Institute for Research on Poverty
Discussion Paper no. 1156-98

An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics

John Fitzgerald
Bowdoin College
E-mail: jfitz@bowdoin.edu

Peter Gottschalk
Boston College
E-mail: peter.gottschalk@bc.edu

Robert Moffitt
Johns Hopkins University
E-mail: moffitt@jhu.edu

March 1998

This research was supported by the National Science Foundation through a grant to the PSID Board of Overseers. We wish to thank Joseph Altonji, Greg Duncan, Guido Imbens, Charles Manski, Gary Solon, Jeffrey Wooldridge, and three anonymous referees for comments on various drafts, as well as seminar participants at Berkeley, Michigan State, New York University, Princeton, Stanford, and the University of Wisconsin. Excellent research assistance was provided by Robert Reville, Lisa Tichy, and Thomas Vanderveen.

Abstract

By 1989, the Michigan Panel Study on Income Dynamics (PSID) had experienced approximately 50 percent sample loss from its initial 1968 membership due to cumulative attrition. We study the effect of this attrition on the unconditional distributions of several socioeconomic variables and on the estimates of several sets of regression coefficients. We provide a statistical framework for conducting tests for attrition bias that draws a sharp distinction between selection on unobservables and on observables and that shows that weighted least squares can generate consistent parameter estimates when selection is based on observables, even when they are endogenous. Our empirical analysis shows that attrition is highly selective and is concentrated among individuals of lower socioeconomic status. We also show that attrition is concentrated among those with more unstable earnings, marriage, and migration histories. Nevertheless, we find that these variables explain very little of the attrition in the sample and that the selection that occurs is moderated by regression-to-the-mean effects from selection on transitory components that fade over time. Consequently, despite the large amount of attrition, we find no strong evidence that attrition has seriously distorted the representativeness of the PSID through 1989, and considerable evidence that its cross-sectional representativeness has remained roughly intact.

An Analysis of Sample Attrition in Panel Data:
The Michigan Panel Study of Income Dynamics

The increased availability of panel data from household surveys has been one of the most important developments in applied social science research in the last 30 years. Panel data have permitted social scientists to examine a wide range of issues that could not be addressed with cross-sectional data or even repeated cross sections. Nevertheless, the most potentially damaging and frequently mentioned threat to the value of panel data is the presence of biasing attrition—that is, attrition that is selectively related to outcome variables of interest.

In this paper we present the results of a study of attrition and its potential bias in one of the most well-known panel data sets, the Michigan Panel Study of Income Dynamics (PSID). The PSID has suffered considerable attrition since it began in 1968—almost 50 percent of initial sample members had attrited by 1989. We study the effect of attrition in the PSID on the means and variances of several important socioeconomic variables—such as individual earnings, educational level, marital status, and welfare participation—and on the coefficients of variables in regressions for these variables. We also examine whether the likelihood of attrition is related to past instability of such behaviors—earnings instability, propensities to migrate or to change marital status, and so on. A companion paper studies the effect of attrition on estimates of intergenerational relationships (Fitzgerald et al., 1997b).

An understanding of the statistical issues is important to understanding our approach. We provide a statistical framework for the analysis of attrition bias which shows that the common distinction between selection on unobservables and observables is critical to the development of tests for attrition bias and adjustments to eliminate it. However, we show that selection on observables is not the same as exogenous selection, because selection can be based on endogenous observables such as lagged dependent variables which are observed prior to the point of attrition. We note that the attrition bias generated by this type of selection can be eliminated with weighted least squares (WLS), using weights

obtained from estimated equations for the probability of attrition, and hence without the highly parametric procedures found in much of the literature. Many of our tests for attrition bias are consequently based on whether lagged endogenous variables affect attrition rates. However, we also conduct an implicit test for selection on unobservables by comparing PSID distributions with those from an outside data source, the Current Population Survey (CPS).

We find that while the PSID has been highly selective on many important variables of interest, including those ordinarily regarded as outcome variables, attrition bias nevertheless remains quite small in magnitude. The major reasons for this lack of effect are that (1) the magnitudes of the attrition effect, once properly understood, are quite small (most attrition is random) and (2) much attrition is based on transitory components that fade away from regression-to-the-mean effects both within and across generations. We also find that attrition-adjusted weights play a small role in reducing attrition bias. We conclude therefore that the PSID has stayed roughly representative through 1989.[1]

I. THE PSID: GENERAL ATTRITION PATTERNS

The PSID began in 1968 with a sample of approximately 4,800 families drawn from the U.S. noninstitutional population (for a general description of the PSID, see Hill, 1992). Since 1968, families have been interviewed annually and a wide variety of socioeconomic information has been collected. Adults and children in the original PSID households or who are descendants of members of those households are followed if they form or join new households, thereby providing the survey the possibility of staying representative of the nonimmigrant U.S. population. A consequence of the self-replenishing nature of the panel is that the sample has grown over time. There were approximately

[1]A similar conclusion was reached by Becketti et al. (1988) for the PSID using data through 1981. See also Duncan and Hill (1989) for an analysis of representativeness in 1980.

18,000 individuals in the 1968 families; by 1989, information on about 26,800 individuals had been collected.[2]

About 60 percent of the 1968 families were drawn from a representative sampling frame of the U.S. called the "SRC" sample, and 40 percent were drawn from a set of individuals in low-income families (mostly in Standard Metropolitan Statistical Areas) known as the "SEO" sample. At the time the survey began, the PSID staff produced weights that were intended to allow users to combine the two samples and to calculate statistics representative of the general population. Those sample weights have been updated periodically to take into account differential mortality as well as differential attrition (see Institute for Social Research, 1992: 82–98, for a recent discussion of nonresponse and other weighting adjustments). We shall discuss the effect of this weight adjustment in our paper.

Table 1 shows response and nonresponse rates of the original 1968 sample members.[3] The first three columns in the table show the number of individuals remaining in the sample by year—the number in a family unit, the portion in institutions—whom we treat as respondents, to be consistent with practice by PSID staff—and their sum, equal to 18,191 individuals in 1968. As the table indicates in the fourth column, about 88 percent of these individuals remained after the second year, implying an attrition rate of 12 percent. The actual number attriting is shown in the fifth column, with conditional attrition rates shown in parentheses below each count. A smaller proportion left the PSID in each year after the first—generally about 2.5 or 3.0 percent annually. By 1989, only 49 percent of the original number were still being interviewed, corresponding to a cumulative attrition rate of 51 percent.

---

[2]Institute for Social Research (1992: Table 14). The PSID also interviews individuals who are not related to a 1968 family but who move into interviewed households, most commonly by marrying a PSID member. Those individuals are termed "nonsample" observations and are assigned a zero weight. Another 11,600 of these individuals had been interviewed by 1989, in addition to the 26,800 mentioned in the text. Generally, such individuals are no longer interviewed if they leave a PSID household. However, all children of a "sample" parent and "nonsample" parent are kept in the survey, which causes the PSID sample size to grow over time; see below.

[3]These attrition rates condition on being interviewed in 1968, the initial year. However, only 76 percent of the families selected to be interviewed were interviewed (Hill, 1992: 25). We return to this issue below in our comparisons with the CPS.

TABLE 1
Response and Nonresponse Rates in the PSID

	Remaining in Sample				Attritors[a]				
Year	In a Family Unit	In an Institution	Total	As a Percentage of 1968 Total	Total	Family Unit Nonresponse	Died	Moved	In from Nonresponse
1968	17807	384	18191	100.0	—	—	—	—	—
1969	15561	367	16028	88.1	2163 (.119)	1797 (.099)	84 (.005)	282 (.016)	—
1970	15126	333	15459	85.0	600 (.037)	351 (.022)	74 (.005)	175 (.011)	31
1971	14767	322	15089	82.9	404 (.026)	208 (.013)	95 (.006)	101 (.007)	34
1972	14400	293	14693	80.8	429 (.028)	190 (.013)	115 (.008)	124 (.008)	33
1973	13969	307	14276	78.5	449 (.031)	247 (.017)	100 (.007)	102 (.007)	32
1974	13581	307	13888	76.3	410 (.029)	229 (.016)	89 (.006)	92 (.006)	22
1975	13226	302	13528	74.4	386 (.028)	200 (.014)	97 (.007)	89 (.006)	26
1976	12785	291	13076	71.9	487 (.036)	310 (.023)	86 (.006)	91 (.007)	35
1977	12377	310	12687	69.7	411 (.031)	234 (.018)	88 (.007)	89 (.007)	22
1978	12078	320	12398	68.2	330 (.026)	210 (.017)	63 (.005)	57 (.004)	41
1979	11718	316	12034	66.2	387 (.031)	224 (.018)	73 (.006)	90 (.007)	23

(table continues)

TABLE 1, continued

Year	Remaining in Sample: In a Family Unit	Remaining in Sample: In an Institution	Remaining in Sample: Total	Remaining in Sample: As a Percentage of 1968 Total	Attritors[a]: Total	Attritors[a]: Family Unit Nonresponse	Attritors[a]: Died	Attritors[a]: Moved	In from Nonresponse
1980	11357	305	11662	64.1	405 (.034)	233 (.019)	90 (.007)	82 (.007)	33
1981	11022	340	11362	62.5	337 (.029)	208 (.018)	77 (.007)	52 (.004)	37
1982	10780	326	11106	61.1	285 (.025)	135 (.012)	88 (.008)	62 (.005)	29
1983	10487	322	10809	59.4	336 (.030)	194 (.017)	83 (.007)	59 (.005)	39
1984	10178	319	10497	57.7	348 (.032)	225 (.021)	93 (.009)	30 (.003)	36
1985	9891	275	10166	55.9	371 (.035)	229 (.022)	96 (.009)	46 (.004)	40
1986	9517	292	9809	53.9	390 (.038)	275 (.027)	84 (.008)	31 (.003)	33
1987	9230	257	9487	52.2	357 (.036)	215 (.022)	94 (.010)	48 (.005)	35
1988	9002	206	9208	50.6	310 (.033)	178 (.019)	95 (.010)	37 (.004)	31
1989	8743	170	8913	49.0	323 (.035)	212 (.023)	79 (.009)	32 (.003)	28

Notes: Excludes new births and nonsample entrants.

[a]Figures in parentheses show attrition rates as a percentage of the total sample remaining in the prior year (column 4).

Table 1 also shows the distribution of the attritors by reason—either because the entire family did not respond ("Family Unit Nonresponse"), because of death, or because of a residential move which could not be successfully followed.[4] The distribution of attrition by reason has not changed greatly over time, although there is a slight increase in the percentage attriting because of death and a slight reduction in the percentage attriting because of mobility. Both of these trends are no doubt a result of the increasing age of the 1968 sample. The final column in the table shows the number of individuals who came back into the survey from nonresponse ("In from Nonresponse") each year. These figures are quite small because, prior to the early 1990s, the PSID did not attempt to locate and reinterview attritors.

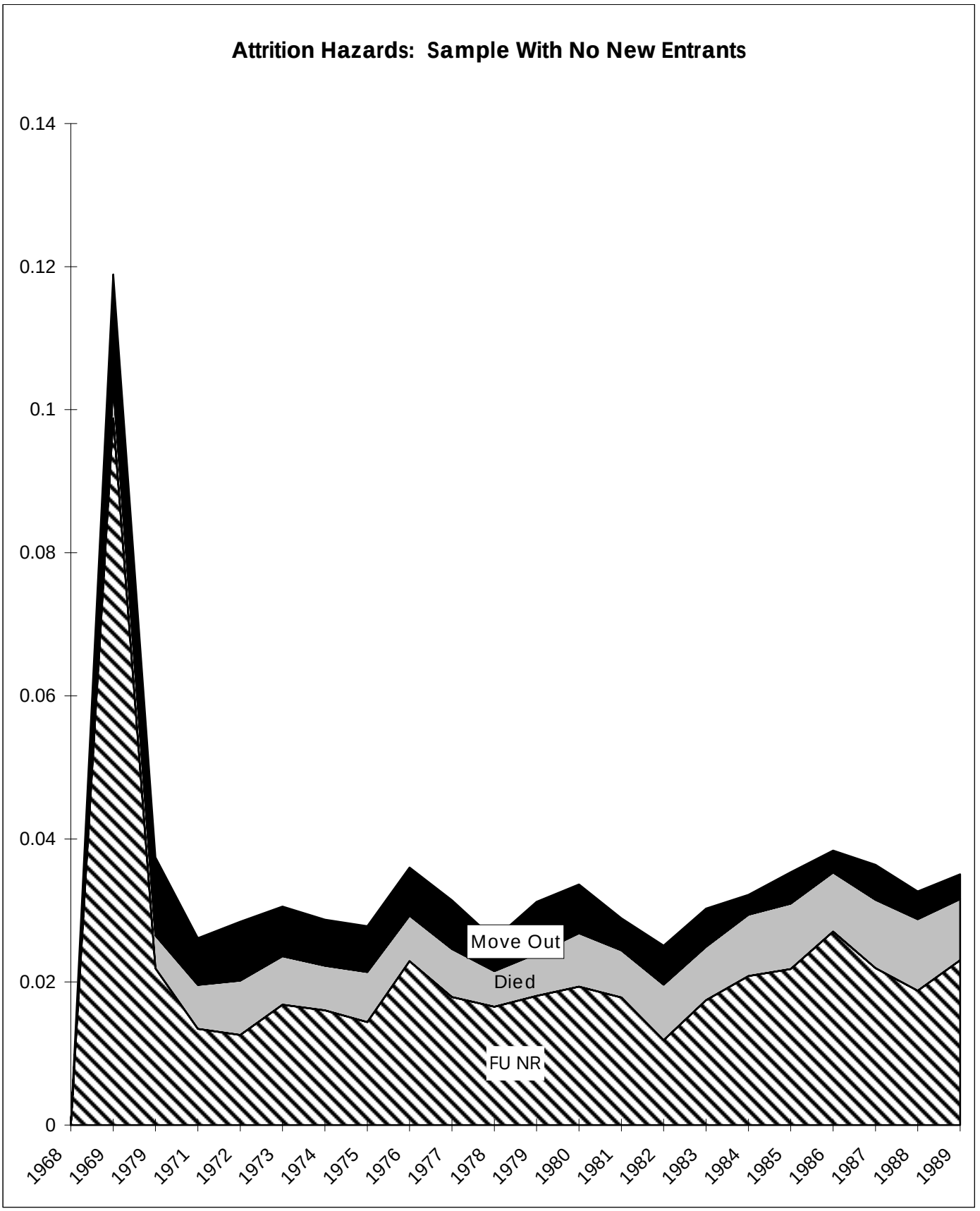
Figure 1 illustrates the overall attrition hazards graphically and clearly shows the spike in the hazard in the first year. It is also more noticeable in the figure that there has been a slight upward trend in attrition rates over time.

