NBER WORKING PAPER SERIES

BOUNDING THE EFFECTS OF SOCIAL EXPERIMENTS: ACCOUNTING FOR ATTRITION IN ADMINISTRATIVE DATA

Jeffrey Grogger

Working Paper 18838
http://www.nber.org/papers/w18838

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2013

The author thanks Dan Black, Bob Lalonde, an anonymous reviewer, seminar participants at the University of St. Gallen, and participants at the Fifth IZA Conference on Labor Market Policy Evaluation for helpful comments. This paper is forthcoming in Evaluation Review. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Bounding the Effects of Social Experiments: Accounting for Attrition in Administrative Data
Jeffrey Grogger
NBER Working Paper No. 18838
February 2013
JEL No. C33,C9,I38

ABSTRACT

Social experiments frequently exploit data from administrative records. However, most administrative data systems are state-specific, designed to track earnings or benefit payments among residents within a single state. Once an experimental participant moves out of state, his earnings and benefits in his state of origin consist entirely of zeros, giving rise to a form of attrition. In the presence of such attrition, the average treatment effect of the experiment is no longer point-identified, despite random assignment. I propose a method to estimate such attrition and, for binary outcomes such as employment, to construct bounds on the average treatment effect. Results from a welfare-reform experiment considered to have sizeable effects appear quite ambiguous after accounting for attrition. The results have important implications for planning social experiments.

Jeffrey Grogger
Irving B. Harris Professor of Urban Policy
Harris School of Public Policy
University of Chicago
1155 E. 60th Street
Chicago, IL 60637
and NBER
jgrogger@uchicago.edu

1. Introduction

Field experiments play an increasingly important role in economic research. Greenberg and Schroder (2004) catalog over 250 social experiments that have been carried out since the late 1960s. Those studies cover a broad array of topics, including job training, welfare programs, and tax compliance initiatives, among others. Examples from the US include the Gary Negative Income Tax Experiment; the National Job Training and Partnership Act Study; the Unemployment Insurance Bonus experiments; and numerous welfare-reform experiments (Greenberg, Moffitt, and Friedman 1981; Bloom, et al, 1997; Meyer 1995; Grogger and Karoly 2005). A sampling of social experiments outside the US includes the Canadian Self-Sufficiency Project; the Dutch Counseling and Monitoring Program; and job-training programs in Norway (Michalopoulos, Tattrie, et al 2002; Gorter and Kalb 1996; Raaum and Torp 2002).

The chief virtue of social experiments stems from random assignment. By assigning participants at random between control and treatment groups, they provide a means to estimate the effects of experimental programs in a manner that is free of the omitted variable bias that may arise in non-experimental designs. Experimental designs also lend themselves to simple analytical methods that are relatively transparent and easy for non-statisticians to comprehend.

Evaluations of social experiments often make use of longitudinal administrative data. This has been increasingly the case in recent years (Greenberg and Schroder 2004). Administrative data come from sources such as unemployment offices and welfare agencies. Depending on the focus of the experiment, evaluation studies may make use of administrative data on a number of different outcomes from a number of different sources. As compared to survey data, administrative data may be much less expensive to obtain.

Despite their cost advantage, administrative data suffer from an important missing data problem. Since programs such as unemployment insurance and welfare are run at the state- or country-level, the administrative data systems used to operate those programs only track activity within the jurisdiction. They are not designed to capture out-of-jurisdiction migration. Thus when a worker moves to a new jurisdiction, she will appear to have zero earnings in her jurisdiction of origin, even if she is working in her new jurisdiction of residence. This has long been a problem in the US, where cross-state migration is common. With increasing international migration, the same problem may increasingly affect European register data.

Cross-jurisdiction migration means that zeros in the data arise for two reasons: because the participant has no earnings or benefits, or because she has moved to a new jurisdiction. Since out-of-jurisdiction migration is not directly observed, and most evaluations obtain data from a single jurisdiction, the analyst faces a corrupted sample. Rather than observing data on the outcomes of interest, the analyst observes a mixture of such outcomes and out-of-jurisdiction migration, without being able to distinguish between them.

If such migration were random, then the observed data would provide a lower bound on the effect of the program. However, if attrition arises from choices made by experimental subjects, it may be non-random. Under non-random attrition, the observed data identify neither the effect of the program nor a lower bound.

In this paper I consider one aspect of the out-of-jurisdiction migration problem. I analyze permanent out-of-jurisdiction migration, which I refer to as attrition. Although this is not the only form of migration, it is extensive. Moreover, it leaves a distinctive trail in the data. When a study participant moves from one jurisdiction to another, all of her subsequent entries in the administrative data systems of her jurisdiction of origin become zeros. I propose using these terminal runs of zeros to impute attrition. I show how the observed data and the imputed attrition indicator can be used to provide bounds on the average treatment effect without imposing assumptions about the process generating the attrition.

These bounds are informative, but their width depends on the information used to impute attrition. More information reduces imputed attrition and with it the width of the bounds. The magnitude of the reduction can be sizeable. In the welfare experiment that I analyze below, imputations based on multiple outcomes cut the width of the bounds by one-half to one-third as compared to imputations based on a single outcome.

The bounds can also be narrowed by imposing additional structure. I focus on a stationarity assumption that allows me to use non-terminal runs of zeros to estimate the share of terminal runs that represent true attrition. A virtue of the stationary bounds is that the underlying stationarity assumption can be tested.

The bounding strategy fits with the approach of Horowitz and Manski (1995) and particularly Horowitz and Manski (2000), who provide bounds on experimental treatment effects in the presence of attrition. However, there are two key differences between the setting that I consider and that of Horowitz and Manski (2000). First, attrition in their setting means that no outcome data are observed. That is different from my setting, where data are observed but the analyst does not know whether they represent the outcome of interest or the consequence of out-of-jurisdiction migration. Second, I assume the availability of longitudinal data, as is often the case in social experiments, which allow bounds to be constructed on the basis of imputed attrition.

The application reported below makes use of data from the Connecticut Jobs First (JF) welfare-reform evaluation. The treatment condition in JF included both a substantial earnings disregard, which allowed recipients to keep most of their benefit after they started working, and a short time limit on benefit receipt. Previous analyses that did not account for the attrition problem indicated that JF raised welfare receipt substantially before the time limit became binding, and reduced it sharply thereafter (Bloom et al. 2002). The analysis below suggests that there is a great deal of ambiguity in the effects of the program once one accounts for attrition.

In the next section I discuss the JF program, the data, and conventional estimates of the average treatment effect (ATE) of the program on employment and transfer receipt. In Section 3 I show that those estimates do not identify the ATE. I discuss the attrition problem, my attrition indicator, and the bounds based on that indicator. Section 4 discusses the bounds that can be obtained under the assumption of stationarity and the specification tests. Empirical results from our application accompany each section.

