THE PANEL STUDY OF INCOME DYNAMICS AFTER FOURTEEN YEARS: AN EVALUATION

by

Sean Becketti

William Gould

Lee Lillard

Finis Welch

UCLA Department of Economics
Working Paper #361
January 1985

The Panel Study of Income Dynamics After Fourteen Years: An Evaluation

ABSTRACT

The Panel Study of Income Dynamics (PSID) is a self-replacing longitudinal data set of families who are interviewed annually by the Institute for Social Research (ISR). These annual interviews began in 1968 and are still continuing. In this paper, we consider whether attrition over the first fourteen years of the PSID has reduced the representativeness of the sample. We find no evidence in favor of this hypothesis. We do find some anomalies in the observation weights supplies by ISR. In addition, we find that a substantial quantity of data is withheld from the public distribution tapes. These are the data for the individuals who do not respont to the most current wave.

The Panel Study of Income Dynamics After Fourteen Years: An Evaluation

I. Introduction

The Panel Study of Income Dynamics (PSID) is a self-replacing longitudinal data set of approximately 5,000 families who are interviewed annually by the Institute for Social Research (ISR). These annual interviews began in 1968 and are still continuing. The socioeconomic characteristics of each of the families and of the roughly 20,000 individuals in these families are recorded in minute detail. Because of its vast scope and because it is a panel, the PSID has become an important data source for scholars in all of the social sciences.

In this paper, we consider whether attrition over the first fourteen years of the PSID has reduced the representativeness of the sample. Entry into and exit from the PSID might be correlated with individual characteristics in a way that biases estimates of behavioral relationships. If this is the case, then investigators may wish to use only the first few waves of the sample in their empirical work. In addition, published results that exploit the longitudinal aspects of the PSID or that are based on data from the later waves may need to be re-examined. Alternatively, if sample entry and exit are completely random, more confidence can be placed in results based on evidence from the PSID, and more ambitious studies of dynamic behavior may be advisable.

There is no well developed statistical theory for testing the null hypothesis that sample dynamics in the PSID are random. One reason for this lacuna is that data sets like the PSID are relatively new phenomena. As a result, this kind of hypothesis has rarely arisen before. Another reason is that the alternative hypothesis, that entry into and exit from the sample are

correlated with some (perhaps unobserved) variable of interest, is too broad. It is impossible to examine every possible relationship between the sample dynamics and other variables.

The impossibility of exhaustively and conclusively testing our null hypothesis leads us to adopt an eclectic approach in our investigation. In the next section, we describe the structure of the PSID in greater detail, and we document some of the important features of the observed sample dynamics. In the section following, we take a closer look at those individuals who leave the PSID before the fourteenth wave and we compare them to those individuals who remain in the sample. In the fourth section of the paper, we compare the original 1968 PSID sample to the sample collected by the March 1968 Current Population Survey (CPS). If the CPS is representative of the U.S. population, then this comparison can tell us whether or not the original PSID families are a representative sample. In section five, we consider whether PSID sample dynamics affect the estimates of some particular behavioral relationships. Finally, we summarize the evidence and conclude that attrition has not reduced the representativeness of the PSID.

II. The Dynamics of Sample Size in the PSID

A. The Structure of the PSID[1]

In 1968, the Survey Research Center (SRC) of the Institute for Social Research (ISR) at the University of Michigan interviewed 4,802 families. Of this group, 2,930 families were selected from SRC's master sampling frame.[2] These families (and/or the members of these families) are called the SRC sample. The remaining 1,872 families were drawn from the Bureau of the Census's Survey of Economic Opportunity (SEO). The members of this latter group are called the SEO sample.

The 1968 SRC sample is a probability sample rather than a random sample of U.S. families, that is, the probability that a family with given characteristics is in the sample is known.[3] As it turns out, these probabilities do not vary greatly across SRC families. Thus very little bias is introduced by treating the SRC sample as a random sample of either families or individuals.

A major goal of the PSID is to study the determinants of poverty. In a random sample of 5,000 families, too few poverty and minority families would be drawn. To overcome this problem, the SEO sample was added to the PSID. The SEO sample is a subset of the approximately 30,000 families interviewed in 1966 and 1967 for the Survey of Economic Opportunity. This subset was selected according to criteria specified by ISR.[4] One important criterion for inclusion in the PSID was that the household had family income in 1966 less than or equal to twice the 1966 poverty line. In addition, these families had to agree to release their names and addresses to ISR.