In a background report (Fitzgerald et al., 1997a), we show cumulative rates of response among 1968 sample members by race, sex, and age. Cumulative nonresponse rates have been highest for races other than black and white, and next highest for blacks. Nonresponse rates are higher among men than among women. Not surprisingly, nonresponse rates are highest among the older 1968 sample members and among respondents initially between ages 16 and 24. Among the oldest 1968 sample members, those 65 and over, only 7 percent were interviewed in 1989. Nonresponse rates are also higher in the SEO subsample than in the SRC subsample, although not by a large amount.

That mortality should have a marked effect on the measured response rate is not surprising, but it does imply that the 51 percent attrition rate in Table 1 overstates sample loss among the living population. When individuals who died while in the PSID are excluded, overall nonresponse rates fall from 51 percent to 45 percent for the entire sample and from 68 percent to 47 percent among those aged

[4]Some of the "Family Unit Nonresponse" observations may have attrited because of migration or mortality unknown to the PSID.

Figure 1


Attrition Hazards: Sample With No New Entrants
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Move Out
Died
FU NR
1968
1969
1979
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989

55–64. When an additional adjustment is made for mortality among attritors after the point of attrition (using national mortality rates by age, race, and sex), the attrition rate for the older population falls another 12 percentage points to 35 percent, and the overall attrition rate falls to 44 percent (i.e., the estimated percentages of still-alive individuals who have left the PSID).[5]

II. STATISTICAL APPROACH

Although a sample loss as high as 44 percent must necessarily reduce precision of estimation, there is no necessary relationship between the size of sample loss from attrition and the existence or magnitude of attrition bias. Even a large amount of attrition causes no bias if it is "random" in a sense we will define formally below. In this section we will outline our approach to addressing this issue by presenting a statistical model that distinguishes between different types of bias. We discuss the different restrictions necessary to detect and correct for each type and outline which types we will address in our empirical work.

Selection on Observables and Unobservables

Attrition bias in the econometric literature is associated with models of selection bias, and the applicability of the selection bias model to attrition was recognized early in the literature (e.g., Heckman, 1979). But recognition of the problem of nonresponse and the bias it can cause dates from much earlier in the survey sampling literature (see Madow et al., 1983, for a review). Here we will present a model tied more closely to econometric formulations than to those in survey sampling studies. Our setup will initially be formulated as a cross-section model but then will be modified for panel data.

---

[5]That is, individuals who died after the point of attrition cannot be identified from the PSID data as having died . This implies that the attrition rates we have calculated, even netting out those who died while in the PSID, overstate the fraction of the living population that has attrited. We use national mortality rates by age, race, sex, and year to estimate the number of attritors who have died, and then recalculate our attrition rates accordingly.

We assume that the object of interest is a conditional population density f(y|x) where y is a scalar dependent variable and x is (for illustration) a scalar independent variable. We will work at the population level and ignore sampling considerations. Define A as an attrition dummy equal to 1 if an observation is missing its value of y because of attrition and 0 if not (we assume for the moment that x is observed for all, as would be the case if it were a time-invariant or lagged variable). We therefore observe (or can estimate) only the density g(y|x,A = 0). The problem is how to infer f from g. By necessity this will require restrictions of some kind.

Although there are many restrictions possible (in fact, an infinite number), we will focus only on a set of restrictions that can be imposed directly on the attrition function, which we define as the probability function Pr(A = 0|y, x, z). Here z is an auxiliary variable which is assumed to be observable for all units (e.g., a time-invariant or lagged variable) but distinct from x, and whose role will become clear momentarily. The variable y is partially unobserved in this function because it is not observed if A = 1.

The key distinction we make is between what we term **selection on observables** and **selection on unobservables**.[6] We say that selection on observables occurs when

$$Pr(A = 0|y, x, z) = Pr(A = 0|x, z) \quad \textbf{(1)}$$

We say that selection on unobservables occurs simply when (1) fails to hold; that is, when the attrition function cannot be reduced from Pr(A = 0|y, x, z).[7]

---

[6]These terms have not, to our knowledge, been utilized in the literature on sample selection models (i.e., models where a subset of the population is missing information on y). However, the terms have been used in the treatment-effects literature, most extensively and explicitly by Heckman and Hotz (1989) but also by Heckman and Robb (1985: 190). The concept of selection on observables, if not the exact term, appears much earlier in the treatment-effects literature. We should also note that the survey sampling literature often uses the terms "ignorable" and "missing at random" selection to describe what we are terming selection on observables (Little and Rubin, 1987).

[7]We could define selection on unobservables to occur when x and z drop out of the probability function, and then define selection on both observables and unobservables to occur when y, x, and z all appear in the function, but we are not particularly interested in the former case and hence will not maintain such usage.

These definitions may be more familiar when they are restated within the textbook parametric model. Letting $E(y|x) = \beta_0 + \beta_1 x$ and $Pr(A = 0|x, z) = F(-\delta_0 - \delta_1 x - \delta_2 z)$, where F is a proper cumulative distribution function (c.d.f.), we can state the model equivalently with error terms ϵ and ν as

$$y = \beta_0 + \beta_1 x + \epsilon \quad , \; y \textit{ observed if } A = 0 \tag{2}$$

$$A^* = \delta_0 + \delta_1 x + \delta_2 z + \nu \tag{3}$$

$$\begin{aligned} A &= 1 \quad \textit{if } A^* \geq 0 \\ &= 0 \quad \textit{if } A^* < 0 \end{aligned} \tag{4}$$

where ν is the random variable whose c.d.f. is F. In the context of this model, selection on unobservables occurs when

$$z \perp\!\!\!\perp \epsilon | x \quad \textbf{but} \quad \upsilon \; \tilde{\perp\!\!\!\perp} \; \epsilon | x \tag{5}$$

and selection on observables occurs when

$$\nu \perp\!\!\!\perp \epsilon | x \quad \textbf{but} \quad z \; \tilde{\perp\!\!\!\perp} \; \epsilon | x \tag{6}$$

where the symbols $\perp\!\!\!\perp$ and $\tilde{\perp\!\!\!\perp}$ denote "is independent of" and "is not independent of," respectively. If $z \perp\!\!\!\perp \epsilon | x$ and $\nu \perp\!\!\!\perp \epsilon | x$, then attrition is "random" and hence estimation on the nonattriting sample causes no bias. The selection on observables case is relatively unfamiliar in the econometrics literature, but we will show that it is relevant for the attrition problem. However, we will first deal with the more familiar case of selection on unobservables.

Selection on Unobservables

We will discuss this model only briefly because of its familiarity. Exclusion restrictions are the usual method of identifying this model, and our major goal here is to discuss the difficulty in finding such restrictions for a nonresponse model in the PSID.

Working from the parametric form of the model, the conditional mean of y in the nonattriting sample can be written

$$\begin{aligned} E(y|x, z, A = 0) &= \beta_0 + \beta_1 x + E(\epsilon|x, z, \nu < -\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h(-\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h'(F(-\delta_0 - \delta_1 x - \delta_2 z)) \end{aligned} \quad \textbf{(7)}$$

where h and h′ are functions with unknown parameters. Moving from the first to the second line of the equation requires that the joint distribution of ϵ and ν be independent of x and z, so that the conditional expectation depends on x and z only through the index. Moving from the second to the third line simply replaces the index by its probability, which is permissible since they have a one-to-one correspondence.

Early implementations of this model assumed a specific bivariate distribution for ϵ and ν, leading to specific forms of the expectation function (e.g., the inverse Mills ratio for bivariate normality), while more recent implementations have relaxed some of the distributional assumptions in the model by estimating functions h or h′ whose arguments are either the attrition index or the attrition probability, respectively (see Maddala, 1983, for a textbook treatment of the early approach and Powell, 1994: 2509–2510, for discussions of the more recent approach). Armed with estimates of the parameters of the attrition index or of the predicted attrition probability, equation (7) becomes a function whose parameters can be consistently estimated.[8]

However, aside from nonlinearities in the h, h′, and F functions, identification of β requires an exclusion restriction, namely, that a z exist satisfying the independence property from ϵ and for which δ_2 is nonzero. Such a variable is often loosely termed an "instrument," although most estimation methods proposed for equation (7) do not take a textbook instrumental-variables form. Finding a suitable instrument for unobservable selection is more difficult in the case of nonresponse than in some other

[8]If nonparametric methods are used to estimate h and h′, not all of the parameters in β (e.g., the intercept) may be identifiable. We should also note at this point that if x is time-varying, then it is necessarily missing for attritors and hence the attrition propensity equation cannot be estimated as we have written it. Additional assumptions are then required to estimate the model. For example, adding time subscripts, one could assume $x(t) = a_0 + a_1 x(t-1) + a_2 z + u(t)$, thus letting x be a function of lagged x and z (alternatively, some different z′ could be specified). Substituting this equation for x(t) into the attrition equation would permit estimation provided x(t-1) is available for all observations. This procedure, however, introduces another potential source of selection bias from nonindependence of u(t) and $\epsilon(t)$.

applications because there are few variables affecting nonresponse that can be credibly excluded from the main equation for y. While this depends on the specific model under consideration, on a priori grounds personal characteristics such as those generally included in x are unlikely to be promising sources of instruments since most such characteristics are related to behavior in general and hence to y.

More promising are variables external to the individual and not under his control, such as characteristics of the interviewer or the interviewing process, or even interview payments. Although we have proposed no explicit behavioral model of attrition, a natural theory would be a simple benefit-cost model in which an individual compares the value of participating in the survey to the value of not participating. Good interviewers or interviewing conditions lower the cost of participation, and interview payments directly increase the value of participation. However, a suitable instrument must vary across respondents, and must vary in a manner independent of y. The staff at the Institute for Survey Research who have administered the PSID have assigned interviewers on the basis of respondent characteristics and have also varied interviewing conditions (length of interview, in-person vs. telephone, number of callbacks, etc.) entirely and only on the basis of respondent characteristics; consequently there is no exogenous component to the variation in treatment. This rules these variables out as instruments.

Moreover, no exogenous variations in interview payments have occurred over the course of the PSID, because payments have been adjusted only for inflation over time and vary within year only on the basis of interview mode. Based on these and other considerations we discuss in our background report (Fitzgerald et al., 1997a), we conclude that there are no instruments for nonresponse in the PSID which are credibly exogenous to behavior in general.[9]

Although we will therefore not test for selection on unobservables directly, or correct for such selection, indirect tests for selection on unobservables can be conducted whenever an outside data set is

[9]Exclusion restrictions are only one form of information. For an example of the use of other types of information, see Manski (1994). Fitzgerald et al. (1997a) provide some simple bounds calculations of one type proposed by Manski.

available containing validation information. Administrative data on some variables (e.g., earnings) are occasionally available, but this is the exception rather than the rule, and they are not available for the PSID.[10] However, the CPS is a heavily used outside data set which is a repeated cross section and hence not subject to the same type of attrition bias as the PSID. The CPS is subject to nonresponse itself, but not of the same order of magnitude as the 50 percent attrition rate in the PSID.[11] Hence we will use the CPS as a comparison data set and compare the marginal distributions of variables in the CPS and PSID to one another as well as compare regression coefficients in the two data sets. If selection on unobservables is present and it biases the coefficients, for example (see equation (7)), estimates from the two data sets will be different. Unfortunately, this method of comparison is useful only for cross-sectionally defined variables and not for variables which make use of the panel nature of the PSID, and hence does not offer a general solution to the problem.[12]

Selection on Observables

As we noted previously, the case of selection on observables is relatively unfamiliar in the econometrics literature. Because of this unfamiliarity, and because, unlike selection on unobservables, it is something we can actually address, we will discuss it at slightly greater length than we did the previous case.

The critical variable in the selection on observables case is z, a variable which affects attrition propensities but is presumed also to be related to the density of y conditional on x (i.e., z is endogenous to y). Such a variable can exist only if the investigator is interested in a "structural" y function which we

---

[10]See Hill (1992: 29) and Bound et al. (1994) for a discussion of validation studies using the PSID.

[11]While the magnitude of nonresponse does not map directly into the amount of bias, as we noted earlier, it would be unlikely for the CPS to be more biased than the PSID given these differences in the amounts of attrition.

[12]Imbens and Hellerstein (1996) show that such outside data sets, if taken as "truth," can be imposed on the data set of interest (e.g., the PSID) and can be used to formally test whether the data distributions in the two data sets are the same. See related work by Imbens and Lancaster (1994) and Hirano et al. (1996) along these lines.

interpret as a function of a variable x that plays a causal role in a theoretical sense; other variables (i.e., z) do not "belong" in the function. More generally, this situation will arise whenever the investigator is interested in (say) the expectation of y conditional on x and simply does not wish to condition on z. In cross-sectional data, for example, the standard Mincerian theory of human capital proposes that earnings are a function of education and experience; other variables which are jointly determined with earnings, like occupation and industry, should not be controlled for to obtain the "correct" estimates. Yet use of any sample that is selected on the basis of occupation and industry (e.g., only certain occupations and industries are included) will clearly bias the estimates of the earnings equation. The variable z is thus an "auxiliary" endogenous variable. As we will discuss below, in the panel data case, a lagged value of y can play the role of z if it is not in the "structural" model and if it is related to attrition.

In the presence of selection on such an endogenous variable, it is easy to show that least squares estimation of equation (2) on the nonattriting sample will generate inconsistent estimates of β and, more generally, that the estimable density $g(y|x, A = 0)$ will not correspond to the complete-population density $f(y|x)$ since the event $A = 0$ is related to y through z. Apart from this selection on observables bias, using as much of the lagged information in the panel as possible helps reduce the amount of residual, unexplained attrition variation left over in the data, and this will reduce the scope for selection on unobservables.

In the Appendix, we show formally that, under the selection on observables restriction given in equation (1), the complete-population density $f(y|x)$ can be computed from the conditional joint density of y and z, which we denote by g:

$$f(y|x) = \int_z g(y, z|x, A = 0)\, w(z, x)\, dz \qquad \textbf{(8)}$$

where

$$w(z,x) = \left[\frac{Pr(A = 0|z,x)}{Pr(A = 0|x)}\right]^{-1} \quad \textbf{(9)}$$

are normalized weights. The numerator of equation (9) inside the brackets is the probability of retention in the sample and is, in the parametric model described above, $F(-\delta_0 - \delta_1 x - \delta_2 z)$. Because both the weights and the conditional density g are identifiable and estimable functions, the complete-population density $f(y|x)$ is estimable, as are its moments such as its expected value ($\beta_0 + \beta_1 x$ in the parametric model).[13] Equation (8) shows that the complete-population density can be derived by weighting the conditional density by the (normalized) inverse selection probabilities; in the parametric model, it can be shown that this implies that WLS can be applied to equation (2) using the weights in equation (9).