2. The Jobs First Program

2.1 The Experiment

JF began in 1996 with the goal of testing the effects of Connecticut's welfare reform program. New applicants to the welfare program were randomized between treatment and control status at the time they applied for aid. Ongoing recipients were randomized when their eligibility was reassessed (which typically took place every six months). The control group received aid under the rules of the Aid to Families with Dependent Children program. As compared to the control group, participants in the treatment group faced a financial work incentive, in that they were able to retain most of their benefits when they found a job. At the same time, the treatment group faced a short, 21-month time limit, whereas the control faced none for four years after random assignment.

Economic theory predicts that the combination of a financial incentive and a time limit should increase employment among welfare recipients. However, it predicts that the two components of the program should have differing effects on welfare receipt. Whereas the time limit should reduce welfare receipt, especially once it starts to bind, the financial incentive may prolong welfare receipt, since it provides the participant with an incentive to combine work and welfare (Grogger and Karoly 2005; Bitler, Gelbach, and Hoynes 2006).

There were 4803 participants included in the evaluation. Of these, 2396 were assigned to the control group and 2407 to the treatment group. The program and its evaluation are discussed in detail by Bloom et al. (2002).

2.2 Data

Outcome data come from two administrative data systems: Connecticut's Unemployment Insurance (UI) system and the state's Eligibility Management System (EMS). The UI system records quarterly earnings for each worker covered by the UI program. Data are reported by employers to the state. Covered employment accounts for roughly 90 percent of all jobs nationwide (US Bureau of Labor Statistics 1989), although it misses roughly 10 to 20 percent of employment among welfare populations (Isaacs and Lyons 2000). Nevertheless, covered employment is a widely used outcome in the evaluation of both welfare- and training-related programs. By this definition, participants are employed if they have positive covered earnings during the calendar quarter and are nonemployed if they have zero covered earnings.

EMS provides data on monthly welfare payments and Food Stamp payments for all recipients in the state. Benefit data are available monthly, but have been aggregated to the quarterly level so as to be comparable to the employment data. A participant is recorded as a welfare recipient if she received a positive welfare benefit in any month during the quarter.

Likewise, she is recorded as a Food Stamp recipient if she received a positive Food Stamp benefit in any month during the quarter.

3. Average Treatment Effects under Attrition

In this section I define the ATE and the conventional ATE as estimated from the observed data. I present estimates of the conventional ATE from JF. I show that, in the presence of attrition, such conventional estimates do not identify the true ATE.

Let T_i denote the participant's treatment status, where $T_i = 1$ denotes assignment to treatment and $T_i = 0$ denotes assignment to control. In the typical social experiment, there may be several outcomes of interest. Attention here is restricted to binary outcomes, which are naturally bounded. These include the outcomes I focus on from JF. Denote the K potential outcomes for the ith participant at time t by

$Y_{it}^k(T_i) \in \{0,1\}$ i=1,…,N; t=0,…,M; and k=1,…,K.

There are N participants in the experiment; the sample period extends for M periods. Period t=0 denotes the time at which random assignment takes place.

Since counterfactual outcomes are unobservable, in the absence of attrition the analyst would observe

$$Y_{it}^k = T_i Y_{it}^k(1) + (1 - T_i) Y_{it}^k(0)$$

Define the average treatment effects of the program at time t as $\Delta_t^k = EY_{it}^k(1) - EY_{it}^k(0)$. This represents the mean effect of the program among the population at risk of treatment. By virtue of random assignment to treatment status, Δ_t^k could be identified from Y_{it}^k, since[1]

[1] I also require the stable unit treatment value assumption, which essentially rules out general equilibrium effects (Angrist, Imbens and Rubin 1996). Implicitly I am also assuming full compliance with the experiment, which is appropriate for this application.

$$\Delta_t^k = EY_{it}^k(1) - EY_{it}^k(0)$$
$$= E(Y_{it}^k \mid T_i = 1) - E(Y_{it}^k \mid T_i = 0).$$

This approach could easily be extended to allow for treatment effects that vary according to observable characteristics, and such heterogeneity would often be of interest in applied settings. I leave it implicit here in order to avoid notational clutter.

The problem for the analyst is that Y_{it}^k itself is observed only prior to the date of attrition, which is not directly observed. Define the latent, unobserved attrition indicator as

$A_{it} = 1$ if i has permanently left the jurisdiction on or before date t

$= 0$ otherwise.

Since I define attrition as permanent migration from the jurisdiction, $A_{it} = 1 \Rightarrow A_{it+j} = 1$, j=1,.., M-t. Once the participant has left the jurisdiction, I no longer observe Y_{it}^k. Rather, I observe zeros instead. Since Y_{it}^k itself takes on zero values, the analyst does not know whether the observed zeros represent valid realizations of Y_{it}^k or attrition.

Given the definition of A_{it}, I can represent the observed data as

$$X_{it}^k = (1 - A_{it})Y_{it}^k \qquad \text{i=1,…,N; t=1,…,M; and k=1,…,K.} \qquad (1)$$

The conventional ATE reported in the literature estimates not Δ_t^k, but rather

$$\Delta_{Xt}^k = E(X_{it}^k \mid T_i = 1) - E(X_{it}^k \mid T_i = 0). \qquad (2)$$

Estimates of Δ_{Xt}^k, obtained as treatment-control contrasts in the sample averages of the outcome variables, are represented by the black lines marked with squares in Figure 1. I follow much of the literature in reporting estimates by quarter after random assignment. The dashed black lines show confidence intervals obtained as the 5th and 95th percentiles of estimates from 500 bootstrap replications.

[Figure 1 about here]

These estimates are similar to those reported by Bloom et al. (2002). They show that JF raised employment in every quarter after random assignment by an amount ranging from roughly 4 to 7 percentage points. All of these estimates are significant in the sense that the low end of the bootstrap confidence interval exceeds zero. The conventional treatment effect estimates for welfare receipt are statistically significantly greater than zero during the first 7 quarters after random assignment. They become significantly negative thereafter. This pattern stems from the fact that time limits began to bind in quarter 8 and has been noted elsewhere in the literature (Bloom et al 2002; Grogger and Karoly 2005; Bitler, Gelbach, and Hoynes 2006). The conventional treatment effects for Food Stamp receipt are significantly positive for the first seven quarters after random assignment and insignificantly different from zero thereafter.

