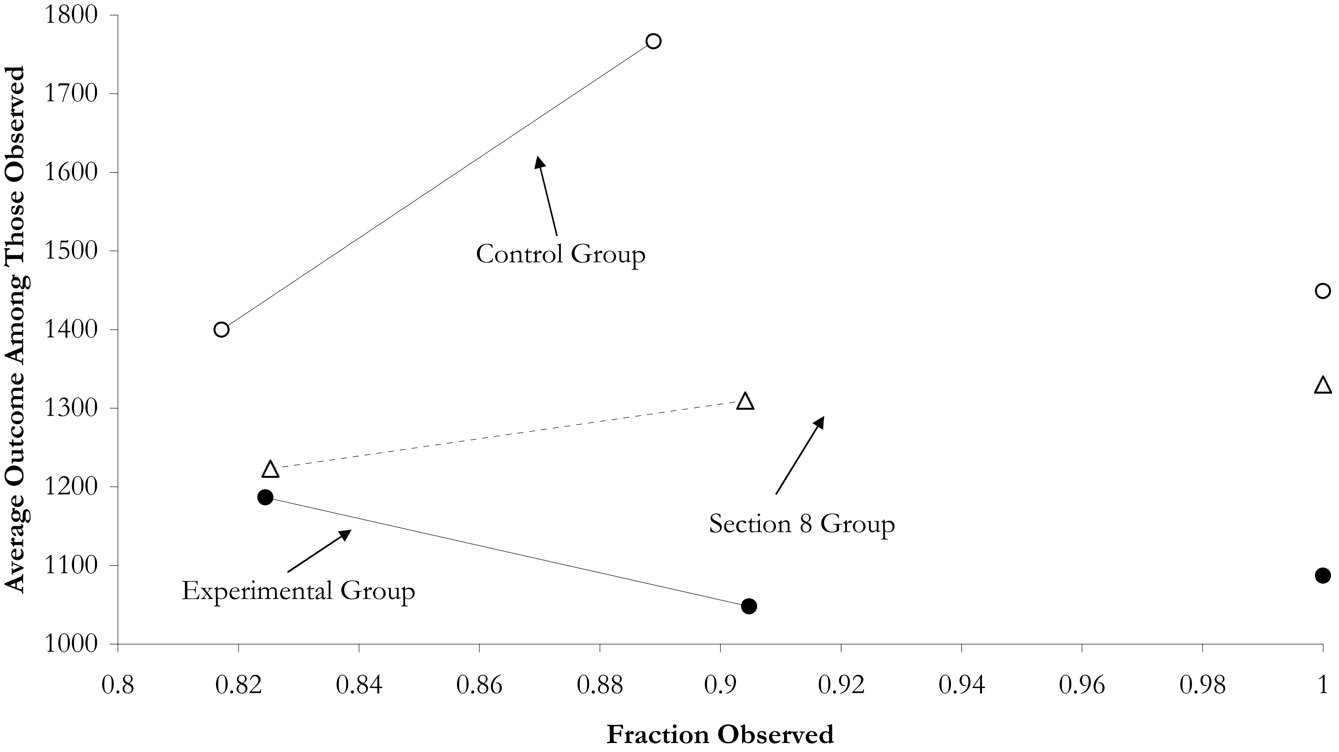


Table 1. Estimated MTO Program Impacts: Fraction of Quarters Employed, Years 1 to 4

	Intention-to-Treat Parameters				Chi-square Statistic on	
Estimator	Experimental	Section 8	Control Mean	Selection Correction	Exclusion of Instruments	Number of Observations
Main Sample, OLS	0.003	-0.018	0.404			1055
(Ignore the problem)	(0.026)	(0.033)	(0.020)			
Lee Bounds	-0.007 0.008	-0.027 -0.014	NA			
(Bound the problem)	(0.026) (0.026)	(0.032) (0.033)				
	[.0137^{e}]	[.0143^{c}]				
Heckman Two-Step	0.007	-0.013	0.329	0.257	12.320	1055, 1248
(Try to fix the problem approximately)	(0.028)	(0.035)	(0.066)	(0.213)	p=0.006	
Heckman MLE	NC	NC	NC	NC	NC	1055, 1248
(Try to fix the problem)						
Full Sample, OLS	0.001	-0.018	0.399			1248
(Ideal solution)	(0.024)	(0.030)	(0.019)			

Note: Standard errors in parentheses. "NC"="no convergence", i.e., failure to converge on the part of of STATA's maximum likelihood routine.

[c] Lee procedure involves trimming from control group.

[e] Lee procedure involves trimming from experimental group.

[s] Lee procedure involves trimming from section 8 group.