We should emphasize that the application of WLS in this case is unrelated to the heteroskedasticity rationale appearing in most econometrics texts. It is also not in conflict with the conventional view among many applied economists that survey weights can be ignored because they do not affect the consistency of ordinary least squares (OLS) coefficients, since survey weights are often intended only to adjust for sample designs which have stratified the population or differentially sampled it by variables that are exogenous. Here, however, selection is indirectly on the dependent variable, and not adjusting for attrition results in loss of consistency.

If z is not a determinant of attrition, the weights in equation (9) equal 1 and hence all conditional densities equal unconditional ones and no attrition bias is present. Alternatively, if y and z are independent conditional on x and A = 0, the density g in equation (8) factors and it can again be shown that the unconditional density $f(y|x)$ equals the conditional density, and there is no attrition bias.

---

[13]As we noted in footnote 8, if contemporaneous x is unobserved and hence the attrition probability equation cannot be estimated, lagged x or additional z variables are required.

While these results are relatively unfamiliar in the econometric literature, they are pervasive in the survey sampling literature, where they form the intellectual justification for the construction and use of attrition-based survey weights (Rao, 1965, 1985; Little and Rubin, 1987: 55–60).[14,15] In the econometrics literature, while weighting formulations are sometimes used as a framework for discussing selection models (e.g., Heckman, 1987), the main point of contact with the models discussed here is the choice-based sampling literature (for discrete y, see Manski and Lerman, 1977, for an early treatment and Amemiya, 1985, for a textbook treatment; for continuous y, see Hausman and Wise, 1981, Cosslett, 1993, and Imbens and Lancaster, 1996). That literature generally considers estimation and identification in samples which are selected directly on the dependent variable, y; weighted maximum likelihood or least squares procedures are often proposed to "undo" the disproportionate endogenous sampling. The difference in the attrition case is that selection is on an auxiliary variable (z) and not on y itself; but otherwise the solutions are closely related.[16]

It should also be noted that simply conditioning on z does not solve the problem. This can be seen most simply by observing that the object of interest in most models is $E(y|x)$, not $E(y|x, z)$. Including z in the regressor set will generate "biased" coefficients on x in a linear regression model, for

---

[14]For an exception, see Cosslett (1993: 31–32). In addition, after the first draft of this paper we discovered an independent treatment of the selection on observables case by Horowitz and Manski (forthcoming), who show that the mean of a function of y can be consistently estimated with weights of the type we have discussed under the same restrictions.

[15]We should note that the weights discussed in the survey sampling literature sometimes differ from the weights in our model in two respects. First, many survey weights—including those in the PSID—are also intended to capture nonrandom sampling at the initial stage (e.g., from stratified designs). That is not the purpose of the weights we have discussed and requires a slightly different formulation to justify. Second, the weights in our model are not the type of "universal" weights generally computed for many survey data sets; "universal" weights are designed to be all-purpose and usable for any variable or model, whereas our weights are model-specific because one can easily imagine using different attrition equations (e.g., with different lagged y's) depending on the model being estimated and its definition of y.

[16]We wish to emphasize that WLS is not the only estimation method—there are many (imputation, generalized method of moments, various forms of maximum likelihood)—nor is it efficient; in addition, there are many issues connected with the use of weights which we do not discuss here. The major advantage of WLS is that it produces consistent estimates and is relatively easy to implement.

example, in the sense that it will not estimate the effect of x on y unconditional on z. Because z is an endogenous variable, it distorts the conditional distribution of y on x. Hence correcting for selection on observables is to be sharply distinguished from the corrections for unobservable selection shown in equation (7), which involve conditioning on functions of x and z; those methods are not appropriate for this case.

Testing

The application of the selection on observables model to attrition in panel data is straightforward if a lagged value of y (e.g., y at the initial wave of the panel, when all observations are present) plays the role of z, assuming that attrition is affected by such a lagged value. Lagged values of y will, assuming serial correlation in the y process, be related to current values of y conditional on x. The use of lagged values of y in this role requires the same distinction we noted earlier between structural and auxiliary determinants of contemporaneous y, because the use of lagged y as a z makes sense only if the investigator is interested, for theoretical or other purposes, in functions of y not conditioned on those lagged values.[17]

As noted previously, two sufficient conditions for the absence of attrition bias on observables are that the weights equal 1 (i.e., z does not affect A) and that z is independent of y conditional on x. Specification tests for selection on observables can be based on either of these two conditions. Thus one test is simply to determine whether candidate variables for z (e.g., lagged values of y) significantly affect A. We will conduct these tests extensively in our empirical work. A second test would be to conduct specification tests for whether OLS and WLS estimates of equation (2) are significantly different, which

[17]An investigator who posits a theoretical (i.e., structural) model that includes all lags of y will necessarily have much reduced scope for selection on observables. Taking this point to its extreme, if there are no observables in the data set that are excluded from the structural y function, there is no role for using observables to adjust for selection. Selection on observables is a data-set-defined and model-defined category, and what is an observable variable in one data set or model may be an unobservable in another.

is an indirect test for whether the identifying variables used in the weights are endogenous (see DuMouchel and Duncan, 1983, for an example of such a test). We will not conduct such tests in our paper but instead leave them for future research. However, we will determine whether using the universal weights provided by the PSID staff affect the estimated coefficients of several models, even though the "model based" weights we have been discussing are not necessarily the same as the PSID universal weights (see footnote 15).

Another test for selection on observables which we will perform is based on an exercise performed by Becketti et al. (1988) and which we term the BGLW test. In the BGLW test, the value of y at the initial wave of the survey, which we denote by y_0, is regressed on x and on future A (i.e., whether the individual later attrites). The test for attrition selection is based on the significance of A in that equation.[18] This test must necessarily be closely related to the test we have already described of regressing A on x and y_0 (which is z in this case); in fact, the two equations are simply inverses of one another.

Formally, suppose that the attrition function is taken as the latent index in the parametric model, i.e.,

$$A^* = \delta_0 + \delta_1 x + \delta_2 z + \upsilon \qquad \textbf{(10)}$$

Inverting this equation, taking expectations, and applying Bayes' Rule, it can be shown that

$$E(y_0|A,x) = \int y_0 \, f(y_0|x) \; w(A,y_0,x) \; dy_0 \qquad \textbf{(11)}$$

where

$$w(A,y_0,x) = \frac{Pr(A|y_0,x)}{Pr(A|x)} \qquad \textbf{(12)}$$

[18]We assume x to be time-invariant. If it is not, this method requires that only the values of x at the initial wave be included in the equation.

which are essentially the same as the weights appearing in (9) but including the probabilities of A = 1 as well as A = 0. Equation (11) shows that if the weights all equal 1, the conditional mean of y_0 is independent of A and hence A will be insignificant in a regression of y on x and A (the conditional mean of y_0 in the absence of attrition bias is $\beta_0 + \beta_1 x$, so a regression of y_0 on x will yield estimates of this equation). As noted previously, the weights will equal 1 only if y_0 is not a determinant of A conditional on x. Thus the BGLW method is an indirect test of the same restriction as the direct method of estimating the attrition function itself.[19]

However, if the weights do not equal 1, it would be difficult to derive an explicit solution for equation (11) from the estimates of (10) that we will obtain in our attrition propensity models. To do so would require conducting directly the integration shown in (11). It would be simpler just to estimate a linear approximation to (11) by OLS, as did Becketti et al., to determine the magnitude of the effect of A on the intercept and coefficients of the equation for y_0 as a function of x. We shall therefore also estimate such equations in our empirical work. However, it should be kept in mind that this is not an independent test of attrition bias separate from that embodied in our estimates of equation (10); it is only a shorthand means of deriving the implications of our estimates of equation (10) for the magnitudes of differences in 1968 y conditional on x between attritors and nonattritors.

Panel Data and Permanent/Transitory Effects

Finally, we wish to relate the selection on observables model we have been discussing to more traditional models of attrition in panel data, and to point out a connection with permanent/transitory distinctions which we will also apply in our empirical work below. The most well-known model of attrition in the econometrics literature is that of Hausman and Wise (1979), which has been generalized

[19]In general, of course, if $v = \alpha + \beta u + \epsilon$, regressing u on v instead of v on u results in a "biased" coefficient on v (i.e., it is not a consistent estimate of the inverse of β). Nothing here contravenes that. The "coefficient" on x in a regression of y on x and A bears no simple relationship to δ_1 or δ_2 in equation (10), as can be seen from equation (11).

and extended by Ridder (1990, 1992), Nijman and Verbeek (1992), Van den Berg et al. (1994), and others (see Verbeek and Nijman, 1996, for a review). These models generally assume a components structure to the error term, sometimes including individual-specific time-invariant effects and sometimes serially correlated transitory effects, for example, and impose restrictions on how attrition relates to the components of the structure. A common assumption in some studies in the literature is that the unobserved components of attrition propensities are independent of the transitory effect but not the individual effect; in that case, simple first-differencing (among other methods) can eliminate the bias.

Our approach differs from this past work because we sharply distinguish between identifiability under selection on observables and on unobservables, a distinction not made in these past studies. Many error components models which allow attrition propensities to covary with individual components of the process can be treated within the selection on observables framework because lagged values of y can be mapped into those components. If we let z in our model stand for a vector of lagged values of y instead of a scalar, we have $\Pr(A = 0 \mid x, y_{t-1}, y_{t-2}, y_{t-3},...,y_0)$ as our attrition function. Assume full observability of those lagged values. Then any model in which the error components of the y process which covary with attrition can be uniquely mapped into the set of t values of lagged y can be captured by our selection on observables model. An example is the autoregressive model:

$$y_t = \beta_0 + \beta_1 x + \epsilon_t \tag{13}$$

$$\epsilon_t = \sum_{\tau=0}^{t-1} \rho_\tau \epsilon_\tau + \omega_t \tag{14}$$

$$A^* = \delta_0 + \delta_1 x + \sum_{\tau=0}^{t-1} \delta_{2\tau} \epsilon_\tau + \upsilon_t \tag{15}$$

Estimation of (13) on the nonattriting sample results in bias because ϵ_t is serially correlated and A* is a function of the lagged values of that error. But solving equation (13) for ϵ_τ in lagged periods, and

substituting into equation (15) for the lagged errors, leads to an equation for A* where only lagged y appear.

This example also illustrates a case in which controlling for lagged observables in the A* equation is not sufficient to avoid attrition bias, for it is necessary that the contemporaneous shock ω_t (i.e., that which is not forecastable from lagged y) be independent of ν_t conditional on the observables. For example, shocks to earnings which occur simultaneously with, not prior to, attrition from the sample, cannot be captured by lagged values of y; attrition bias from this source falls under the selection on unobservables rubric discussed earlier. However, a full conditioning on the available data on the history of y reduces the scope of possible unobservable selection, as we noted earlier, because it isolates the only remaining source of such bias to contemporaneous, nonforecastable shocks.

The general form of our attrition probability $\Pr(A = 0 | x, y_{t-1}, y_{t-2}, y_{t-3}, ..., y_0)$ is capable of capturing a large variety of alternative forms of attrition dependence on lagged y other than the simple linear form portrayed in the autoregressive case. For example, the mean of a set of lagged values of $\mathbf{y}$, $\bar{\mathbf{y}}$, is a consistent estimator (as $T \to \infty$) for the individual effect, after conditioning on observables x and assuming mean-zero transitory disturbances. The deviations of each value of y_τ from $\bar{y}$ represent transitory disturbances in each period τ. By estimating flexible forms of the attrition function which contain both $\bar{y}$ and the deviations of lagged y from $\bar{y}$ in different periods, we can determine whether attrition probabilities covary with "permanent" levels of y and with transitory shocks one period, two periods, and more periods back in time. The variance of y_τ over any specified length of past periods is yet another transform of lagged y values which may covary with attrition; this would occur if it is variability per se, not the mean or value of any set of individual disturbances, that affects whether

individuals stay in or out of the sample.[20] We will test these and other transforms of lagged y in our models.

Summary of Analyses to be Conducted

To summarize, in the following analysis of the PSID we will (i) conduct tests for the presence of attrition on unobservables by comparing cross-sectional marginals and regression coefficients in the CPS and the PSID; (ii) conduct tests for the presence of selection on observables by estimating attrition equations as a function of lagged y values as well as by regressing first-period y on future attrition; and (iii) conduct tests for "dynamic" attrition effects by estimating attrition equations as a function of lagged permanent, transitory, and other moments of the lagged y distribution.

We should note at this point that a problem with implementing procedures using lagged values of y is that those measures are available for the full sample only at the initial year of the PSID, 1968. Conditioning on values of y after 1968 necessarily opens the door to bias because some attrition has already occurred and estimation must be restricted to observations for which all data on all lagged variables in the equation are available. Consequently, for the most part, we will restrict our tests of lags to only those available in the first year, 1968. While this approach necessarily ignores much of the information in the PSID on attritors prior to the point of attrition, it yields results least subject to the post-1968 attrition bias problem. Our dynamic attrition analysis will be an exception, for there we will estimate attrition hazards—that is, probabilities of exit conditional on being in the sample the previous period—as a function of all the lags available up to each decision point. That analysis will be conducted ignoring the potential bias induced by this sample restriction (usually called "unobserved heterogeneity"

[20]Formal modeling of the error process of y could be conducted here, but we will leave that for future research and will only test various transforms of lagged y in a reduced-form context.

in duration analyses); consequently, no "structural" interpretation will be given to the estimated coefficients in those attrition equations.[21]

III. OBSERVABLE CORRELATES OF ATTRITION IN THE PSID

Rather than begin our analysis with the comparison of the PSID to the CPS, we will first examine the observable correlates of attrition in the PSID, primarily focusing on characteristics, any one of which could be a "y" or an "x," in 1968. We will also estimate attrition probability equations as a function of 1968 characteristics for selected "y" variables and will conduct BGLW tests in this section.

The latest year of the PSID available at the time our data files were created was 1989. We focus on the seemingly simple question of whether 1968 characteristics differ between those who were present in 1989 and those who were not (hence the distributions of x and y conditional on A, in a tabular form).[22] For our analysis sample, we take every individual who was present in a PSID household in 1968, or about 18,000 individuals, as noted previously. We disaggregate the sample by sex and 1968 household headship status, and focus on five population subgroups: male heads, wives, female heads, male nonheads, and female nonheads. This asymmetric treatment of men and women is required by the gender-specific definitions of headship in the PSID, and the division of groups by headship in the first place is required because sharply differential amounts of information were collected on heads and nonheads (many variables are not available for the latter group).[23] We also exclude subfamily heads from

---

[21]Note, however, that a bias in the structural coefficients of attrition hazards does not affect the consistency of the WLS estimator using the predicted probabilities from those equations as weights. The selection on observables model does not require independence of z and ν in equation (3).