However, estimates of Δ_{Xt}^k do not identify the average treatment effects Δ_t^k. To see this, write the conditional expectation of Y_{it}^k in terms of A_{it} as

$$E(Y_{it}^k \mid T_i) = P(A_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, A_{it}=0) + P(A_{it}=1 \mid T_i)E(Y_{it}^k \mid T_i, A_{it}=1). \quad (3)$$

Writing the conditional expectation of X_{it}^k similarly yields

$$\begin{aligned} E(X_{it}^k \mid T_i) &= P(A_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, A_{it}=0) + P(A_{it}=1 \mid T_i)E(1-A_{it} \mid T_i, A_{it}=1) \\ &= P(A_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, A_{it}=0) \\ &\le E(Y_{it}^k \mid T_i) \end{aligned} \quad (4)$$

where the last line follows since both $P(A_{it}=1 \mid T_i)$ and $E(Y_{it}^k \mid T_i, A_{it}=1)$ are bounded between zero and one. This implies that

$$\Delta_{Xt}^k = P(A_{it}=0 \mid T_i=1)E(Y_{it}^k \mid T_i=1, A_{it}=0) - P(A_{it}=0 \mid T_i=0)E(Y_{it}^k \mid T_i=0, A_{it}=0), \quad (5)$$

which does not identify Δ_t^k. Δ_{Xt}^k would provide a lower bound for Δ_t^k (in an absolute value sense) if attrition were independent of treatment status, since in that case $P(A_{it} \mid T_i) = P(A_{it})$. However, if attrition is affected by treatment, then Δ_{Xt}^k cannot be used to point-identify Δ_t^k. Estimates of Δ_{Xt}^k can be used to provide bounds on Δ_t^k that are valid for general departures from random attrition.[2]

4. Bounds based on imputed attrition

4.1 Hypothetical bounds based on A_{it}

To motivate my approach, it is instructive to consider how one might construct bounds if A_{it} were observable. Define

$Z_{it}^k = (1 - A_{it})Y_{it}^k + A_{it}$ i=1,…,N; t=1,…,M; and k=1,…,K. Then

$$\begin{aligned} E(Z_{it}^k \mid T_i) &= P(A_{it} = 0 \mid T_i)E(Y_{it}^k \mid T_i, A_{it} = 0) + P(A_{it} = 1 \mid T_i)E(A_{it} \mid T_i, A_{it} = 1) \\ &= P(A_{it} = 0 \mid T_i)E(Y_{it}^k \mid T_i, A_{it} = 0) + P(A_{it} = 1 \mid T_i) \\ &\geq E(Y_{it}^k \mid T_i), \end{aligned} \tag{6}$$

where the final inequality follows since

$$E(Z_{it}^k \mid T_i) - E(Y_{it}^k \mid T_i) = P(A_{it} = 1 \mid T_i)[1 - E(Y_{it}^k \mid T_i, A_{it} = 1)]$$

and $P(A_{it} = 1 \mid T_i)$ and $E(Y_{it}^k \mid T_i, A_{it} = 1)$ are bounded between zero and one. Combining (4) and (6) yields

$$E(X_{it}^k \mid T_i) \leq E(Y_{it}^k \mid T_i) \leq E(Z_{it}^k \mid T_i),$$

from which it follows that

$$E(X_{it}^k \mid T_i = 1) - E(Z_{it}^k \mid T_i = 0) \leq \Delta_t^k \leq E(Z_{it}^k \mid T_i = 1) - E(X_{it}^k \mid T_i = 0), \tag{7}$$

---

[2] When Δ_t^k varies as a function of covariates, the bounds are valid for the average treatment effect within the population at risk of treatment for given values of the covariates.

or equivalently,

$$\Delta_{Xt}^{k} - P(A_{it} = 1 | T_i = 0) \le \Delta_t^k \le \Delta_{Xt}^k + P(A_{it} = 1 | T_i = 1). \qquad (8)$$

Equation (8) shows that the width of the bound is $P(A_{it} = 1 | T_i = 1) + P(A_{it} = 1 | T_i = 0)$.

The approach yields bounds because X_{it}^k and Z_{it}^k complement one another. Whereas X_{it}^k effectively replaces post-attrition values of Y_{it}^k with zeros, Z_{it}^k replaces them with ones. Of course, the approach is not feasible empirically because the attrition indicator A_{it} is not observed. In the next section I introduce imputed attrition indicators based on observable data and show that they can be used in a similar way to construct bounds for Δ_t^k.

4.2. Imputing attrition

The imputed attrition indicators are quite intuitive. Since attrition implies that $X_{it}^k = 0$ for all periods following the date of departure, runs of zeros at the end of the sample period are suggestive of attrition. I use these terminal runs of zeros to impute attrition.

Although our primary interest is in terminal runs of zeros, it is useful for discussing the stationary bounds below to adopt notation for runs of arbitrary initial duration d. Such runs can be defined for any of the K observable outcome sequences. Define binary indicators of d-period runs of zeros for the kth outcome beginning at date t as

$$a_{it}^k(d) = I\left(\sum_{j=0}^{d-1} X_{it+j}^k = 0\right)$$

where *I*(*A*)=1 if the event *A* is true and *I*(*A*)=0 otherwise. Thus $a_{it}^k(d) = 1$ if $X_{it+j}^k = 0$ for j=0,…,*d*-1. Data from all K outcome sequences can be pooled by defining

$$a_{it}(d) = I\left(\sum_{k=1}^{K} a_{it}^{k}(d) = K\right).$$

The terminal runs of zeros that I use to impute attrition are indicated by $a_{it}^{k}(M-t+1)$ for single outcomes and $a_{it}(M-t+1)$ for multiple outcomes. For ease of notation, I define $a_{it}^{k} \equiv a_{it}^{k}(M-t+1)$ and $a_{it} \equiv a_{it}(M-t+1)$. I focus attention on the multiple-outcome indicator a_{it} for the rest of the theoretical development, although the same analysis could be carried out in terms of the single-outcome indicators a_{it}^{k}.

4.3. Bounds based on imputed attrition

The observed data X_{it}^{k}, which are defined in terms of the latent attrition indicator A_{it}, can be defined equivalently in terms of the imputed attrition indicator a_{it}. To see this, let

$$x_{it}^{k} = (1-a_{it})Y_{it}^{k} \qquad \text{i=1,…,N; t=1,…,M; and k=1,…,K.} \qquad (9)$$

Equations (1) and (9) imply that $X_{it}^{k} = x_{it}^{k}$ whenever $A_{it} = a_{it}$. $A_{it} \neq a_{it}$ implies that $A_{it} = 0$ and $a_{it} = 1$, since $A_{it} = 1 \Rightarrow a_{it} = 1$. This arises when $X_{it+j}^{k} = 0$, for k = 1,…,K and j=0,…, M-t-1, even though the participant has not yet left the jurisdiction. In this case, $X_{it}^{k} = 0$ and $x_{it}^{k} = 0$ (since $a_{it} = 1$), so the two sequences are equivalent. Thus x_{it}^{k} identifies Δ_{Xt}^{k}.