In 1968 ISR calculated weights for each family that are supposed to represent the ex ante probability that a family appears in the PSID. These weights were updated in 1972, 1978, 1979, 1980, and 1981 to take account of differential non-response rates in the succeeding waves of the PSID. From 1978 on, weights have been calculated annually for each individual in the sample.

A substantial number of individuals in each wave after 1968 are assigned a weight of zero. These are the so-called non-sample persons, that is, persons who entered the sample after 1968 through marriage or living arrangements with a sample person. ISR assigns a weight of zero to these individuals to indicate that they are unable to calculate the probability that any particular non-sample person enters the PSID. Obvious additional complications arise when a child is born to parents one of whom is a non-sample person. ISR

assigns the child an individual weight that is the average of the parents' weights, that is, the child receives one-half the sample parent's weight.

The most important feature of a non-sample person is that ISR makes no attempt to continue interviewing them if they stop residing with a sample person. A sample person, on the other hand, is pursued even if they leave their original sample family. Such sample persons are called splitoffs by ISR.

ISR distributes the PSID data in four ways: single year family tapes, single year family-individual tapes, merged family tapes, and merged family-individual tapes. Single year tapes contain interviews from a single year (or "wave") of the PSID; merged tapes contain data from every year of the survey. Family tapes contain one record for each family in the panel; family-individual tapes contain one record for each person in the panel. In each family, one person is designated as the head. Thus a family tape can be derived from the corresponding family-individual tape by deleting the records of all non-heads.

All of these public distribution tapes have a rectangular format, that is, if a family or individual does not participate in every year of the survey, their answers are coded as missing values for every variable in all waves in which the family or individual does not participate. ISR does not make any attempt to reduce the size of the data sets by compressing out redundant missing values. However, and this is crucial, no record appears for any family or individual that does not respond to the most current wave. Thus, in the 1981 wave merged family-individual tape released by ISR, there are records for only 60 percent of the 30,957 persons ever interviewed by ISR.[5]

B. Attrition from the PSID

Individuals enter and exit the PSID in a variety of ways. Aside from those individuals initially selected and interviewed, some individuals enter the data set by being born to sample members, while others attach themselves to sample members through other means, typically through marriage. Some individuals leave the data set because they have died, some refuse to answer additional questionnaries, some cannot be found, and a large number (the non-sample persons who leave sample households) are simply not pursued by ISR.[6]

Tables 2.1 and 2.2 display the number of years that an individual remains in the PSID by the wave in which they enter.[7] Table 2.1 presents this duration data separately for the SRC and SEO subsamples, while Table 2.2 presents the data separately for sample and non-sample persons. In both tables, each entry gives the percent of the indicated group that attrites after the number of years listed at the top of that column. For example, in Table 2.2 we see that 3.61 percent of the sample persons who first appear in wave 5 remain in the sample for exactly four interviews. By adding up entries along that row, we see that 20.99 percent (9.93+4.74+2.71+3.61) of the sample persons who first appear in wave 5 respond to four or less interviews. The last entry in each row gives the percentage of that row that is still in the PSID. Thus, we see that 70.65 percent of the sample persons who first appear in wave 5 are still responding to the interviews. The entries in the bottom row give the number of persons in each column. Thus 836 sample persons respond to exactly four interviews.

From these two tables, we see that those individuals who do attrite are most likely to do so after their first interview. Slightly more than 12 percent of the individuals interviewed in the initial (1968) wave of the PSID never appear in later waves.[8] Furthermore, the probability that a person ever

TABLE 2.1

Duration in the PSID by Entry Wave for SRC and SEO Subsamples

Wave	Sample	Duration to Exit 1	2	3	4	5
1	SRC	14.51	2.68	2.21	1.81	2.21
	SEO	10.10	3.36	1.92	2.59	2.47
2	SRC	18.30	11.60	6.96	2.84	2.84
	SEO	24.90	14.20	10.89	3.50	4.28
3	SRC	12.64	7.67	4.29	4.29	4.51
	SEO	19.38	10.31	5.57	6.19	4.95
4	SRC	13.63	6.33	2.68	4.38	3.16
	SEO	23.87	6.85	4.50	6.26	6.46
5	SRC	11.02	6.21	5.01	3.41	2.20
	SEO	20.00	8.25	5.79	4.74	4.21
6	SRC	9.62	7.93	3.85	3.85	4.09
	SEO	16.91	7.93	5.22	6.47	3.55
7	SRC	14.83	5.98	3.35	3.35	3.59
	SEO	13.05	7.08	8.85	3.54	4.20
8	SRC	14.91	5.73	5.28	1.83	3.67
	SEO	13.33	8.00	3.78	4.89	3.56
9	SRC	8.39	6.45	7.10	4.73	4.30
	SEO	13.01	7.89	6.82	6.18	5.76
10	SRC	12.33	7.49	5.95	3.52	70.70
	SEO	14.60	9.97	7.56	5.67	62.20
11	SRC	12.34	6.55	4.53	76.57	
	SEO	16.02	9.45	7.60	66.94	
12	SRC	11.90	9.48	78.63		
	SEO	13.84	13.84	72.33		
13	SRC	11.52	88.48			
	SEO	14.55	85.45			
14	SRC	100.00				
	SEO	100.00				
Total	SRC	2489	1072	813	617	654
Total	SEO	2455	1393	917	794	763