[22]In our background report (Fitzgerald et al., 1997a), we also conduct analyses of the middle year, 1981, because that was the latest year analyzed by Becketti et al. (1988). The issue that analysis addresses is whether any attrition bias we find has arisen since the Becketti et al. study was conducted.

[23]The PSID makes no distinction between male heads similar to that made between wives and female heads, for all married women are automatically classified as wives. The PSID also incorporates cohabitation to a degree: any couple living together in a "partner" status for more than one interview is then and thereafter treated as

the PSID because they were defined inconsistently over time and also differently than in the CPS, whose comparisons to the PSID are an important part of our analysis.

For the bulk of our work, we include the SEO oversample together with the SRC representative sample. We therefore use PSID-constructed 1968 sample weights whenever appropriate.[24] However, we also provide estimates on the SRC sample alone and show that attrition effects are sometimes worse for that sample than for the combined SEO-SRC sample.

Distributions of 1968 Characteristics

Table 2 shows the mean values of 1968 characteristics of men aged 25–64 and household heads in 1968 by their attrition status as of 1989—"Always In" versus "Ever Out" by that year.[25] As the first two columns indicate, attritors and nonattritors have many significant differences in characteristics. Attritors are more likely to be on welfare, less likely to be married, and are older and more likely nonwhite. In addition, attritors have lower levels of education, fewer hours of work, less labor income, and are less likely to own a home and more likely to rent.[26] The clear implication of this pattern is that attritors are concentrated in the lower portion of the socioeconomic distribution. The second moments for labor income in the table indicate that the variance of labor income is greater among attritors than among nonattritors, and, interestingly, that the attritor labor income distribution is more dispersed at the upper

---

"married"—the male is classified as a "head" and the female is classified as a "wife." We include them in our sample.

[24]These weights reflect only the sample design of the PSID (and initial nonresponse) and contain no adjustments for attrition. Hence they are not the types of weights we were discussing in Section II. However, they must be used because the SEO observations were sampled on variables that are correlated with income, which is closely related to many of our dependent variables.

[25]Because only a tiny fraction of attritors ever return (see Table 1), those individuals who were "Always In" between 1968 and 1989 are almost identical to the set of individuals present in 1989, and the set of individuals who were "Ever Out" between 1968 and 1989 is almost identical to those who were nonresponse in 1989.

[26]All monetary figures in the paper are in real 1982 dollars using the personal consumption expenditure deflator. The top and bottom 1 percent of the labor income variable is excluded to circumvent top-coding problems and to avoid distortion from outliers.

TABLE 2
1968 Characteristics by Attrition Status: Male Heads, Aged 25–64

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Welfare participation (%)	0.8	1.3	1.4	1.2
Marital status (%):				
Married	95.8	90.1*	87.1	98.1
Never married	2.4	3.7*	4.9	0.4
Widowed	0.3	1.5*	2.0	0.1
Divorced/separated	1.2	4.6*	5.9	1.3
Percent with annual hours worked > 0	98.7	94.1*	95.7	89.8
Annual labor income	21345	17011	17277	16298
Annual labor income for those w/ income > 0	21631	18152	18106	18281
Annual hours worked for those w/ hours > 0	2378	2246	2268	2182
Variance of log annual labor income for those w/ income > 0	.248	.529	.481	.667
Labor income quintile ratios for those w/ labor income > 0:				
Quintile 20/median	.658	.611	.615	.558
Quintile 40/median	.886	.905	.923	.865
Quintile 60/median	1.101	1.139	1.123	1.164
Quintile 80/median	1.392	1.498	1.462	1.493
Education (%):				
< 12	31.5	52.5*	50.8	57.2
12	32.8	25.6*	27.3	21.0
12–15	15.8	11.5*	11.5	11.5
16+	19.9	10.4*	10.4	10.4
Race (%):				
White	92.7	88.3*	87.4	90.7
Black	6.6	10.7*	11.5	8.0
Region (%):				
Northeast	24.7	25.8	26.9	22.3
North Central	32.2	27.5*	26.5	30.1
South	26.7	30.1*	29.6	31.2
West	16.4	16.7	17.0	15.7

(table continues)

TABLE 2, continued

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Age	40.7	45.6*	43.1	52.1
Tenure (%):				
Own home	74.9	62.9*	58.0	75.9
Rent	21.5	33.8*	38.9	20.2
Number of children in family	2.0	1.5	1.6	1.3
Sample size	1238	1533	1116	417

Note: Sample weights used.

*Significantly different from “Always In” at 10 percent level.

tail than is the nonattritor distribution. This suggests that, to some degree, some high labor-income families may be more likely to attrite than middle-income families.[27]

The last two columns in Table 2 provide an assessment of the effect of mortality. The third and fourth columns disaggregate the "Ever Out" subsample into those "Not Dead" and those "Dead" according to whether individuals died while in the PSID (as noted previously, some individuals die after attriting, of which we have no knowledge). Comparing the third column (not dead) with the first two shows that the gap between the "Always In" and "Ever Out" is sometimes narrowed by excluding the dead from the attritors, but rarely by very much; indeed, in some circumstances, the gap even increases. The latter occurs when mortality is related to a variable opposite in sign to its relation to attrition conditional on being alive; consequently, ignoring mortality actually makes the selectiveness of attrition seem milder than it actually is.

Tables 3 and 4 show the corresponding tables for wives and female heads.[28] The general findings are the same as for male heads: attritors and nonattritors frequently differ in their characteristics, and the differences cannot be explained by mortality. A few of the details do differ across demographic groups, however. Female heads have much larger differences in welfare participation, for example (female heads also have higher participation rates in the U.S. welfare system than other groups). Interestingly, the variance of labor income is smaller among attritors than nonattritors among female heads, although the differences among women are not significant. We conclude that the many significant differences in attritors and nonattritors in the PSID appear broadly across all headship and gender groups.

---

[27]A similar finding was reported by Becketti et al. (1988).

[28]In our background report (Fitzgerald et al., 1997a), we also provide tabulations for nonheads.

TABLE 3
1968 Characteristics by Attrition Status: Wives, Aged 25–64

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Welfare participation (%)	1.1	1.6	1.4	2.2
Percent with annual hours worked > 0	47.7	44.0	44.4	42.3
Annual labor income	36308	3299	3366	2960
Annual labor income for those w/ income > 0	7653	7509	7580	7128
Annual hours worked for those w/ hours > 0	1311	1315	1342	1173
Variance of log annual labor income for those w/ income > 0	1.546	1.624	1.548	2.014
Labor income quintile ratios for those w/ labor income > 0:				
Quintile 20/median	.240	.218	.222	.216
Quintile 40/median	.800	.611	.622	.557
Quintile 60/median	1.205	1.164	1.667	1.195
Quintile 80/median	2.000	1.637	2.078	1.670
Education (%):				
< 12	30.5	45.6*	44.7	50.0
12	49.1	38.7*	39.9	32.8
12–15	10.7	10.2	9.6	12.7
16+	9.8	5.5*	5.8	4.5
Race (%):				
White	92.0	89.5*	90.0	86.6
Black	7.4	9.4*	8.7	12.5
Region (%):				
Northeast	23.9	27.3*	28.3	22.4
North Central	31.7	26.4*	25.1	32.8
South	28.0	31.2*	31.8	27.9
West	16.5	15.1	14.8	16.9
Age	40.9	44.5*	43.5	49.6
Tenure (%):				
Own home	77.8	69.1*	67.9	75.5
Rent	18.8	28.5*	29.6	22.6

(table continues)

TABLE 3, continued

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Number of children in family	2.0	1.5	1.6	1.4
Sample size	1377	1043	847	196

Note: Sample weights used.

*Significantly different from “Always In” at 10 percent level.

TABLE 4
1968 Characteristics by Attrition Status: Female Heads, Aged 25–64

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Welfare participation (%)	4.3	10.5*	10.0	17.9
Marital status (%):				
Married	1.4	1.8	1.7	2.5
Never married	21.2	14.6*	14.7	13.1
Widowed	38.7	39.1	39.1	39.0
Divorced/separated	36.7	40.8	40.6	43.8
Percent with annual hours worked > 0	80.4	67.4*	67.0	73.7
Annual labor income	8199	6950	7167	3482
Annual labor income for those w/ income > 0	10214	10296	10679	4723
Annual hours worked for those w/ hours > 0	1593	1645	1676	1203
Variance of log annual labor income for those w/ income > 0	1.426	1.185	1.045	1.739
Labor income quintile ratios for those w/ labor income > 0:				
Quintile 20/median	.316	.424	.471	.438
Quintile 40/median	.737	.800	.838	.653
Quintile 60/median	1.163	1.178	1.178	2.483
Quintile 80/median	1.553	1.468	1.440	5.724
Education (%):				
< 12	45.1	49.2	46.8	88.4
12	28.3	32.4	33.7	11.6
12–15	13.8	9.6*	10.2	0.00
16+	12.8	8.8*	9.3	0.00
Race (%):				
White	80.3	76.0*	77.3	55.4
Black	18.8	23.2*	21.9	44.6
Region (%):				
Northeast	25.2	26.2	26.3	24.8
North Central	30.0	24.6*	25.6	9.3
South	25.8	27.7	25.9	57.5
West	19.0	21.4	22.2	8.4

(table continues)

TABLE 4, continued

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Age	44.9	47.4*	47.2	50.4
Tenure (%):				
Own home	45.0	40.3	40.3	40.7
Rent	50.3	55.9*	55.8	58.2
Number of children in family	1.3	1.0	1.0	1.8
Sample size	502	526	475	51

Note: Sample weights used.

*Significantly different from “Always In” at 10 percent level.

Attrition Probits

The first multivariate analysis we present consists of estimates of binary-choice models for the determinants of attrition, using the same data in the tables we have been presenting (i.e., whether having ever been nonresponse by 1989 as a function of 1968 characteristics). We therefore estimate probit equations for the probability of having ever been nonresponse by 1989.[29] As in Tables 2–4, the sample consists of all 1968 respondents aged 25–64, and all regressors are measured in 1968.

We shall also make a distinction between "x" and "y" in this analysis by focusing on three "y" variables: labor income, marital status, and welfare participation (female heads only). We select these three because they are some of the more common dependent variables used by economists and sociologists and therefore their relations to attrition are of particular interest. Our tabular analysis in Tables 2–4 showed some evidence of significant attrition effects for these key variables, which should generate some cause for concern for analysts who study these outcomes.[30] One issue that can be addressed in a multivariate analysis is whether these effects are attenuated when a set of other socioeconomic variables is controlled for in a regression framework.

Table 5 shows a set of expanding specifications of attrition probits which focus on the effect of our first "y," labor income, on the attrition of male heads. The first two columns of Table 5 show the effect of labor income on attrition without conditioning on any other regressors ("No labor income" is a dummy equal to 1 if the individual has no labor income). The results show that the 1968 labor income levels of male heads have a very strong correlation with future nonresponse. Attrition probabilities are quadratic in labor income (lowest at middle income levels and greatest at high and low income levels), a

[29]Although we do not estimate a dynamic model of year-by-year attrition, these estimates can be viewed as a model of cumulative attrition that reflects the working-out of a year-by-year model. Since all the regressors are held at their 1968 values, our equation can be viewed as an approximation to the reduced-form model.

[30]To repeat a point from Section II, the concern arises because the 1968 values of these variables are likely to covary with their later values.

TABLE 5
Ever-Out Attrition Probits
Male Heads Aged 25–64, Focus on Labor Income

	Model 1		Model 2		Model 3		Model 4	
	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X
Intercept	.334*	.128	.360*	.139	1.770*	.671	1.130*	.417
	(.059)		(.096)		(.454)		(.518)	
Labor income[a]	-.0239*	-.0092	-.0272*	-.0105	-.0192*	-.0073	-.0237*	-.0088
	(.0030)		(.0103)		(.0108)		(.0120)	
No labor income	.284*	.110	.254	.100	.291	.110	.181	.067
	(.160)		(.177)		(.180)		(.186)	
Labor income squared[b]		.009	.003	.018	.006	.022	.008	
			(.025)		(.026)		(.026)	
Black					.074	.028	.037	.014
					(.066)		(.081)	
Other race					.356	.134	.198	.073
					(.248)		(.251)	
Age					-.088*	-.033	-.039	-.014
					(.022)		(.024)	
Age squared[c]					.107*	.041	.054*	.020
					(.025)		(.028)	
Education < 12 years					.200*	.076	.208*	.077
					(.690)		(.071)	
Some college					-.114	-.043	-.195*	-.072
					(.096)		(.097)	
College degree					-.305*	-.116	-.384*	-.142
					(.107)		(.109)	
Northeast							-.051	-.019
							(.939)	
North Central							-.139	-.051
							(.091)	
South							-.120	-.044
							(.088)	
In SEO sample							-.070	-.025
							(.080)	

(table continues)

TABLE 5, continued

	Model 1		Model 2		Model 3		Model 4	
	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X
Lives in rural area (SMSA < 1000)							-.271* (.072)	-.100
Number of children in family						-.033*	-.012 (.017)	
Presence of child < 6							.095 (.061)	.035
Owns house							-.310* (.068)	-.114
Might move in future							-.015 (.072)	-.006
Income/Needs ratio							.031 (.033)	.012
R^2	.028		.028		.044		.068	
Sample size	2253		2253		2253		2253	
Number ever out	1074		1074		1074		1074	
Log likelihood	-1516.05		-1515.99		-1490.27		-1453.02	

Notes: Excludes known dead. Characteristics measured in 1968. ∂P/∂X signifies the effect of a unit change in the variable on the probability of attrition evaluated at the mean. R^2 equals 1 minus the ratio of the log likelihood of the fitted function to the log likelihood of a function with only an intercept.

* Significant at 10 percent level.

[a]Coefficients multiplied by 10^3.

[b]Coefficients multiplied by 10^8.

[c]Coefficients multiplied by 10^2.

pattern also found by Becketti et al. (1988), as noted earlier. Individuals with no labor income at all have higher attrition rates as well. The third column in the table shows that when "standard" earnings-determining variables are added—race, age, and education—labor income remains a significant determinant of attrition. Implicitly, therefore, the residual in a labor income equation containing these regressors is correlated with attrition. When a large number of other variables—income/needs, home ownership, SEO status, and others—are added, the labor income effects remain.

Table 6 shows the coefficients on the earnings variables in these models (except for the first) for wives and female heads, and also the coefficients for other 1968 "y" variables.[31] For female heads and wives, labor income effects are much weaker. For neither group is there much of an effect of labor income on nonresponse except for the effects of having no labor income at all, which continues to have a positive effect on nonresponse. For wives, even this effect is relatively weak when the larger set of covariates is included in the equation. When the earnings variables are replaced by our other two "y" variables—1968 marital status and welfare participation—rather similar patterns are found. Again, there are some significant coefficients on these variables when nothing else is controlled for, but in all cases those effects fall to insignificance at conventional levels in the most expanded specification.