Proceeding along the lines of section 4.1, define

$$z_{it}^{k} = (1-a_{it})Y_{it}^{k} + a_{it} \qquad \text{i=1,…,N; t=1,…,M; and k=1,…,K.} \qquad (10)$$

Evaluating the expectations of (9) and (10) yields

$$\begin{aligned} E(x_{it}^{k} \mid T_i) &= P(a_{it}=0 \mid T_i)E(Y_{it}^{k} \mid T_i, a_{it}=0) + P(a_{it}=1 \mid T_i)E(1-a_{it} \mid T_i, a_{it}=1) \\ &= P(a_{it}=0 \mid T_i)E(Y_{it}^{k} \mid T_i, a_{it}=0) \end{aligned} \qquad (11)$$

and

$$\begin{aligned} E(z_{it}^k \mid T_i) &= P(a_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, a_{it}=0) + P(a_{it}=1 \mid T_i)E(a_{it} \mid T_i, a_{it}=1) \\ &= P(a_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, a_{it}=0) + P(a_{it}=1 \mid T_i). \end{aligned} \tag{12}$$

Writing the conditional expectation of Y_{it}^k along the lines of (2), but in terms of a_{it} rather than A_{it}, yields

$$E(Y_{it}^k \mid T_i) = P(a_{it}=0 \mid T_i)E(Y_{it}^k \mid T_i, a_{it}=0) + P(a_{it}=1 \mid T_i)E(Y_{it}^k \mid T_i, a_{it}=1). \tag{13}$$

Comparing (11) and (12) to (13) yields

$$E(x_{it}^k \mid T_i) \le E(Y_{it}^k \mid T_i) \le E(z_{it}^k \mid T_i),$$

from which it follows that

$$E(x_{it}^k \mid T_i=1) - E(z_{it}^k \mid T_i=0) \le \Delta_t^k \le E(z_{it}^k \mid T_i=1) - E(x_{it}^k \mid T_i=0), \tag{14}$$

or equivalently,

$$\Delta_{Xt}^k - P(a_{it}=1 \mid T_i=0) \le \Delta_t^k \le \Delta_{Xt}^k + P(a_{it}=1 \mid T_i=1). \tag{15}$$

The width of the bound is $P(a_{it}=1 \mid T_i=1) + P(a_{it}=1 \mid T_i=0)$. This is wider than the hypothetical bounds above, since $A_{it}=1 \Rightarrow a_{it}=1$. As a general rule, one would expect the width of the bound to depend on the amount and nature of the information used to impute attrition. The next section illustrates.

4.3 Terminal Runs of Zeros and Estimated Bounds on Δ_t^k

Table 1 displays the sample distribution of terminal runs of zeros by the quarter in which the run begins. The first three pairs of columns report tabulations based on terminal runs of zeros in each of the separate outcome sequences. These represent tabulations of a_{it}^k .

[Table 1 about here]

Within each pair of columns, the first column reports the entry hazard for a terminal run, that is, the probability that a terminal run begins at time t, given that it had not yet begun at time t-1. The entry hazard for terminal runs based on single outcome sequences can be written as $P[a_{it}^{k}=1\,|\,a_{i,t-1}^{k}(1)=0)]$. The second column reports the probability of imputed attrition, which can be written in terms of the entry hazards as

$$P(a_{it}^{k}=1)=\sum_{j=0}^{t-1}P[a_{it-j}^{k}=1\,|\,a_{i,t-j-1}^{k}(1)=0)]\,. \tag{16}$$

Thus the first entry in the first column shows that 17.2 percent of the sample began a terminal run of zero employment in the first quarter after random assignment. By definition of a terminal run, these runs were fifteen quarters long. In the second quarter, the probability that a 14-quarter run of zero employment began was 0.6 percent, so the cumulative probability of attrition in quarter 2 was 17.8 percent.

Since the probability of imputed attrition is defined as the cumulative probability of terminal runs of zeros, it grows over the sample period. Based on the employment data alone, attrition is imputed to 42.7 percent of the sample by the end of the sample period. Imputed attrition is even higher based on the other outcomes. Imputed attrition amounts to 75 percent of the sample based on the welfare receipt indicator and 57.3 percent based on the Food Stamp receipt indicator.

Terminal runs of zeros based on all three outcome series, corresponding to a_{it}, appear in the final pair of columns. Combining the three data series to impute attrition has a dramatic effect on the distribution of terminal runs of zeros. Based on all three outcomes, only 1 percent of the sample began a terminal run of zeros in the first quarter of the sample period, as compared to 17 percent based on the employment outcome alone. Based on the multiple-outcome attrition indicator, only 20.8 percent of the sample appears to leave by the end of the sample period.

Combining data series to impute attrition can greatly reduce the amount of attrition that appears in the sample. Although this reduction is particularly large due to the negative correlation among our outcomes, multiple-outcome attrition indicators will yield lower levels of imputed attrition than their single-outcome counterparts provided that the underlying outcomes are not perfectly correlated. This is important, because the analysis above suggests that the width of our bounds will be roughly twice the probability of imputed attrition. It has implications for the design of social experiments, as I elaborate in Section 6.

Figure 1 presents estimated bounds based on equation (15), which are constructed from the multiple-outcome attrition indicator a_{it}. These are depicted as solid gray lines with circular plotting symbols near the top and bottom of each graph. Confidence intervals appear as dashed gray lines. These are based on bootstrapped standard errors derived from 200 bootstrap samples, for which individuals were sampled with replacement.

Consider first the employment results in the top figure. According to the conventional treatment effects, JF raised employment in every quarter after random assignment by an amount ranging from roughly 4 to 7 percentage points. However, the bounds show that there is considerable ambiguity associated with these results. Between quarters 5 and 6, the half-width of the bound becomes greater than the estimate of Δ_{Xt}^{k}. By (15), this implies that the bound covers zero, so the sign of Δ_{t}^{k} is ambiguous. Factoring in sampling variation, the lower bounds are statistically significant in only the first four quarters. Nevertheless, ambiguity owing to attrition is a much larger problem than uncertainty owing to sampling variation. Ignoring attrition makes the data appear much more informative than they truly are.

Turning to the estimates for welfare receipt in the middle figure, the conventional treatment effects were significantly greater than zero during the first 7 quarters after random assignment and then became significantly negative, as discussed above. At the same time, attrition gives rise to considerable ambiguity. The lower bound exceeds zero during the first 7 quarters, but the upper bound is never less than zero in the latter part of the sample period. Strictly speaking, there is ambiguity owing to attrition as to whether the effect of JF on welfare receipt became negative once the time limit became binding.

The bottom figure depicts the effects of JF on Food Stamp receipt. The conventional treatment effects are significantly positive for the first seven quarters after random assignment and insignificantly different from zero thereafter. However, the lower bound is never significantly greater than zero.

5. Stationary Bounds

One can tighten the bounds, and reduce the ambiguity in the estimates, by relaxing the implicit assumption that all terminal runs are equally likely to reflect attrition. Whereas the long terminal runs that begin early in the sample period seem consistent with out-of-jurisdiction migration as discussed in Section 1, the short terminal runs beginning late in the sample period may just as well stem from fluctuations in Y_{it}^k. This may be a particular problem in light of the fluctuations in employment and transfer receipt that are typical of welfare recipients (Bane and Ellwood 1994; Isaacs and Lyon 2000).