(continued)

Table 2.1 (continued)

Wave	Sample	Duration to Exit 6	7	8	9	10	11	12	13	14
1	SRC	2.09	1.82	2.09	1.63	1.88	1.64	1.60	2.27	61.58
	SEO	2.07	2.26	3.00	2.55	1.57	2.38	2.66	2.72	60.36
2	SRC	3.09	4.64	3.61	1.55	2.84	0.52	2.06	39.18	
	SEO	3.70	2.72	1.36	1.95	1.36	1.56	1.56	28.02	
3	SRC	2.71	4.06	1.81	1.58	2.93	1.81	51.69		
	SEO	3.09	2.27	0.82	2.47	2.27	1.65	41.03		
4	SRC	2.68	2.43	2.68	1.22	2.43	58.39			
	SEO	2.15	1.17	3.52	2.35	1.76	41.10			
5	SRC	5.21	3.81	1.80	3.61	57.72				
	SEO	2.11	3.68	2.28	2.63	46.32				
6	SRC	3.85	4.09	1.92	60.82					
	SEO	5.64	5.64	2.71	45.93					
7	SRC	1.20	4.78	62.92						
	SEO	4.65	4.87	53.76						
8	SRC	3.21	65.37							
	SEO	6.44	60.00							
9	SRC	69.03								
	SEO	60.34								
Total	SRC	616	560	512	444	501	406	389	368	5861
Total	SEO	601	571	564	495	430	437	443	385	5353

Note: This table shows the attrition rates, expressed as a percentage, for the SEO and SRC subsamples. For example the number in the first row and first column on this page indicates that 2.09 percent of SRC subsample persons who entered in wave 1 were last seen in wave 6. Skipping to the end of the row, we see that 61.58 percent of the original wave 1 SRC subsample was still in the panel as of the fourteenth wave. At the bottom of the table we present the overall sample sizes. Thus, 616 persons from the SRC sample were last seen in wave 6. Attrition is broadly defined in the table and includes exiting the sample for all reasons including death.

TABLE 2.2

Duration in the PSID by Entry Wave and Sample/Non-Sample Status

Wave	Status	Duration 1	2	3	4	5
1	Sample	12.38	3.01	2.07	2.19	2.33
	Non-Sample	.	.	.	.	.
2	Sample	1.98	0.40	2.37	2.77	3.95
	Non-Sample	29.89	18.03	11.86	3.39	3.54
3	Sample	1.59	2.38	4.23	4.76	2.38
	Non-Sample	26.18	13.64	5.45	5.64	6.36
4	Sample	7.24	3.49	2.68	2.14	3.49
	Non-Sample	27.50	8.74	4.37	7.65	6.01
5	Sample	9.93	4.74	2.71	3.61	2.03
	Non-Sample	19.97	9.11	7.35	4.47	4.15
6	Sample	0.00	0.00	2.96	4.73	1.48
	Non-Sample	21.72	12.75	5.57	5.57	5.21
7	Sample	0.00	2.55	4.25	1.70	3.68
	Non-Sample	23.40	9.28	7.54	4.64	4.06
8	Sample	3.77	9.55	8.79	4.27	5.78
	Non-Sample	22.54	4.71	1.02	2.66	1.84
9	Sample	2.76	1.76	4.52	4.27	3.77
	Non-Sample	16.60	11.19	8.77	6.34	5.97
10	Sample	2.88	3.36	2.88	2.88	88.01
	Non-Sample	20.84	12.60	9.53	5.98	51.05
11	Sample	4.66	3.56	3.84	87.95	
	Non-Sample	21.19	11.37	7.90	59.54	
12	Sample	4.83	4.62	90.55		
	Non-Sample	19.46	17.60	62.94		
13	Sample	3.26	96.74			
	Non-Sample	21.06	78.94			
14	Sample	100.00				
	Non-Sample	100.00				
Total Sample		2993	1189	929	836	879
Total Non-Sample		1951	1276	801	575	538