Table 7 shows the coefficients in attrition probits when all three types of y variables are included. Although including the variables singly gives the best specification for comparison with the BGLW specification (which inverts the attrition probit to solve for a single y), there is no reason not to include all available data in an attrition probit intended for weight construction, or for general interest.[32] The results in the table indicate that very little is changed when multiple y variables are included; most

---

[31]The full set of regression coefficients on all models is available in Fitzgerald et al. (1997a).

[32]As we stressed in Section II, all these y variables are potentially "endogenous" in the sense that they might be related to a contemporaneous y of interest, and adding more lagged y variables to the attrition equations increases the chances of capturing such endogeneity. But it is only through the existence of such endogeneity that weights can reduce attrition bias.

TABLE 6
Ever-Out Attrition Probits: Other Results

	Model 2		Model 3		Model 4	
	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X
Wives, 25-64, Focus on Labor Income						
Labor income[a]	.0010	.0004	.0056	.0021	.0016	.0006
	(.0166)		(.0168)		(.0172)	
No labor income	.133	.051	.128	.048	.135	.049
	(.083)		(.085)		(.086)	
Labor income squared[b]	.011	.004	.021	.008	.030	.011
	(.073)		(.074)		(.075)	
Female Heads, 25-64, Focus on Labor Income						
Labor income[a]	-.0010	-.0004	-.0018	-.0007	-.0035	-.0013
	(.0195)		(.0201)		(.0214)	
No labor income	.438*	.171	.424*	.162	.424*	.160
	(.125)		(.128)		(.133)	
Labor income squared[b]	.009	.004	.0186	.007	.033	.012
	(.073)		(.074)		(.078)	
Men, 25-64, Focus on Marital Status						
Married	-.436*	-.165	-.192	-.0710	-.156	-.058
	(.134)		(.140)		(.142)	
Widowed	-.130	-.049	.054	.020	.026	.009
	(.234)		(.238)		(.239)	
Divorced/separated	.259	-.098	.255	.094	.288	.106
	(.191)		(.193)		(.194)	
Women, 25-64, Focus on Marital Status						
Married	-.182*	-.069	-.036	-.014	-.039	-.015
	(.101)		(.104)		(.106)	
Widowed	-.024	-.009	.0425	.0160	.065	.024
	(.123)		(.125)		(.126)	
Divorced/separated	.090	.034	.114	.043	.131	.049
	(.112)		(.114)		(.115)	

(table continues)

TABLE 6, continued

	Model 2		Model 3		Model 4	
	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X
Female Heads, 18-54, Focus on Welfare						
Welfare receipt	.270*	.106	.214	.083	.0704	.027
	(.139)		(.143)		(.149)	

Notes: Excludes known dead. Characteristics measured in 1968. ∂ P/∂X signifies the effect of a unit change in the variable on the probability of attrition evaluated at the mean.

*Significant at 10 percent level.

[a]Coefficients multiplied by 10^3.

[b]Coefficients multiplied by 10^8.

TABLE 7
Ever-Out Attrition Probits
Multiple-Focus Variables

	Female Heads 18–54		Men 25–64		Women 18–54	
	Coeff.	∂P/∂X	Coeff.	∂P/∂X	Coeff.	∂P/∂X
Labor Income						
Labor income[a]	-.0350	-.0130	-.0199*	-.0073	.0001	.0000
	(.0022)		(.0120)		(.0013)	
No labor income	.431*	.162	.203	.071	.221*	.082
	(.141)		(.179)		(.071)	
Labor income squared[b]	-.003	-.001	.002	.006	.000	.000
	(.008)		(.003)		(.001)	
Marital Status[c]						
Married	—	—	-.156	-.060	-.039	-.015
			(.142)		(.106)	
Widowed	.141	.053	.026	.009	.065	.024
	(.164)		(.239)		(.126)	
Divorced/separated	.249*	.094	.288	.106	.131	.049
	(.121)		(.194)		(.115)	
Welfare						
Welfare receipt	.070	.027	-.239	-.088	.083	.031
	(.149)		(.213)		(.109)	

Notes: Excludes known dead. Characteristics measured in 1968. ∂P/∂X signifies the effect of a unit change in the variable on the probability of attrition evaluated at the mean. Other variables included are those in Model 4 in Table 5.

*Significant at 10 percent level.

[a]Coefficients multiplied by 10^3.

[b]Coefficients multiplied by 10^8.

[c]Omitted category for female heads is never-married.

effects are insignificant, with the absence of labor income continuing to be the one variable with often-significant effects even after controlling for other regressors.

We should also note that the R-squareds from these probits are extremely small.[33] In Table 5 they never exceed .068 and in the models in Tables 6 and 7 they range from .028 to .071; they are even lower in Models 1, 2, and 3 when fewer other regressors are conditioned on. Thus, even in those cases where significant correlates of attrition are found, they explain very little of the variation in attrition probabilities in the data. One implication of this result is that weights based on these equations would, in all likelihood, have little effect on estimated outcome equations.[34]

We conclude from these results that the unconditional effects of labor income, welfare participation, and marital status covary significantly with attrition probabilities, consistent with our conclusions from the tabular analysis in Tables 2–4 (although the BGLW form of the test, reported next, corresponds more closely to Tables 2–4). However, we also find that, in a majority of the cases, these effects fall to insignificance at conventional levels when a sufficiently broad set of covariates are conditioned on. The main exceptions to this occur for various specifications of labor income models, particularly for male heads but occasionally as well for female heads and for women in general, and for the occasional other model. Thus these results provide support for some concern for cross-sectional attrition bias in the PSID for unconditional distributions, and for conditional distributions for earnings, especially of male heads.

---

[33]The R-squared measure we use is defined in the notes to Table 5 and is a common measure of fit in binary-choice models. This measure has recently been shown to have desirable properties relative to other measures (Cameron and Windmeijer, 1997) and can be interpreted as the proportionate reduction in uncertainty from the fitted model, where uncertainty is defined by an entropy measure.

[34]This statement must be qualified because even weights with very small variance could have a large impact if they are sufficiently highly correlated with the error term and the regressors.

BGLW Tests

In this section we report tests for attrition bias adapted from the work of Becketti et al. (1988), which we termed BGLW tests in our discussion of testing in a previous section. As we discussed there, these tests estimate the effects of future attrition on 1968 outcomes and should be thought of as an inversion of our attrition probits. Apart from nonlinearities and some differences in the stochastic assumptions, the results should have the same general tenor as the attrition probits but will show more directly the degree to which regression coefficients in typical outcome equations are affected.

Table 8 shows 1968 log labor income regressions for male heads.[35] Separate regressions are estimated for individuals who were always in the sample through our final year, 1989, and for the total sample in 1968. We compare the total sample and the nonattriting sample—not attritors and nonattritors—because the issue is how different parameter estimates would be from those in the total sample if only the nonattriting sample is used.[36] We show results separately when the SEO sample is included and excluded. For male heads, none of the coefficients on the variables of most past research interest—Black, Education < 12 years, College degree, Age, and Age squared—are significantly different between the total and nonattriting samples in estimates including the SEO, and the magnitudes of the differences in the coefficients are seldom large from a substantive research point of view. Significant differences do appear for the “Other race” and “Some college” variables (and one of the region variables) for reasons we have not been able to determine. More significant differences appear for

[35]Individuals with zero labor income are excluded. While this introduces some noncomparability with our attrition probits as well as raising well-known selection issues, we wish to maintain correspondence with the bulk of the earnings function literature, which also generally conditions on positive income.

[36]The two sets of differences are transforms of one another, but they have different standard errors. Under the null of equality of the true coefficient vectors, the variance of the difference in the coefficients is the difference in the separate variances (the variance in the smaller sample must be larger, necessarily, under the null).

TABLE 8
1968 Log Labor Income Regressions
Male Heads

	SRC and SEO Combined			SRC Only		
	Total	Always In	Difference	Total	Always In	Difference
Intercept	8.24*	8.38*	.14	8.28*	8.35*	.08
	(.197)	(.232)	(.12)	(.23)	(0.26)	(.13)
Black	-.249*	-.272*	-.022	-.173*	-.195*	-.022
	(.044)	(.056)	(.035)	(.055)	(0.070)	(.043)
Other race	-.221	-.246	.196*	-.393*	-.193	.200*
	(.136)	(.173)	(.106)	(0.164)	(.184)	(.0830)
Education < 12 years	-.293*	-.271*	.023	-.291*	-.244*	.047*
	(.034)	(.039)	(.019)	(.040)	(.045)	(.020)
Some college	.101*	.068*	-.033*	.103*	.098*	-.005*
	(.037)	(.039)	(.014)	(.042)	(.044)	(.001)
College degree	.271*	.283*	.012	.311*	.334*	.024*
	(.043)	(.045)	(.011)	(.050)	(.050)	(.008)
Age	.080*	.074*	-.059	.080*	.079*	-.001
	(.009)	(.011)	(.061)	(.011)	(.013)	(.007)
Age squared[a]	-.948*	-.856*	.092	-.947*	-.922*	.003
	(.108)	(.132)	(.075)	(.125)	(.149)	(.081)
Northeast	.076*	.110*	.034	.088*	.065	-.022
	(.039)	(.045)	(.022)	(.047)	(.052)	(.023)
North Central	.045	.006	-.039*	.013	-.056	-.069*
	(.038)	(.043)	(.020)	(.043)	(.048)	(.021)
South	-.076*	-.105*	-.028	-.111*	-.147*	-.036
	(.039)	(.045)	(.023)	(.045)	(.051)	(.025)
Sample size	2182	1159		1406	788	
R^2	.19	.24		.22	.26	
F-statistic[b]	50.5	35.7		38.8	27.8	
Variance of error	.326	.220		.285	.194	

Notes: Standard errors in parentheses. Sample excludes known dead. SRC+SEO is weighted.
*Significant at 10 percent level.
[a]Coefficients multiplied by 10^3.
[b]F-statistic for hypothesis that all coefficients except the intercept are equal to zero.

the estimates when the SEO is excluded, but these again are not large in magnitude. In summary, at least for the SRC-SEO combined sample, we find very few important effects of attrition on the coefficients.[37,38]

In our background report (Fitzgerald et al., 1997a), we show estimates of labor income equations for wives and female heads; marital status probits for men and women; and welfare-status probits for female heads, all estimated in 1968 separately for the total and nonattriting samples. For wives, the labor income results are essentially similar to those for men, although some significant differences in the magnitude (though not the sign) appear for the education coefficients. For female heads, the only significant labor-income differences are for the coefficients on age, but the separate coefficients for the total and nonattritor samples are each insignificant (a sign that female heads have very flat age-earnings profiles), so it is not clear how important this result is. In the marital-status probits, some significant differences appear for men (Black coefficient) and women (Education coefficient), generating somewhat more concern for these outcome variables than for labor income. The welfare probits show no significant differences in any of the coefficients.

Wald tests for the joint significance of the differences in all slope coefficients and intercepts generally reject the hypothesis of equality between the vectors. However, when tests are conducted for the equality of the slope coefficients allowing the intercepts to differ, most fail to reject equality. The estimated intercept differences (i.e., constraining all coefficients on the other regressors to be the same for the two groups) are shown in Table 9. Thus we conclude that, while the coefficients on "standard" variables in labor income and welfare-participation equations and, to a lesser extent, marital-status

---

[37]Similar findings were reported by Becketti et al., (1988). However, their analysis only went through 1981 and, in addition, they tested the difference in coefficients between attritors and nonattritors whereas we properly test between the total sample and nonattritors.

[38]We calculated White standard errors for the coefficients but found them to be only 5 percent higher, at most, than those shown. We therefore do not calculate them for the remainder of the analysis.

TABLE 9
1968 Income, Marital Status, and Welfare Equations:
Difference in Total and Always-In Samples, Intercept-Only Model

	SRC+SEO	SRC Only
Labor Income Regressions:		
Male heads	-.059*	-.053*
	(.012)	(.013)
Wives	.016	.007
	(.028)	(.034)
Female heads	.091*	.122*
	(.037)	(.061)
Marital-Status Probits:		
Men	-.232*	-.232*
	(.037)	(.044)
Women	-.063*	-.078*
	(.022)	(.028)
Welfare-Status Probits:		
Female heads	-.264*	-.383*
	(.087)	(.186)

Notes: Models include all variables shown in Table 8 but allow the intercept to differ for the Total and Always-In samples. Coefficient equals Total intercept minus Always-In intercept. Standard errors in parentheses. Sample excludes known dead. SRC+SEO is weighted.

*Significant at the 10 percent level.

equations, are unaffected by attrition, there are still differences in the levels of these outcome variables conditional on the regressors.

IV. CROSS-SECTIONAL COMPARISONS TO CENSUS DATA

The second part of our analysis compares cross-sectional distributions and regression coefficients between the PSID and the CPS, allowing us to conduct a more direct analysis of the existence of attrition bias for these types of variables. Comparing the PSID and the CPS has some difficulties, however. The most important is that the sampling frames are not identical, because the CPS includes individuals and families who have immigrated to the U.S. since 1968, while the PSID excludes those families.[39] This issue is of some importance and, consequently, we will present some tabulations on the characteristics of immigrants since 1968 taken from the Decennial Census in 1990. Second, many of the variables are defined differently in the two data sets (headship and labor income, for example), and hence this will generate some noncomparability.

Tables 10 and 11 show comparisons for male heads aged 25–64 in the PSID and CPS in 1968 and 1989, respectively. Table 10 compares the two data sets in 1968 and is thus relevant to the issue of whether the approximate 25 percent nonresponse in the drawing of the PSID sample systematically biased the first wave of the data. The table indicates that the distributions of age, race, education, marital status, and regional location in the CPS and PSID were roughly in line in 1968, both for the SRC sample and the combined (weighted) SRC-SEO sample.[40] A few miscellaneous divergences appear (e.g., in the educational distribution) which may be a result of different questionnaire wording. As for labor force and

---

[39]The PSID Latino supplemental sample, which includes a few immigrants, was not begun until 1990.

[40]The PSID weights in 1968 were not obtained from direct poststratification against Census or CPS distributions but were derived from combining the weights from the University of Michigan's SRC sampling frame and the Census Bureau's SEO sampling weights. The weights for the combined SRC-SEO sample were set to make the combined SRC-SEO sample representative.