Table 2 helps to illustrate this problem. Column (1) shows simply that the length of the terminal run is perfectly negatively correlated with the time that it begins. Columns (2) and (3) report the hazard and probability of a terminal run of zeros by quarter after random assignment, taken from columns (7) and (8) of Table 1. The hazard rises toward the end of the sample period, which could be an artifact of the fixed end of the sample period.

[Table 2 about here]

To account for this artifact, I use data from the entire sample period to estimate the share of runs of a given initial duration that eventually become terminal runs. I then use this estimate to adjust the estimated attrition probability. For example, to adjust the hazard in quarter 15, when all new runs are one quarter long and are necessarily terminal, I estimate from the full data set the share of all one-quarter runs that eventually become terminal. The estimate is 42.7 percent. I use this number to adjust the hazard and the cumulative probability of attrition at quarter 15.

This approach comes at the cost of assuming a particular form of stationarity. As I illustrate below, I require the conditional probability of attrition, given an initial duration of length d, to be the same throughout the sample period. Fortunately, a partial test for this assumption can be constructed from the data.

5.1 Adjusting the Attrition Hazards

The question is whether attrition at time t can be predicted from a run of length d that begins at time t. The attrition hazard at t can be written as

$$P[A_{it}=1 \mid a_{i,t-1}(1)=0)] = P[A_{it}=1 \mid a_{it}(d)=1, a_{i,t-1}(1)=0)]P[a_{it}(d)=1 \mid a_{i,t-1}(1)=0)] \qquad (17)$$

for t=1,...,M-d+1. The left side of (17) is the probability that attrition takes place at t, since $a_{i,t-1}(1)=0$ implies that attrition could not have taken place prior to t. The right side expresses this as the product of two terms. The first is the probability that attrition takes place at t, given that a run of length d (at least) begins at t. The second is the probability that a run of length d (at least) begins at t. The first term can be thought of as an adjustment factor that adjusts for the probability that a run of length at least d actually reflects attrition.

The key stationarity assumption is that the first term in constant in t, that is, that a run of at least d periods be equally likely to predict attrition for all values of t. This requires the relationship between true zeros and attrition to be the same for all t.

To estimate the key adjustment factor, one must replace the unobservable A_{it} with the observable a_{it}. In principle, a separate estimate of the adjustment factor could be obtained from the sample counterpart of $P[a_{it} = 1 | a_{it}(d) = 1, a_{i,t-1}(1) = 0]$ for each value of t=1, ..., M-d. In practice, one would expect lower values of t to yield lower estimates, since they make use of longer follow-up periods. Whereas estimates based on runs beginning at t=1 make use of M-d follow-up periods of data, estimates based on runs beginning at t=M-d make use of a single follow-up period. I return to this issue below when I discuss stationarity tests.

Column (4) of Table 2 reports estimates of $P[a_{it} = 1 | a_{it}(d) = 1, a_{i,t-1}(1) = 0]$. For each value of d, they average over all runs beginning between t=1 and t=M-d. Not surprisingly, these adjustment factors rise in d. All 15-period runs must begin in quarter 1, so the adjustment factor equals 1. Runs of at least 14 periods can begin in either quarter 1 or quarter 2. The entry of 1 in the second row of column (4) tells us that all runs of length at least 14 were in fact terminal runs, regardless of when they began.

Shorter runs are much less predictive of attrition. As mentioned above, among all single-period runs in the data, only 42.7 percent were terminal, as shown in the last row of column (4). About three quarters of all four-period runs were terminal.

Column (5) of Table 2 reports estimates of the product on the right side of (17). This is simply the product of the numbers in columns (2) and (4). This adjusted hazard rises much less strongly over time than its unadjusted counterpart. This suggests that much of the increase in the unadjusted hazard was not due to an increase in the attrition hazard over time, but due to the fact that a run is more likely to be terminal, the closer it begins to the end of the sample period.

The adjusted probability of imputed attrition, defined as the cumulative sum of the adjusted attrition hazards, is reported in column (6). Early in the sample period, it differs little from the unadjusted probability displayed in column (3). As time from random assignment increases, the two probabilities become more different. By the end of the sample period, the adjusted probability of imputed attrition is just over three-fourths of the unadjusted probability.

Stationary bounds based on the adjusted probability of imputed attrition are displayed in Figure 2 as solid gray lines. For comparison purposes, the general bounds from Figure 1 appear as solid black lines. I have omitted confidence intervals to avoid clutter. The stationary bounds are somewhat narrower than the general bounds, particularly toward the end of the sample period.

[Figure 2 about here]

5.2 Testing for Stationarity

The stationarity tests take advantage of the multiplicity of estimates for the adjustment factor $P[a_{it} = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0]$. In principle, one could test whether

$$P[a_{it} = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0] = P[a_{i,t-j} = 1 \mid a_{i,t-j}(d) = 1, a_{i,t-j-1}(1) = 0] \tag{18}$$

for j=1,...M-d. In practice, there may be reasons to adjust this procedure. As discussed above, empirical estimates of $P[a_{it} = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0]$ may differ by t not only because of non-stationarity, but also because runs that begin later in the sample period are more likely to be terminal runs simply because they involve fewer follow-up periods. To deal with this problem, I fix the follow-up period at a particular value s, and base our test on

$$P[a_{it}(d+s) = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0] = P[a_{i,t-j}(d+s) = 1 \mid a_{i,t-j}(d) = 1, a_{i,t-j-1}(1) = 0] \qquad (19)$$

for j=1,...,M-d-s. This limits the number of tests that can be conducted, particularly of those involving runs that begin close to the end of the sample period. Panel A of Table 3 presents estimated values of $P[a_{it} = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0]$ from (18) for odd values of d ranging from 1 to 13 by quarter after random assignment. The other three panels of the Table present estimated values of $P[a_{it}(d+s) = 1 \mid a_{it}(d) = 1, a_{i,t-1}(1) = 0]$ from (19) for the same values of d and the arbitrarily chosen values of s=3, 5, and 7.

[Table 3 about here]

Results in panel A show clear pattern: for fixed initial duration d, the conditional probability of terminal runs is generally higher at the end of the sample period than at the beginning. Particularly consistent are the high conditional probabilities at time t=15-d. F-statistics for tests of stationarity of the conditional attrition probabilities are shown at the bottom of each column. These have been calculated so as to account for the presence of multiple observations from each participant. Also shown are p-values for the F-tests that have not been corrected for multiple comparisons. For an overall test of size 0.05, a Bonferroni correction would lead one to reject the null if any of the unadjusted p-values was greater than 0.05/7=0.0071. The null of stationarity is clearly rejected by the estimates in panel A.

As discussed above, the rejection could arise either from true non-stationarity in the conditional attrition probabilities or from variation in the length of the follow-up period. The pattern in Panel A is consistent with the latter interpretation, since one would expect estimated conditional probabilities to be higher, the shorter the follow-up period used to estimate them. Panels B through D present estimates based on fixed follow-up periods for all values of t and d.