Table 2. Estimated MTO Program Impacts: Annualized Earnings, Years 1 to 4

	Intention-to-Treat Parameters						Chi-square Statistic on	
Estimator	**Experimental**		**Section 8**		**Control Mean**	**Selection Correction**	**Exclusion of Instruments**	**Number of Observations**
Main Sample, OLS	442.54		-75.18		4904.87			1055
(Ignore the problem)	(517.76)		(622.94)		(405.62)			
Lee Bounds	201.01	514.39	-431.40	-27.12	NA			
(Bound the problem)	(437.06)	(519.61)	(593.23)	(625.88)				
	[.0137[e]]		[.0143[c]]					
Heckman Two-Step	479.55		-36.91		4333.66	1961.07	12.320	1055, 1248
(Try to fix the problem approximately)	(520.57)		(658.02)		(1230.66)	(3988.10)	p=0.006	
Heckman MLE	NC		NC		NC	NC	NC	1055, 1248
(Try to fix the problem)								
Full Sample, OLS	473.32		-190.94		4888.68			1248
(Ideal solution)	(475.58)		(562.69)		(366.74)			

Note: Standard errors in parentheses. "NC"="no convergence", i.e., failure to converge on the part of of STATA's maximum likelihood routine.

[c] Lee procedure involves trimming from control group.

[e] Lee procedure involves trimming from experimental group.

[s] Lee procedure involves trimming from section 8 group.

Table 3. Estimated MTO Program Impacts:
TANF Recipiency, Year 5

	Intention-to-Treat Parameters				Chi-square Statistic on	
Estimator	Experimental	Section 8	Control Mean	Selection Correction	Exclusion of Instruments	Number of Observations
Main Sample, OLS (Ignore the problem)	-0.093 (0.030)	-0.034 (0.030)	0.378 (0.024)			1055
Lee Bounds (Bound the problem)	-0.104 -0.089 (0.030) (0.030) [.0137[e]]	-0.044 -0.031 (0.039) (0.039) [.0143[s]]				
Heckman Two-Step (Try to fix the problem approximately)	-0.097 (0.031)	-0.038 (0.039)	0.437 (0.074)	-0.201 (0.239)	12.320 p=0.006	1055, 1248
Heckman MLE (Try to fix the problem)	-0.094 (0.030)	-0.035 (0.038)	0.392 (0.034)	-0.046 (0.082)	12.590 p=0.006	1055, 1248
Full Sample, OLS (Ideal solution)	-0.084 (0.027)	-0.019 (0.036)	0.356 (0.022)			1248

Note: Standard errors in parentheses. "NC"="no convergence", i.e., failure to converge on the part of of STATA's maximum likelihood routine.

[c] Lee procedure involves trimming from control group.

[e] Lee procedure involves trimming from experimental group.

[s] Lee procedure involves trimming from section 8 group.

Table 4. Estimated MTO Program Impacts: TANF Amount, Year 5

	Intention-to-Treat Parameters				Chi-square Statistic on	
Estimator	Experimental	Section 8	Control Mean	Selection Correction	Exclusion of Instruments	Number of Observations
Main Sample, OLS	-367.03	-257.33	1508.586			1055
(Ignore the problem)	(145.08)	(179.76)	(118.06)			
Lee Bounds	-477.02 -351.69	-244.88 -345.46				
(Bound the problem)	(140.35) (145.59)	(173.96) (180.65)				
	[.0137[e]]	[.0143[s]]				
Heckman Two-Step	-373.85	-264.38	1613.86	-361.41	12.320	1055, 1248
(Try to fix the problem approximately)	(142.72)	(180.40)	(337.23)	(1092.93)	p=0.006	
Heckman MLE	-368.50	-258.85	1531.34	-78.15	12.350	1055, 1248
(Try to fix the problem)	(140.87)	(178.59)	(159.89)	(396.10)	p=0.006	
Full Sample, OLS	-361.85	-118.63	1448.91			1248
(Ideal solution)	(131.17)	(173.51)	(106.77)			

Note: Standard errors in parentheses. "NC"="no convergence", i.e., failure to converge on the part of of STATA's maximum likelihood routine.

[c] Lee procedure involves trimming from control group.

[e] Lee procedure involves trimming from experimental group.

[s] Lee procedure involves trimming from section 8 group.

Table 5. Tests for Additive Separability Under Normality (Equality of Slopes)

	Control Slope	*Experimental Slope*	*Section 8 Slope*	*Test Statistic* $X^2(2)$	*Probability Value Under Null of Additive Separability*
Fraction of Quarters Employed, Years 1 to 4	0.090	0.026	1.081	1.884	0.61
	(0.4111)	(0.2750)	(0.7221)		
Annualized Earnings, Years 1 to 4	2864	-1314	9681	0.978	0.39
	(7956)	(5450)	(9894)		
TANF Recipiency, Year 5	-0.482	-0.227	0.214	0.882	0.36
	(0.5481)	(0.3157)	(0.5280)		
TANF Amount, Year 5	-3299	1098	-701	1.981	0.63
	(2935)	(1369)	(2435)		

Notes: The slopes are calculated as the difference in the observed means for Z=0 and 1 (those not subject to intense followup and those subject to intense follow up respectively) divided by the respective difference in the inverse mills ratio terms. The t