TABLE 10
Characteristics of Male Heads Aged 25–64: 1968 PSID and CPS

	CPS	PSID	
		Weighted (SRC and SEO)	Unweighted (SRC only)
Age	43.7	43.3	43.6
Race			
White	0.91	0.9	0.91
Black	0.08	0.09	0.08
Hispanic	—	—	—
Education			
< 12	0.42	0.43	0.41
12	0.32	0.29	0.3
13–15	0.11	0.14	0.14
16+	0.15	0.15	0.15
Marital Status			
Never married	0.03	0.03	0.03
Married	0.92	0.93	0.94
Divorced/separated	0.03	0.03	0.03
Widowed	0.01	0.01	0.01
Region			
Northeast	0.25	0.25	0.22
North Central	0.28	0.3	0.31
South	0.29	0.28	0.3
West	0.18	0.17	0.17
Own Home	—	0.69	0.71
Labor Force			
Positive weeks worked	0.96	0.96	0.96
Weeks worked[a]	—	—	—
Annual hours worked[a]	—	—	—
Earnings[a]			
Real wage and salary	$19478	—	—
Real labor income	—	$20460	$20709

(table continues)

TABLE 10, continued

	CPS	PSID	
		Weighted (SRC and SEO)	Unweighted (SRC only)
Wage and Salary Distribution[a]			
Log variance[b]	0.452	0.389	0.354
Ratios of percentile points to median[b]			
20th percentile	0.671	0.667	0.667
40th percentile	0.886	0.893	0.907
60th percentile	1.114	1.087	1.107
80th percentile	1.429	1.373	1.4
Welfare Participation	0.02	0.01	0.01

[a]Workers only.

[b]PSID figures use labor income rather than wage and salary income.

earnings, neither the CPS nor the PSID has unbracketed variables for weeks worked or annual hours worked in 1968, so only the fraction of those with positive weeks worked can be compared, and in this dimension the PSID again lines up with the CPS. In addition, the PSID unfortunately did not obtain an unbracketed earnings variable in 1968, so we must rely on a measure of labor income, which includes some earned income other than wages and salaries.[41] The means of the two earnings measures are about $1,000 apart in the two data sets, and a bit further apart if the SRC sample is used. Whether this is a result of the difference in the measures cannot be ascertained. Table 10 also shows measures of dispersion in the two data sets, although these are also contaminated by the differences in measures. The log variance of earnings is considerably smaller in the PSID than in the CPS, but the measures of percentile points are not far apart, suggesting that differences at the very lowest percentiles are driving the difference.[42]

Statistical tests for the differences in the distributions almost always reject equality of the distributions because the standard errors from the CPS, with its very large sample sizes, are extremely small. However, the magnitudes of the differences in most of the variables are small from a substantive research point of view, so we shall continue to make comparisons along this dimension rather than through formal statistical tests.[43]

Table 11 shows the comparable distributions in 1989. This table has two columns for the combined SEO-SRC PSID sample, one using 1968 weights and one using the 1989 weights calculated by

---

[41]The PSID procedure for creating labor income is described in Institute for Social Research (1972: 307+). We exclude from our calculations those with zero wage and salary income and those who said on a separate question that they were self-employed. Our CPS wage and salary measure therefore also excludes individuals with self-employment income.

[42]The log variance is sensitive to changes in the lower tail of the distribution.

[43]However, on the more important issue of differences in regression coefficients, we will rely more heavily on tests of differences. See below.

TABLE 11
Characteristics of Male Heads Aged 25–64: 1989 PSID, CPS, and PUMS

	PUMS			PSID		
	With Immigrants	Without Immigrants	CPS	1989 Weights (SRC and SEO)	1968 Weights (SRC and SEO)	Unweighted (SRC Only)
Age	42.4	42.7	42	42	42	42.2
Race						
White	0.86	0.89	0.89	0.9	0.92	0.93
Black	0.08	0.08	0.08	0.09	0.07	0.06
Hispanic	0.07	0.05	0.07	0.03	0.02	0.01
Education						
< 12	0.17	0.16	0.17	0.18	0.18	0.17
12	0.28	0.29	0.36	0.29	0.29	0.29
13–15	0.27	0.28	0.19	0.23	0.23	0.23
16+	0.27	0.27	0.28	0.29	0.3	0.31
Marital Status						
Never married	0.1	0.1	0.1	0.08	0.09	0.08
Married	0.79	0.79	0.79	0.81	0.81	0.82
Divorced/separated	0.1	0.1	0.09	0.09	0.09	0.09
Widowed	0.01	0.01	0.01	0.01	0.01	0.01
Region						
Northeast	0.2	0.19	0.2	0.22	0.23	0.2
North Central	0.25	0.26	0.25	0.28	0.28	0.3
South	0.34	0.35	0.34	0.31	0.31	0.32
West	0.21	0.2	0.21	0.18	0.18	0.17
Own Home	0.72	0.74	0.71	0.73	0.74	0.75

(table continues)

TABLE 11, continued

	PUMS			PSID		
	With Immigrants	Without Immigrants	CPS	1989 Weights (SRC and SEO)	1968 Weights (SRC and SEO)	Unweighted (SRC Only)
Labor Force						
Positive weeks worked	0.92	0.92	0.89	0.93	0.93	0.94
Weeks worked[a]	48.1	48.3	49	46.6	46.6	46.7
Annual hours worked[a]	2156	2164	2165	2172	2176	2199
Earnings[a]						
Real wage and salary	$24239	$24582	$22970	$23481	$23645	$23905
Real labor income	—	—	—	$24090	$24273	$24537
Wage and Salary Distribution[a]						
Log variance	0.63	0.61	0.624	0.501	0.491	0.452
Ratios of percentile points to median						
20th percentile	0.557	0.571	0.566	0.582	0.571	0.589
40th percentile	0.857	0.886	0.868	0.873	0.873	0.875
60th percentile	1.117	1.143	1.132	1.163	1.143	1.143
80th percentile	1.5	1.525	1.509	1.519	1.5	1.5
Welfare Participation	0.02	0.02	0.02	0.01	0.01	0.01

[a]Workers only.

the PSID staff and including an attrition adjustment.[44] Some differences between the PSID and CPS appear but they are not large and are often narrowed slightly by the weights. For example, the higher attrition rate for blacks can be seen from the slightly lower percentage black for the 1968-weight PSID (.07) versus the 1989-weight PSID (.09). The SRC-only sample is the worst (.06), no doubt because no attrition-adjusted weights have been calculated for that sample. Nevertheless, for race, age, education, marital status, and region, the differences between the CPS and the PSID, and among the different PSID samples, is quite small and gives an overall impression of fairly strongly continued representativeness of the PSID for male heads, even through 1989.

In addition, the PSID has a wage and salary earnings variable in 1989 which allows a better comparison with the CPS on this score than was the case for 1968. In 1989 the two are within $500 of each other, only half of the $1,000 difference in 1968. The continued difference with the labor income variable suggests that much of the 1968 difference was indeed a result of noncomparability of variables. For earnings itself, the 1989-weight PSID is the closest to the CPS, followed by the 1968-weight PSID and then by the SRC-only, which is the furthest from the CPS.

As for dispersion, the log variance measures in the PSID are still smaller in 1989 when comparable measures are used (the SRC-only sample continues to be the furthest from the CPS). Again, however, the percentile point measures are reasonably close in the different data sets, perhaps suggesting that the log variance measures are affected by outliers at the bottom of the distribution. The percentile measures show strong increases in dispersion over time (compare Tables 10 and 11), consistent with the evidence now recognized of increasing earnings inequality among men in the U.S. This comparability was also noted previously by Gottschalk and Moffitt (1992).

---

[44]The construction of these attrition-adjusted weights is described in Institute for Social Research (1992: 82–98). The variables included in the attrition equation are age, gender, race, education, number of children, region, lagged family income, and others.

It is necessary to reconcile these findings, which indicate that the PSID has roughly maintained representativeness through 1989 for the unconditional means and distributions of major sociodemographic lines, with those from the previous analysis indicating significant differences between attritor and nonattritor unconditional characteristics in 1968 (Tables 2–4).[45] Taking both results at face value, they necessarily imply that the differences in the value of the variables for the two samples in 1968 must have converged over time. Further investigation of this possibility reveals it indeed to be the case, as we demonstrate in Tables 12 and 13. Table 12 shows the characteristics of PSID males who were 25–40 in 1968 and therefore were 46–61 in 1989, but including in the 1968 sample only those men who responded in 1989. Consequently, the sample is composed of the same individuals in both years (unlike Tables 10 and 11, the former of which includes some men who have attrited or died by 1989 and the latter of which includes a second generation). The table also shows CPS tabulations of men in these same age groups in the same years. It is clear that, while time-invariant characteristics such as race must necessarily remain as far apart between the data sets in 1989 as they were in 1968, this is not the case for time-varying characteristics. Indeed, the distributions of education and marital status change over time for the PSID men in a way that reduces the initial selection and moves the distributions closer to the CPS. The initial selection in the PSID on men with more education is offset by a slower rate of growth of education over the life cycle among nonattriting individuals in the PSID than in the CPS, and the initial selection on married men is partly offset by a more rapid decline in marriage rates in the PSID than in the CPS. The analysis of earnings is complicated by the noncomparability of measures, but the growth of labor income in the PSID was much smaller than the growth of earnings in the CPS, thus partly offsetting the initial selection on relatively high-income men in the PSID.

---

[45]Actually, the differences are a bit exaggerated because Tables 2–4 compare attritors to nonattritors instead of the total sample to nonattritors, which is the implicit comparison in the CPS analysis. At an approximate attrition rate of 50 percent, the differences shown in Tables 2–4 should be halved for comparison with the CPS. This by itself somewhat reduces the perceived seriousness of the discrepancy.

TABLE 12
Characteristics of Males Aged 25–40 in 1968 and 46–61 in 1989 PSID and CPS

	CPS		PSID	
	25–40 in 1968	46–61 in 1989	25–40 in 1968[a]	46–61 in 1989
Age	32.4	53.1	32.8	53.8
Race				
White	0.89	0.87	0.93	0.92
Black	0.09	0.1	0.06	0.06
Education				
< 12	0.31	0.25	0.25	0.27
12	0.38	0.36	0.34	0.3
13–15	0.13	0.14	0.17	0.18
16+	0.18	0.24	0.22	0.26
Marital Status				
Never married	0.12	0.06	0.02	0.02
Married	0.83	0.8	0.95	0.86
Divorced/separated	0.04	0.12	0.01	0.1
Widowed	0.02	0.02	0.01	0.02
Region				
Northeast	0.24	0.21	0.26	0.25
North Central	0.28	0.25	0.3	0.28
South	0.3	0.35	0.29	0.31
West	0.18	0.19	0.15	0.16
Own Home	—	0.89	0.66	0.86
Earnings[b]				
Real wage and salary	$18429	$24694	—	$25464
Real labor income	—	—	$21265	$24638

Note: PSID sample includes SEO and SRC, and both years use 1968 weights.

[a]Sample includes only those responding in 1989.

[b]Workers only.

The simplest explanation for this pattern is that the time series processes for education, marital status, and earnings contain a serially correlated component which at least partly regresses to the mean, and that selection is at least partly based on that component. The existence of autoregressive moving average (ARMA) errors, after a time-invariant or even unit root component has been controlled for, has been amply demonstrated in the literature on earnings dynamics (MaCurdy, 1982; Abowd and Card, 1989; Moffitt and Gottschalk, 1995); the transitory components in these models do not fade out very quickly over time, at least in levels. In the next section, where we more directly examine attrition dynamics, we will show explicitly that attrition is based on lagged shocks which are deviations from average levels, although contemporaneous shocks cannot be directly examined.

A similar regression-to-the-mean effect appears to be at work in the PSID across generations, although milder in magnitude (see Fitzgerald et al., 1997b, for a fuller examination of intergenerational attrition issues). Table 13 shows the original Table 11 for 1989 split out between those 25–45 and those 46–64; the former were mostly children in 1968 and hence constitute the "second generation" that was implicitly contained in Table 11. The CPS-PSID differences are often slightly narrower for the younger generation than for the older, as can been seen from the percentage with less than 12 years of education, the percentage married, and the percentage owning a home. The pattern is not uniform across all categories, however. Nevertheless, for many categories the data are consistent with an intergenerational model with similar serially correlated mean-regressing components.

Returning to Table 11, it can be seen that a second explanation for the comparability with CPS is a small role played by the updating of the PSID weights for attrition on observables. The PSID staff readjusts its weights over time to take into account both differential mortality by age, race, and sex and differential nonresponse (Institute for Social Research, 1992: 82–98). The latter adjustment is based on an estimated nonresponse model in which nonresponse probabilities for different time intervals since 1968 are made a function of past socioeconomic characteristics such as age, race, sex, income, family

TABLE 13
Characteristics of Male Heads Aged 25–45 and 46–64 in 1989 PSID and CPS

	Age 25–45		Age 46–64	
	CPS	PSID	CPS	PSID
Age	34.9	34.8	54.6	55.3
Race				
White	0.88	0.92	0.89	0.92
Black	0.08	0.07	0.08	0.06
Education				
< 12	0.12	0.12	0.25	0.28
12	0.36	0.3	0.35	0.29
13–15	0.22	0.26	0.15	0.17
16+	0.3	0.32	0.25	0.26
Marital Status				
Never married	0.14	0.12	0.04	0.01
Married	0.76	0.78	0.84	0.88
Divorced/separated	0.09	0.09	0.1	0.09
Widowed	0	0	0.02	0.02
Region				
Northeast	0.2	0.21	0.21	0.25
North Central	0.25	0.28	0.25	0.28
South	0.34	0.31	0.34	0.29
West	0.22	0.18	0.19	0.17
Own Home	0.64	0.66	0.83	0.88
Earnings[a]				
Real wage and salary	$22096	$23162	$24878	$25262
Real labor income	—	$23622	—	$25890

Note: PSID sample uses SRC-SEO and 1968 weights.

[a]Workers only.

structure, urban/rural location, and regional location. The predicted nonresponse probabilities from the model are used to adjust the weights for each member of the sample on the basis of his or her characteristics. This procedure is capable, in principle, of adjusting for attrition on observables, as discussed in Section II, even though these are "universal" weights rather than model-specific weights.[46]

Comparison of the columns for 1989-weight and 1968-weight estimates in Table 11 shows that this adjustment has an effect on the PSID means for only a few variables. The adjustments are generally (but not always) in the "right" direction—that is, moving the PSID means closer to those in the CPS. This is particularly the case for the race distribution, where the percentage white is improved by this adjustment. The labor force and income variables are likewise moved slightly toward the CPS by the weight adjustment.[47] Nevertheless, the magnitude of the changes resulting from the weight adjustment is generally quite small. The major reason for this result is that, despite the correlation of observables with attrition propensities, attrition remains mostly noise. This was clear from the low R-squared values reported in our attrition probits. The variances of the predicted attrition rates from those probits are small, which necessarily implies that the variance of attrition-adjusted weights is small; weighting may have little effect in this case (subject to the caveat mentioned previously).