Not surprisingly, the conditional attrition probabilities estimated from fixed follow-up periods are higher than their counterparts in panel A, which utilize all available follow-up data. Shorter follow-ups increase the appearance of attrition. This is evident in moving from Panel B, with the shortest follow-up period and the highest conditional probabilities, to Panel D, with the longest fixed follow-up period and the lowest conditional probabilities.

Fixing the follow-up period also reduces the extent to which the estimated conditional probabilities rise with t, holding d (and s) constant. Indeed, for short initial runs of 1 to 3 periods, the estimated conditional attrition probabilities are lower toward the end of the sample period than at the beginning.

The significance levels of the F-tests for stationarity fall as well. Nevertheless, the tests still reject the null at the 5 percent level in the case of panels B and C, applying Bonferroni corrections. In both cases, the rejections come in the case of the shortest initial runs, where d=1. Only in the case of the longest follow-up period (panel D) does the stationarity test fail to reject. On the one hand, one might wish to weight the evidence from the longest follow-up the most, since longer fixed follow-ups correspond more closely to the notion of attrition as described in Section 1. On the other hand, the longest follow-ups do not allow for stationarity tests toward the end of the sample period, where the gain in terms of tightening the bounds is greatest.

More generally, how one weighs mixed evidence from the stationarity tests should be based on some notion of the loss function attached to tighter bounds on Δ_t^k. If the tighter stationary bounds resolved fundamental ambiguities stemming from the width of the general bounds, stringent rejection criteria for the stationarity tests might be viewed as appropriately cautious. With less at stake, the distinction might be less important.

6. Discussion

Attrition due to out-of-jurisdiction migration poses an important problem for evaluating social experiments. Many such evaluations make use of administrative data. Administrative sources provide longitudinal data that are free of survey response errors and are generally cheaper to obtain than survey data. However, most administrative files cover residents of a single jurisdiction, and most evaluations avail themselves only of data from the jurisdiction in which the experiment is conducted. When a participant leaves the jurisdiction, the entries in her administrative records become made up of zeros. Since zeros are valid values in many cases, and migration is not directly observed, this means the analyst must deal with corrupted data. Conventional estimates based on corrupted data do not identify the average treatment effect.

I propose using terminal runs of zeros to impute attrition and using the imputed attrition indicators to estimate bounds for average treatment effects. The amount of information used to impute attrition plays a key role in determining the width of the bounds. In the empirical example, bounds based on individual outcomes, such as employment or welfare receipt, are too wide to be very meaningful. Bounds based on all available outcomes lead to dramatic reductions in imputed attrition and hence in the width of the bounds. Negative correlation among the outcomes is particularly helpful, but adding any imperfectly correlated outcomes should help produce tighter bounds.

This finding has important implications for the design of experimental evaluations based on administrative data. It indicates that one should collect as much data as possible to reduce the width of the bounds. This is particularly important for experiments that focus primarily on a single outcome, such as employment. Even if other outcomes are of no direct interest given the objectives of the experiments, they may be invaluable for reducing the ambiguity that arises due to attrition. Data from unconventional sources such as school or Department of Motor Vehicle records may provide valuable additional information, as would data from nearby jurisdictions. Data from other time periods or locations within the state might also be used.

In terms of the experiment I analyze, attrition leads to considerable ambiguity regarding its effects. Conventional analyses that did not account for attrition showed the Jobs First welfare reform program to have strong and significant positive effects on employment. However, by the sixth quarter after random assignment, the half-width of the estimated bounds exceeded the magnitude of the conventional ATE.

To tighten the bounds I propose an approach which is valid under a particular stationarity assumption. The stationary bounds were somewhat smaller than the general bounds in the case of the JF program, especially toward the end of the sample period. I also provide tests for the stationarity assumption.

Considerable resources are expended on social experiments, often with the belief that such experiments provide the strongest possible evidence regarding the effects of social programs. Due to attrition, however, results from experimental evaluations may involve much more ambiguity than has previously been recognized. More extensive data collection can help to reduce that ambiguity.

References

The author thanks Dan Black, Bob Lalonde, seminar participants at the University of St. Gallen, participants at the Fifth IZA Conference on Labor Market Policy Evaluation, a referee and the editor for helpful comments. Any errors are his alone.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996) Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 91:444-455.

Bane MJ, Ellwood D (1994) Welfare realities: from rhetoric to reform. Cambridge, Harvard University Press

Bitler M, Gelbach J, Hoynes H (2006) What mean impacts miss: distributional effects of welfare reform experiments. Am Econ Rev 96:988–1012

Bloom D, Scrivener S, Michalopoulos C, Morris P, Hendra R, Adams-Ciardullo D, WalterJ, Vargas W (2002) Jobs first: final report on Connecticut's welfare reform initiative. New York, MDRC

Bloom H, Orr L, Bell S, Cave G, Lin W, Bos J (1997) The benefits and costs of JTPS Title II-A programs: key findings from the national Job Training Partnership Act study. J Hum Resources 32:549-576

Gorter C, Kalb G (1996) Estimating the effect of counseling and monitoring the unemployed using a job search model. J. Hum Resources 31:590-610

Greenberg D, Moffitt R, and Friedmann J (1981) Underreporting and experimental effects on work effort: evidence from the Gary Income Maintenance Experiment." Rev Econ Stat 63:581-589

Greenberg D, Schroeder M (2004) Digest of social experiments. Washington DC, Urban Institute Press

Grogger J, Karoly L (2005) Welfare reform: effects of a decade of change. Cambridge, MA, Harvard University Press

Isaacs J, Lyon M (2000) A cross-state examination of families leaving welfare: findings from the ASPE funded leavers studies. Washington DC, Division of Data and Technical Analysis, Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services

Horowitz J, Manski C (1995) Identification and robustness with contaminated and corrupted data. Econometrica 62:281-302

Horowitz J, Manski C (2000) Nonparametric analysis of randomized experiments with missing covariate and outcome data. J Am Stat Assoc 95:77-84

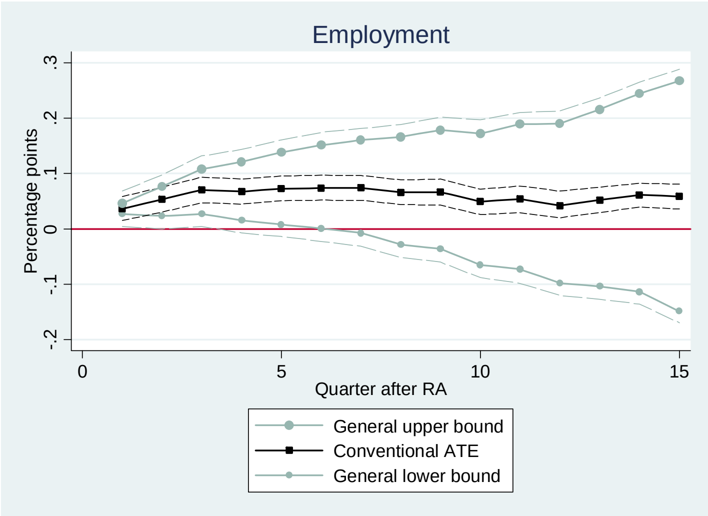
Meyer B (1995) Lessons from the US unemployment insurance experiments. J Econ Lit 33:91-131

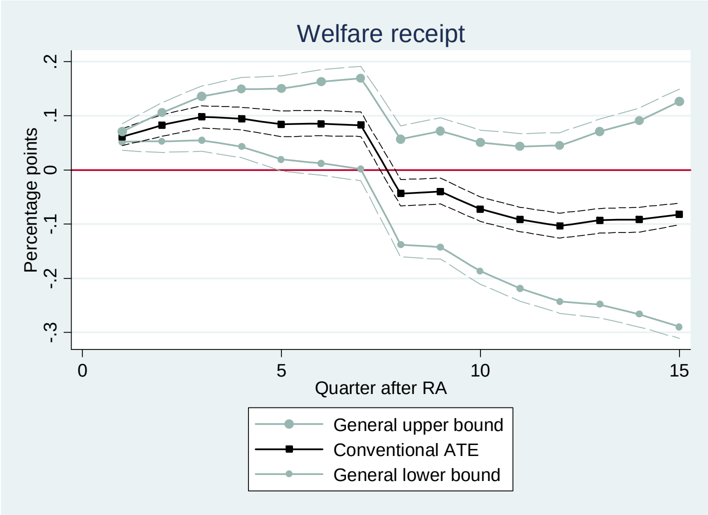
Michalopoulos C, Tattrie D, et al (2002) Making work pay: final report on the Self-Sufficiency Project for long-term welfare recipients. Ottawa: Social Research and Demonstration Corporation

Raaum O, Torp H (2002) Labor market training in Norway—effect on earnings. Labour Econ 9:207-247.

US Bureau of Labor Statistics (1989) Handbook of labor statistics. Washington DC, Government Printing Office

Figure 1: Estimated Average Treatment Effects and General Bounds





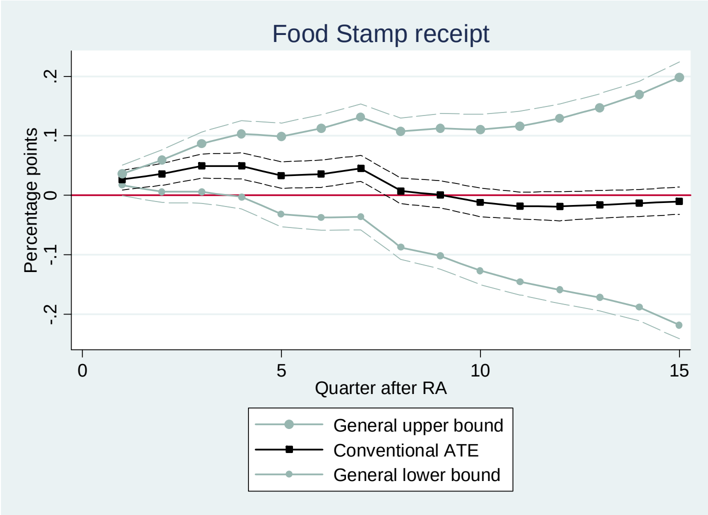


Figure 2: Comparing General and Stationary Bounds

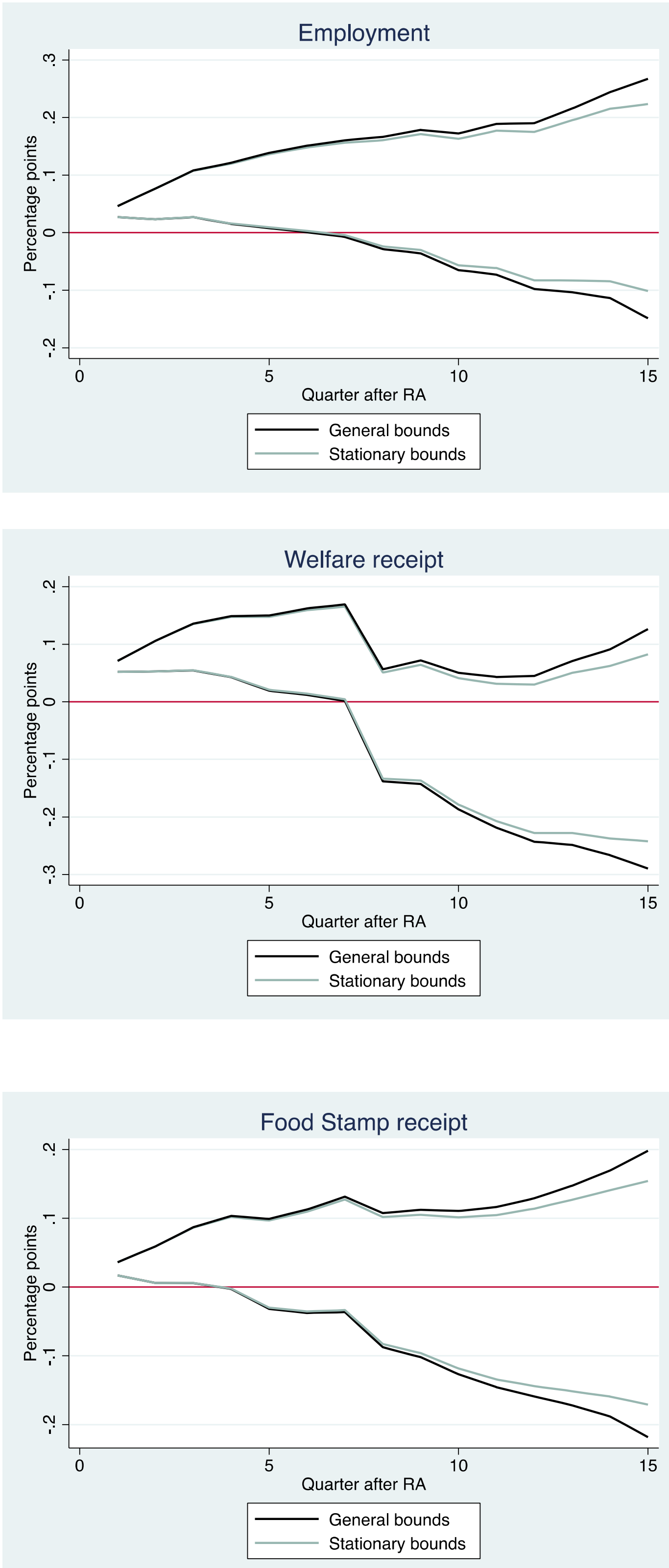


Table 1: Imputed Attrition, Based on Different Outcome Variables, by Quarter After Random Assignment