Although we have now provided explanations for the closeness of the CPS and PSID cross-sectional distributions, we note that there are some remaining differences. These can be further narrowed once immigration into the U.S. since 1968 is accounted for. The importance of immigration is illustrated

---

[46]We say "in principle" because it is necessary that the nonresponse model be properly specified for the adjustment to restore representativeness. It is worth emphasizing that no outside benchmarks from the CPS or other data set are used for these nonresponse adjustments. The adjustments are all "internal" and result only in a multiplication factor being applied to the prior year's weights to obtain current weights. See footnote 44.

[47]However, Table 11 also suggests a problem with the PSID weight because time-invariant characteristics, such as race, are capable of perfect attrition adjustment since the true population means of those variables must be the same as they were in 1968; hence it is easy to calculate a weight that perfectly restores the 1968 mean. But if the weights are based on nonresponse models which are parametric functions of several variables (like race), and hence smooth them over, the resulting weights will never fully adjust any single variable, even time-invariant ones. This is a problem with all universal weights.

in Table 11, which shows means for male heads in 1989 taken from the 1990 Decennial Census Public Use Microdata Sample (PUMS). Although the CPS did not, as of 1989, ask date-of-immigration questions, the Decennial Census did. The PUMS figures in the table introduce some additional complications because the PUMS means without immigrants are not always equal to those of the CPS, in part because of sampling error in the CPS and in part because the 1989 CPS sampling frame is based on that of the 1980, not the 1990, Census. Nevertheless, in several instances the PUMS tabulations indicate that immigrant/nonimmigrant differences in characteristics are in the direction that would explain some of the CPS-PSID differences. Immigrants are disproportionately nonwhite, for example, possibly explaining the remaining gap between the CPS and PSID; and immigrants have lower labor force activity and earnings, consistent with the direction of the PSID-CPS gap (i.e., higher labor force activity and earnings levels in the PSID). Thus, while the evidence is not conclusive, it does suggest that immigration is part of the explanation for the remaining PSID-CPS difference for some variables.

CPS-PSID comparisons for other demographic groups—wives, female heads, male nonheads, and female nonheads (see Fitzgerald et al., 1997a) indicate that the results for wives are quite similar to those for male heads and, if anything, the CPS-PSID differences are even smaller. The results for female heads again show small CPS-PSID differences, with a few exceptions.

We conclude from this examination, therefore, that, despite the seemingly large differences in characteristics of attritors and nonattritors in the PSID, it nevertheless remains cross-sectionally representative of the nonimmigrant U.S. population.

CPS-PSID Regression Comparisons

Table 14 shows estimates of cross-sectional log earnings equations for male heads in the PSID and CPS in 1968, 1981, and 1989, using 1989 values for the independent variables as well as dependent variables. In general, the differences in parameter estimates are larger than might be expected on the

TABLE 14
PSID and CPS Log Earnings Regressions: Male Heads

	1968		1981		1989	
	PSID	CPS	PSID	CPS	PSID	CPS
Intercept	8.642*	8.456*	8.478*	7.545*	8.066*	7.560*
	(.015)	(.065)	(.086)	(.071)	(.067)	(.080)
Black	-.229*	-.393*	-.159*	-.283*	-.278*	-.241*
	(.032)	(.014)	(.043)	(.016)	(.048)	(.017)
Other race	-.102	-.264*	.144	-.210*	.046	-.210*
	(.099)	(.040)	(.111)	(.030)	(.125)	(.028)
Low education	-.288*	-.271*	-.244*	-.313*	-.140*	-.366*
	(.026)	(.010)	(.037)	(.012)	(.046)	(.015)
Some college	.028	.119	.016*	.101*	.167*	.119*
	(.027)	(.014)	(.033)	(.013)	(.039)	(.013)
College degree	.247*	.248*	.293*	.263*	.442*	.390*
	(.032)	(.013)	(.036)	(.012)	(.040)	(.012)
Age	.061*	.070*	.063*	.105*	.078*	.101*
	(.007)	(.003)	(.009)	(.003)	(.011)	(.004)
Age squared[a]	-.007*	-.008*	-.006*	-.011*	-.008*	-.011*
	(.001)	(.004)	(.001)	(.001)	(.001)	(.001)
Northeast	.054*	.035*	-.016	.060*	.080*	.148*
	(.029)	(.012)	(.035)	(.014)	(.014)	(.016)
North Central	.092*	.012	.070*	.046*	-.067	.057*
	(.028)	(.012)	(.033)	(.013)	(.041)	(.015)
South	.102*	-.177*	-.067*	-.039*	-.099*	-.013
	(.029)	(.012)	(.034)	(.013)	(.041)	(.014)

Notes: Standard errors in parentheses. Combined SRC-SEO sample (weighted) is used for PSID. Omitted categories for dummies are white, 12 years of education, and West.

*Significant at 5 percent level.

[a]Coefficients multiplied by 10.

basis of the unconditional means which, as we just demonstrated, are quite close to one another. The regression coefficients in the three years show generally similar signs, but a number of differences are sizable in magnitude. Two of these—for "Other race" and "Some college"—are probably due to differences in definitions of other race and of post-high-school education.[48] The same type of difference appears for earnings regressions of wives and female heads (see Fitzgerald et al., 1997a).

Table 15 shows F-statistics and chi-square statistics for the significance of the differences between PSID and CPS earnings regressions as well as probit equations for marital status and welfare participation in each year. For the log earnings regressions for male heads, the full set of coefficients, those excluding the constant, and those excluding the constant and the regional coefficients are significantly different in the two data sets in 1968. However, interestingly, the size and significance of the test statistics tend to fall over time. Indeed, by 1989, the coefficients other than the constant and region are insignificantly different in the two data sets. This finding suggests that attrition is not the cause of these differences in coefficient vectors. We speculate that the initial selectivity of who consented to be a part of the PSID (a 25 percent nonresponse rate) could have generated the 1968 differences we observe. That the dissimilarity then tends to fade out over the length of the PSID may be the result of the regression-to-mean phenomenon we demonstrated earlier for the unconditional means. This is an area for future research.[49]

---

[48]In the PSID, "Hispanic" was coded as a racial category prior to 1985 whereas in the CPS, "Hispanic" comes from a separate ethnicity question. For our regressions, we recoded "Hispanic" to "White" in the PSID in years prior to 1985. For the "Some college" variable, the treatment of junior colleges and vocational schools is different in the two data sets. On the other hand, these coefficients are also those for which differences appeared in Table 8.

[49]Becketti et al. (1988) found the same result: through 1981, the F-statistics for the difference in earnings regression coefficients (they did not examine other dependent variables) tended to fall over time. They speculated that the cause might be a result of their inclusion of nonsample individuals after 1968. However, we exclude nonsample individuals and find the same result, so we conclude that the pattern is a result of something else. We should also note that the patterns in Table 15 are unaltered by either the exclusion of the SEO sample or estimation without weights.

TABLE 15
Significance Tests for CPS-PSID Differences

	1968	1981	1989
Earnings: Male Heads			
All coeffs	11.3*	8.9*	5.6*
All coeffs but constant	3.7*	2.5	3.4*
All coeffs but constant & region	4.1*	3.0	4.0
Earnings: Female Heads			
All coeffs	2.8*	2.6*	3.9*
All coeffs but constant	1.2	1.3	1.6
All coeffs but constant & region	1.6	1.5	2.2
Earnings: Wives			
All coeffs	1.5	8.1*	4.8*
All coeffs but constant	1.5	0.9	2.4
All coeffs but constant & region	1.5	0.9	2.3
Marital Status: Males			
All coeffs	124.6*	96.4*	96.1*
All coeffs but constant	23.0*	23.5*	18.3*
All coeffs but constant & region	14.7*	22.0*	13.6
Marital Status: Females			
All coeffs	21.1*	16.2	27.1*
All coeffs but constant	20.5*	9.1	22.1*
All coeffs but constant & region	7.5*	8.7	13.5
Welfare Participation: Female Heads			
All coeffs	107.7*	25.8*	28.7*
All coeffs but constant	42.0*	23.9*	18.4
All coeffs but constant & region	33.2*	17.2*	14.2

Note: Earnings statistics are F-statistics; marital status and welfare participation are chi-square statistics.

*Significant at 5 percent level.

Table 15 shows somewhat similar patterns in the test statistics for other demographic groups and for other dependent variables, although the size of the statistics is sometimes smaller and sometimes larger. For the earnings equations for both wives and female heads, the coefficients in the two data sets are insignificantly different from one another when the constant is excluded (and when both the constant and the region coefficients are excluded) in all three years. For the other dependent variables, the test statistics are larger than for earnings but, like the male head earnings statistics, generally fall over time. In addition, in 1989 not a single test statistic for any group or any dependent variable is significant when coefficients other than the constant and region are compared.[50]

In any case, the major finding of our analysis is that, while the PSID-CPS differences in regression coefficients are larger than would be expected after our examination of the unconditional means, these differences go back to 1968. Further investigation, particularly of the causes of the initial difference in 1968, would be warranted in future research.

V. DYNAMIC ATTRITION MODELS

In the final section of our analysis, we explore the dynamic attrition issues discussed in Section II concerning the effects of permanent and transitory components of lagged "y" variables and make use, in general, of the full y-history by estimating year-by-year attrition hazards through 1989. This exercise has interest for two reasons. First, for the development of weights based on estimated attrition functions, these equations may be superior to those based only on the levels of the 1968 variables. However, given the results of our analysis thus far, attrition bias in the PSID does not appear to be very severe for cross-sectionally defined variables. The second reason is therefore more important, because these equations

[50]This general pattern of falling test statistics might be thought to be partly the result of declining sample sizes, but in fact the combined CPS-PSID sample size increases over time because the CPS has been gradually expanded, more than enough to outweigh PSID attrition.

have implications for attrition bias in equations used in past and future PSID studies which involve dynamic, or panel-defined, outcome variables rather than cross-sectionally defined ones (earnings and employment dynamics, welfare and marital status transition models, etc.). If "y" in our models in Section II is reinterpreted as such a dynamic outcome variable, then that approach implies that if lags of those variables are significant determinants of attrition then analyses which attempt to model the contemporaneous values of those variables on the nonattriting sample may produce inconsistent parameter estimates (namely, if the lagged values of those variables covary with the contemporaneous values). Because there is no counterpart to the CPS for panel-defined variables in the PSID, this can be our only (indirect) test of attrition bias for PSID dynamic analyses.

Although we have not developed a formal model of the causes of attrition, it is plausible to hypothesize that not only are individuals of low socioeconomic status likely to attrite (as our results on levels of the relevant variables have demonstrated thus far) but also that individuals with a recent change in earnings, marital status, and other variables are more likely to attrite. Taking this notion one step further, we hypothesize that individuals more likely to attrite are those observed over their full past history to have had above-average rates of fluctuations in earnings, above-average numbers of transitions in marital status, or above-average rates of geographic migration—to take the three which we will examine. We conjecture that it is plausible to suppose that disruption in general may be related to attrition because it may make individuals either more difficult to locate by the PSID field staff, less receptive to participation in the panel, or both.

To investigate this issue, we estimate attrition functions with a latent index of the form:

$$A_{it}^{*} = f(y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}) + x_{i0}\theta + \upsilon_{it} \qquad \textbf{(16)}$$

where the outcome variable, A_{it}, equals 1 if the individual attrites at time t, conditional on still being a respondent at t-1. The vector X_{i0} consists of time-invariant "x" variables, with coefficient vector θ. Equation (13) allows the lagged dependent variables to affect current attrition propensities in a general

way (function f) but, in our empirical work, we test functions which transform the lagged y into only four different summary variables: (a) the individual-specific mean of the variable over all years since 1968, (b) the individual-specific variance of the variable over all years since 1968, (c) deviations of lagged variables from the individual-specific means, and (d) durations of time spent in various states defined by the variables in question.

The first of these measures tests whether attrition is affected by individual-specific mean levels of earnings, marital status, and other variables (we include family structure and geographic mobility as well). This analysis should yield broadly similar findings to those in Section III because they only replace the 1968 values of these variables with their means over a period of years. The second of the statistics measures individual heterogeneity in turnover (labor market, marital, geographic location, etc.). As we noted previously, if attrition covaries with lagged values for these variables, then it follows that models estimated on nonattritors but using the contemporaneous counterparts to these measures as dependent variables (turnover, durations, transition rates, etc.) will be biased provided that the contemporaneous and lagged measures covary as well. The third of the measures tests whether lagged changes (“shocks”) to these variables affect attrition. This is logically separate from the question of individual heterogeneity in turnover. It relates closely to the issue of whether transitory events affect later attrition, although we cannot be sure of that interpretation because we cannot, by definition, determine whether recent events will persist in the future if the individual attrites (and hence whether the events will, in retrospect, be seen to be permanent or transitory shocks). This analysis has implications for bias in the estimation of transition rate models for contemporaneous variables on the nonattriting sample. The fourth measure is more familiar and tests whether durations in a state (marriage, migration) affect attrition propensities; these equations have implications for the estimation of contemporaneous models for the length of spells.

For our models we pool all observations on individuals aged 25–64 in original 1968 sample families for all years 1970–1989 for which they are observed.[51] We estimate logits for whether the individual attrites in the next period as a function of the four summary measures discussed above defined as of the current period. We also include 1968 variables for education, age, and other socioeconomic characteristics. In some runs we include year dummies, which fully capture duration dependence.

Table 16 shows a series of estimated attrition equations focusing on lagged earnings. Column (1) shows that attrition propensities for men are significantly negatively affected both by lagged mean earnings as well as earnings in the prior period. The latter implies that negative deviations of current earnings from mean earnings raise the likelihood of attrition. Column (2) shows that the effect of deviations does not extend back beyond the current period of the observation. Column (3) tests the effect of the individual-specific variance and finds that attrition rates are positively affected by variances, even conditioning on mean earnings in the current period and prior period. Column (4) shows that this result is robust to the inclusion of age and year dummies, because it could be the case that if attrition rates vary with calendar year or age, this might create spurious estimates since earnings vary with year and age.[52] However, column (5) shows that the inclusion of several standard socioeconomic variables (education, race, etc.) is sufficient to render insignificant the effect of lagged mean earnings on attrition rates, a result not surprising inasmuch as permanent earnings are likely to be more predictable by such regressors than are earnings deviations or earnings variances. The latter two remain significant even after inclusion of the additional regressors. The last column shows, in addition, that there are no significant effects of this kind for women. We speculate that earnings are not as good a predictor of instability of other

[51]We omit 1968 and 1969 so that we can construct at least two lagged variables for individuals last observed in 1970. We also make no adjustment to the standard errors for the pooled nature of the data (as we noted earlier, there are no adjustments for unobserved heterogeneity). However, year-by-year estimation of the models reveals qualitatively similar results; hence the standard error issue does not affect our conclusions.

[52]The year dummies show no significant duration dependence in the hazard after 1970.