	Employment		Welfare receipt		Food Stamp receipt		All three outcomes	
Quarter after RA	Hazard for terminal run $P(a_{it}^k = 1 \mid a_{it-1}^k(1) = 0)$ (1)	Cumulative probability for terminal run of zeros $P(a_{it}^k = 1)$ (2)	Hazard for terminal run $P(a_{it}^k = 1 \mid a_{it-1}^k(1) = 0)$ (3)	Cumulative probability for terminal run of zeros $P(a_{it}^k = 1)$ (4)	Hazard for terminal run $P(a_{it}^k = 1 \mid a_{it-1}^k(1) = 0)$ (5)	Cumulative probability for terminal run of zeros $P(a_{it}^k = 1)$ (6)	Hazard for terminal run $P(a_{it}^k = 1 \mid a_{it-1}^k(1) = 0)$ (7)	Cumulative probability for terminal run of zeros $P(a_{it} = 1)$ (8)
1	0.172	0.172	0.086	0.086	0.077	0.077	0.010	0.010
2	0.006	0.178	0.058	0.143	0.040	0.116	0.017	0.027
3	0.009	0.187	0.054	0.197	0.037	0.153	0.014	0.041
4	0.010	0.197	0.046	0.243	0.038	0.191	0.013	0.053
5	0.013	0.210	0.044	0.288	0.034	0.226	0.012	0.065
6	0.012	0.222	0.033	0.321	0.026	0.252	0.010	0.075
7	0.010	0.232	0.039	0.360	0.031	0.283	0.009	0.084
8	0.014	0.245	0.087	0.447	0.047	0.330	0.014	0.097
9	0.012	0.257	0.041	0.488	0.029	0.359	0.010	0.107
10	0.015	0.272	0.052	0.540	0.033	0.391	0.012	0.119
11	0.017	0.289	0.043	0.582	0.034	0.425	0.012	0.131
12	0.022	0.311	0.053	0.636	0.039	0.464	0.013	0.144
13	0.029	0.340	0.039	0.675	0.031	0.495	0.016	0.160
14	0.036	0.376	0.037	0.712	0.036	0.532	0.019	0.179
15	0.051	0.427	0.038	0.750	0.041	0.573	0.029	0.208

Table 2: Adjustments for Stationarity, Imputing Attrition Based on All Three Outcomes

Quarter after RA	Length of terminal spells (d) (1)	Hazard for terminal run $P(a_{it}=1 \mid a_{it-1}(1)=0)$ (2)	Cumulative probability for terminal run of zeros $P(a_{it}=1)$ (3)	Adjustment for stationarity $P(a_{i,t}=1 \mid a_{i,t}(d)=1, a_{i,t-1}(1)=0)$ (4)	Adjusted hazard for terminal run (4)x(2) (5)	Adjusted cumulative probability for terminal run of zeros (Sum of (5)) (6)
1	15	0.010	0.010	1.000	0.010	0.010
2	14	0.017	0.027	1.000	0.017	0.027
3	13	0.014	0.041	0.977	0.014	0.040
4	12	0.013	0.053	0.947	0.012	0.052
5	11	0.012	0.065	0.927	0.011	0.064
6	10	0.010	0.075	0.929	0.009	0.073
7	9	0.009	0.084	0.903	0.008	0.080
8	8	0.014	0.097	0.874	0.012	0.092
9	7	0.010	0.107	0.846	0.008	0.101
10	6	0.012	0.119	0.814	0.009	0.110
11	5	0.012	0.131	0.778	0.010	0.120
12	4	0.013	0.144	0.734	0.010	0.129
13	3	0.016	0.160	0.653	0.010	0.139
14	2	0.019	0.179	0.558	0.011	0.150
15	1	0.029	0.208	0.427	0.013	0.163

Table 3: Estimates of Adjustment Factor $P[a_{it}(d+s)=1 \mid a_{it}(d)=1, a_{i,t-1}(1)=0]$, by Period (t), Minimum Initial Length of Run (d), and Length of Follow-Up (s)

Quarter after RA (t)	A. No adjustment for length of follow-up period							B. Follow-up (s) = 3 quarters					
	Minimum initial length of run (d)							Minimum initial length of run (d)					
	1	3	5	7	9	11	13	1	3	5	7	9	11
1	0.315	0.495	0.667	0.754	0.790	0.821	0.979	0.541	0.72	0.87	0.934	0.864	0.821
2	0.439	0.651	0.781	0.804	0.891	0.921	0.976	0.578	0.817	0.914	0.873	0.946	0.921
3	0.419	0.578	0.684	0.788	0.893	0.957		0.675	0.784	0.806	0.859	0.907	
4	0.414	0.682	0.779	0.845	0.984	1.000		0.589	0.852	0.844	0.845	0.983	
5	0.424	0.648	0.843	0.952	0.983			0.547	0.736	0.857	0.952		
6	0.356	0.603	0.691	0.758	0.887			0.553	0.821	0.853	0.758		
7	0.333	0.667	0.792	0.955				0.444	0.746	0.811			
8	0.445	0.663	0.844	0.985				0.555	0.704	0.844			
9	0.392	0.681	0.839					0.525	0.725				
10	0.401	0.714	0.917					0.453	0.714				
11	0.386	0.694						0.458					
12	0.432	0.840						0.431					
13	0.532												
14	0.681												
15													
Mean	0.427	0.653	0.778	0.846	0.903	0.927	0.977	0.53	0.765	0.852	0.869	0.927	0.883
Average N	144	88	73	69	67	69	66	145	90	77	74	72	73
F	4.93	3.06	3.18	8.32	5.1	7.52	0.01	3.22	1.55	0.98	2.95	3.11	2.87
p-value	0.0000	0.0005	0.0009	0.0001	0.0002	0.0001	0.9255	0.0002	0.1274	0.4437	0.0125	0.027	0.0925

(continued)

Table 3
(continued)

Quarter after RA (t)	C. Follow-up (s) = 5 quarters					D. Follow-up (s) = 7 quarters			
	Minimum initial length of run (d)					Minimum initial length of run (d)			
	1	3	5	7	9	1	3	5	7
1	0.459	0.645	0.826	0.836	0.78	0.411	0.613	0.739	0.754
2	0.551	0.762	0.848	0.853	0.891	0.513	0.706	0.829	0.804
3	0.569	0.681	0.745	0.8		0.494	0.629	0.694	
4	0.517	0.739	0.779	0.845		0.448	0.682	0.779	
5	0.482	0.659	0.843			0.432	0.648		
6	0.485	0.744	0.691			0.439	0.603		
7	0.373	0.683				0.341			
8	0.473	0.663				0.445			
9	0.417								
10	0.401								
11									
12									
13									
14									
15									
Mean	0.478	0.699	0.791	0.834	0.848	0.445	0.65	0.762	0.785
Average N	144	94	81	80	75	148	99	87	82
F	2.44	0.98	1.71	0.32	3.1	1.68	0.75	1.84	0.53
(p-value)	0.0093	0.4471	0.13	0.8123	0.08	0.1107	0.589	0.1393	0.4657

Notes: F-statistics account for presence of multiple observations per participant