TABLE 16
Dynamic Attrition Models with Focus on Lagged Earnings
(Logit Coefficients)

	Males					Females
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{y}$	-.20*	-.24*	-.28*	-.26*	-.07	.23
	(.07)	(.08)	(.08)	(.08)	(.09)	(.14)
y_{t-1}	-.22*	-.17*	-.18*	-.20*	-.15*	-.11
	(.06)	(.08)	(.06)	(.06)	(.07)	(.11)
y_{t-2}	—	-.09	—	—	—	—
		(.09)				
Var(y)	—	—	.32*	.33*	.38*	-.04
			(.09)	(.09)	(.09)	(.23)
Time dummies and age	n	n	n	y	y	y
Other characteristics[a]	n	n	n	n	y	y
R^2	.018	.017	.020	.025	.043	.018

Notes: Dependent variable is 1 if individual attrites in next period, 0 if not. $\bar{y}$ is the mean earnings from 1968 to current period; y_{t-1} and y_{t-2} are earnings in the current period and one period back; and var(y) is the variance of earnings from 1968 to the current period. The coefficients on the first three variables are multiplied by 10^4 and the coefficient on the fourth is multiplied by 10^8. Standard errors in parentheses. For R-squared definition, see Table 5 notes.

*Significant at 10 percent level.

[a]Education, race, region, age of youngest child, rural residence, homeowner.

behaviors for women as they are for men because there are considerably more planned fluctuations in earnings for women.[53]

These results, therefore, are consistent, at least for men, with attrition being selective on stability. Therefore it should be expected that measures of second moments, of turnover and hazard rates, and of related variables should be smaller in the nonattriting PSID sample than in the population as a whole.

Tables 17 and 18 show that this result extends to marital, family structure, and migration behavior. Table 17 demonstrates that men recently experiencing a transition out of marriage (due to divorce, separation, or widowhood) are more likely to attrite than those not experiencing such a transition. In addition, men who have experienced larger numbers of marital transitions in the past are more likely to attrite. Interestingly, however, no effects of this kind appear for females. Table 18 shows that men who have split off from other families are more likely to attrite—although the effects are insignificant when other characteristics are controlled—and that men who have moved recently or who show a high average propensity to move are more likely to attrite. Again, however, no significant effects appear for women.

Although these results clearly demonstrate a tendency for men with more unstable histories to attrite, the seriousness of the problem for the PSID is difficult to judge. The R-squared values in these attrition equations are uniformly very small, as shown in the tables, which implies that attrition along these dimensions may not have a large effect on the comparable contemporaneous measures on the nonattriting sample from selection on these observables. This cannot be known for certain because the size of the bias depends not only on the R-squared values but also on the size of the relation of these lagged instability measures both with the regressors in the main outcome equation of interest and with the error term in that equation (recall the model of Section II). However, weights based on these

[53]We thank a referee for suggesting as well that the female results may reflect the existence of married-couple households in which the husband's earnings are the dominant factor affecting the family's attrition.

TABLE 17
Dynamic Attrition Models with Focus on Lagged Marital Status
(Logit Coefficients)

	Males				Females
	(1)	(2)	(3)	(4)	(5)
$\bar{y}$	-.24	-.22	-.31	-.21	-.14
	(.19)	(.20)	(.19)	(.20)	(.19)
y_{t-1}	-.81*	-.72*	-.67*	-.72*	-.15
	(.15)	(.15)	(.15)	(.16)	(.17)
n_{tr}	—	.20*	—	.21*	.02
		(.05)		(.09)	(.09)
Duration	—	—	-.04*	.00	-.01
			(.01)	(.02)	(.02)
Other characteristics[a]	n	n	n	y	y
R^2	.022	.024	.023	.043	.009

Notes: Dependent variable is the same as in Table 16. $\bar{y}$ is the average probability of being married from 1968 to the current period; y_{t-1} is a married dummy for the current period; n_{tr} is the number of marital transitions from 1968 to the current period; and "Duration" is the number of years since the last marital transition. All equations contain age and year dummies. Standard errors in parentheses. For R-squared definition, see Table 5 notes.

*Significant at 10 percent level.

[a]See Table 16.

TABLE 18
Dynamic Attrition Models with Focus on Splitoff and Migration
(Logit Coefficients)

	Splitoff				Migration		
	Male			Female	Male		Female
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Splitoff							
Split in t-1	.73*	.74*	.35	-.05	—	—	—
	(.37)	(.37)	(.37)	(.59)			
Ever Split Off	—	.28*	-.04	.00	—	—	—
		(.15)	(.16)	(.18)			
Migration							
$\bar{y}$	—	—	—	—	.90*	.77*	-.02
					(.29)	(.30)	(.36)
y_{t-1}	—	—	—	—	.41*	.28*	.13
					(.12)	(.12)	(.13)
Duration	—	—	—	—	-.02*	-.01	-.00
					(.01)	(.02)	(.01)
Other characteristics[a]	n	n	y	y	n	y	y
R^2	.006	.007	.036	.017	.015	.040	.017

Notes: Dependent variable is the same as in Table 16. $\bar{y}$ is the average number of moves from 1968 to the current period; y_{t-1} is a dummy for having moved in the current period; and "Duration" is the number of years since the last move. All equations include age and year dummies. Standard errors in parentheses. For R-squared definition, see Table 5 notes.

*Significant at 10 percent level.

[a]See Table 16.

equations could be developed which would capture dynamic effects more adequately than do the current, universal PSID weights, and these could be used in specification tests to see the importance of their effect on estimates of outcome equations. Nevertheless, this approach cannot capture any bias from selection on unobservables in such equations (unfortunately, as previously noted, there is no equivalent to the CPS for these variables with which to gauge the presence of such selection).

VI. CONCLUSIONS

Our study of attrition in the PSID has yielded several findings:

- The observed baseline characteristics of those who later do and do not attrite from the PSID are quite different; these differences are often statistically significant. Attritors tend to have lower earnings, lower education levels, lower marriage propensities, and appear generally to be drawn from the lower tail of the socioeconomic distribution.

- These unadjusted differences fall in magnitude and are usually rendered statistically insignificant as determinants of attrition propensities after conditioning on a number of other socioeconomic characteristics. In one leading case, however—earnings for male heads—a significant relationship continues to exist even after such conditioning.

- In a regression context, attrition appears primarily to affect intercepts rather than slopes of regressions for earnings and welfare participation, but also some slopes for marital-status regressions.

- Cross-sectional comparisons of unconditional moments between the PSID and the CPS show a close correspondence all the way through 1989. We reconcile the seemingly inconsistent findings of, on the one hand, significant measured correlates of attrition and, on the other hand, continued cross-sectional representativeness by showing that regression-to-the-mean effects are present that cause initial differences in characteristics to fade away over time both within and across generations. A small role is also played by PSID weights used to adjust for attrition related to observables, although, because attrition is mostly noise, the weights do not alter PSID means by a very large amount. We also find that some portion of the remaining CPS-PSID difference is a result of the exclusion from the PSID sampling frame of individuals who have immigrated to the U.S. since 1968.

- Regression coefficients in models for earnings, marital status, and welfare participation in the CPS and the PSID are usually quite similar in sign and magnitude but not always so, and the differences in coefficient vectors as a whole are usually significant in the baseline year (1968). However, the test statistics for the difference in coefficient vectors fall over time and imply that, by 1989, the CPS and PSID coefficients are insignificantly different as a whole.

- We find evidence that attrition propensities are correlated with individual-specific levels of turnover and instability in earnings, in marital status, and in geographic mobility. We also find that recent unfavorable events along these dimensions—a drop in earnings, a marital dissolution, or a geographic move—induce more attrition. The magnitudes of the effects of these variables on attrition, as measured by R-squareds, are not large, which suggests that they are unlikely to induce significant bias in studies which have such dynamic measures as outcome variables. As noted earlier, however, this conclusion depends on model-specific correlations, and we recommend that authors of these types of studies be aware of possible attrition biases and check the sensitivity of their results accordingly.

APPENDIX

Let f(y, z|x) be the complete-population joint density of y and z and let g(y, z|x, A = 0) be the conditional joint density. Then

$$g(y,z|x,A=0) = \frac{g(y,z,A=0|x)}{Pr(A=0|x)}$$

$$= \frac{Pr(A=0|y,z,x)\, f(y,z|x)}{Pr(A=0|x)}$$

$$= \frac{Pr(A=0|z,x)\, f(y,z|x)}{Pr(A=0|x)}$$

$$= \frac{f(y,z|x)}{w(z,x)}$$

where w(z, x) is given in equation (9) in the text. Hence

$$f(y,z|x) = w(z,x)\; g(y,z|x,A=0).$$

Integrating both sides over z gives equation (8) in the text.

References

Abowd, J., and D. Card. 1989. “On the Covariance Structure of Earnings and Hours Changes.” *Econometrica* 57 (March): 411–445.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge: Harvard University Press.

Becketti, S., W. Gould, L. Lillard, and F. Welch. 1988. “The Panel Study of Income Dynamics after Fourteen Years: An Evaluation.” *Journal of Labor Economics* 6 (October): 472–492.

Bound, J., C. Brown, G. Duncan, and W. Rogers. 1994. “Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data.” *Journal of Labor Economics* 12 (July): 345–368.

Cameron, A. C., and F. A. G. Windmeijer. 1997. “An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models.” *Journal of Econometrics* 77 (April): 329–342.

Cosslett, S. 1993. “Estimation from Endogenously Stratified Samples.” In *Handbook of Statistics*, vol. 11, *Econometrics,* eds. G. S. Maddala, C. R. Rao, and H. D. Vinod. New York: North-Holland.

DuMouchel, W., and G. Duncan. 1983. “Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples.” *Journal of the American Statistical Association* 78 (September): 535–543.

Duncan, G., and D. Hill. 1989. “Assessing the Quality of Household Panel Data: The Case of the Panel Study of Income Dynamics.” *Journal of Business and Economic Statistics* 7 (October): 441–452.

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1997a. A Study of Sample Attrition in the Michigan Panel Study of Income Dynamics. Mimeographed, Johns Hopkins University.

Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1997b. The Impact of Attrition on the Estimation of Intergenerational Relationships in the PSID. Mimeographed, Johns Hopkins University.

Gottschalk, P., and R. Moffitt. 1992. “Earnings and Wage Distributions in the NLS, CPS, and PSID.” Part I of Final Report to the U.S. Department of Labor. Providence: Brown University.

Hausman, J., and D. Wise. 1979. “Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment.” *Econometrica* 47 (March): 455–473.

Hausman, J., and D. Wise. 1981. “Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment.” In *Structural Analysis of Discrete Data with Econometric Applications,* eds. C. Manski and D. McFadden. Cambridge: MIT Press.

Heckman, J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47 (January): 153–161.

Heckman, J. 1987. “Selection Bias and Self-Selection.” In *The New Palgrave: A Dictionary of Economics*, eds. J. Eatwell, M. Milgate, and P. Newman, vol. IV. London: Macmillan.

Heckman, J., and V. J. Hotz. 1989. “Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs.” *Journal of the American Statistical Association* 84 (December): 862–874.

Heckman, J., and R. Robb. 1985. “Alternative Methods for Evaluating the Effects of Interventions.” In *Longitudinal Analysis of Labor Market Data*, eds. J. Heckman and B. Singer. New York: Cambridge University Press.

Hill, M. 1992. *The Panel Study of Income Dynamics: A User’s Guide.* Newbury Park, CA: Sage Publications.

Hirano, K., G. Imbens, G. Ridder, and D. Rubin. 1996. “Combining Panel Data Sets with Attrition and Refreshment Samples.” Mimeographed, Harvard University, September.

Horowitz, J., and C. Manski. Forthcoming. “Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations.” *Journal of Econometrics*.

Imbens, G., and J. Hellerstein. 1996. “Imposing Moment Restrictions from Auxiliary Data by Weighting.” Mimeographed, Harvard University, July.

Imbens, G., and T. Lancaster. 1994. “Combining Micro and Macro Data in Microeconometric Models.” *Review of Economic Studies* 61 (October): 655–680.

Imbens, G., and T. Lancaster. 1996. “Efficient Estimation and Stratified Sampling.” *Journal of Econometrics* 74 (October): 289–318.

Institute for Social Research. 1972. *A Panel Study of Income Dynamics: Study Design, Procedures, and Available Data, 1968–1972 Interviewing Years*, vol. 1. Ann Arbor, MI.

Institute for Social Research. 1992. *A Panel Study of Income Dynamics: Procedures and Tape Codes, 1989 Interviewing Year,* vol. 1, *Procedures and Tape Codes*. Ann Arbor, MI.

Little, R., and D. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

MaCurdy, T. 1982. “The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis.” *Journal of Econometrics* 18 (January): 83–114.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Madow, W., H. Nisselson, I. Olkin, and D. Rubin, eds. 1983. *Incomplete Data in Sample Surveys,* 3 volumes. New York: Academic Press.

Manski, C. 1994. “The Selection Problem.” In *Advances in Econometrics: Sixth World Congress,* vol. 1, ed. C. Sims. New York: Cambridge University Press.

Manski, C., and S. Lerman. 1977. “The Estimation of Choice Probabilities from Choice Based Samples.” *Econometrica* 45 (November): 1977–1988.

Moffitt, R., and P. Gottschalk. 1995. “Trends in the Autocovariance Structure of Earnings in the U.S.: 1969–1987.” Working Paper 355, Johns Hopkins University.

Nijman, T., and M. Verbeek. 1992. “Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function.” *Journal of Applied Econometrics* 7 (July-September): 243–257.

Powell, J. 1994. “Estimation of Semi-Parametric Models.” In *Handbook of Econometrics,* vol. IV, eds. R. Engle and D. McFadden. New York: North-Holland.

Rao, C. R. 1965. “On Discrete Distributions Arising Out of Methods of Ascertainment.” In *Classical and Contagious Discrete Distributions*, ed. G. P. Patil. Oxford, NY: Pergamon Press.

Rao, C. R. 1985. “Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent?” In *A Celebration of Statistics*, eds. A. Atkinson and S. Fienberg. New York: Springer-Verlag.

Ridder, G. 1990. “Attrition in Multi-Wave Panel Data.” In *Panel Data and Labor Market Studies*, eds. J. Hartog, G. Ridder, and J. Theeuwes. New York: North-Holland.

Ridder, G. 1992. “An Empirical Evaluation of Some Models for Non-Random Attrition in Panel Data.” Mimeographed, University of Groningen.

Van den Berg, G., M. Lindeboom, and G. Ridder. 1994. “Attrition in Longitudinal Panel Data and the Empirical Analysis of Dynamic Labour Market Behavior.” *Journal of Applied Econometrics* 9 (October-December): 421–435.

Verbeek, M., and T. Nijman. 1996. “Incomplete Panels and Selection Bias.” In *The Econometrics of Panel Data*, eds. L. Matyas and P. Sevestre. Boston: Kluwer.