(continued)

Table 2.2 (continued)

Wave	States	Duration 6	7	8	9	10	11	12	13	14
1	Sample	2.08	2.03	2.53	2.07	1.73	2.00	2.11	2.49	60.99
	Non-Sample	.	.	.	.	.	.	.	.	.
2	Sample	2.37	3.95	1.19	1.19	2.37	1.58	2.77	73.12	
	Non-Sample	3.85	3.39	2.77	2.00	1.85	0.92	1.39	17.10	
3	Sample	2.12	2.38	1.59	2.38	2.38	1.59	72.22		
	Non-Sample	3.45	3.64	1.09	1.82	2.73	1.82	28.18		
4	Sample	1.88	1.34	1.61	2.68	1.34	72.12			
	Non-Sample	2.73	2.00	4.19	1.28	2.55	32.97			
5	Sample	0.90	2.71	0.90	1.81	70.65				
	Non-Sample	5.43	4.47	2.88	3.99	38.18				
6	Sample	3.85	4.44	1.78	80.77					
	Non-Sample	5.39	5.21	2.69	35.91					
7	Sample	2.55	3.68	81.59						
	Non-Sample	3.29	5.61	42.17						
8	Sample	2.46	64.75							
	Non-Sample	7.79	60.05							
9	Sample	82.91								
	Non-Sample	51.12								
Total Sample		772	753	778	684	651	646	668	642	11214
Total Non-Sample		445	378	298	255	280	197	164	111	.

Note: The above table shows the attrition rates, expressed as a percentage, by sample/non-sample status. For example, the number in the first row and first column on this page indicates that 2.08 percent of sample persons who entered in wave 1 were last seen in wave 6. Skipping to the end of the row, we see that 60.99 percent of the original wave 1 sample were still in the panel as of wave 14. At the bottom of the table we present the overall sample sizes. Thus, 772 sample persons were last seen in wave 6.

attrites after answering more than one questionnaire is reasonably small. For example, the probability that an individual who participated in the 1968 wave ever attrites, given that this individual does not attrite after the first interview, is only 30 percent.

Members of the SEO subsample are generally more likely to attrite than members of the SRC subsample. Non-sample persons are much more likely to attrite than sample members. However, we suspect that this latter differential is mostly due to the ISR policy of not following non-sample persons if they leave their sample household.

Table 2.3 displays the distribution of lengths of participation in the PSID.[9] This distribution is reported separately for the sample and non-sample persons and, within these two groups, for those individuals who have attrited and those who are still participating. The participation lengths for individuals who are still participating in the PSID are censored observations; the final lengths of participation for these persons will be at least as long as the observed length in 1981.

In Table 2.3 we see that the length of participation appears to be roughly uniformly distributed for those sample individuals who do not attrite after their first interview. This pattern of attrition is consistent with the hypothesis that attrition takes place randomly - at least after the most mobile and least responsive individuals are screened out by one interview. For non-sample persons, the distribution of participation lengths appears to be a smoothly declining function of participation length.

There are several factors which may explain this difference in the shapes of the distributions for sample and non-sample persons. Sample individuals who enter after the 1968 wave are typically born into the sample. Non-sample persons typically enter the sample by marrying or entering into a living

TABLE 2.3

Length of Participation in the PSID

(in years)

Frequency Percent Row Percent Column Percent	Sample		Non-Sample		
	Completed	Open	Completed	Open	Total
1	2567	448	1481	505	5001
	8.31	1.45	4.79	1.63	16.18
	51.33	8.96	29.61	10.10	
	29.32	3.01	40.35	14.03	
2	721	507	800	530	2558
	2.33	1.64	2.59	1.72	8.28
	28.19	19.82	31.27	20.72	
	8.24	3.41	21.80	14.73	
3	542	440	433	368	1783
	1.75	1.42	1.40	1.19	5.77
	30.40	24.68	24.28	20.64	
	6.19	2.96	11.80	10.23	
4	557	327	258	305	1447
	1.80	1.06	0.83	0.99	4.68
	38.49	22.60	17.83	21.08	
	6.36	2.20	7.03	8.47	
5	554	382	212	314	1462
	1.79	1.24	0.69	1.02	4.73
	37.89	26.13	14.50	21.48	
	6.33	2.57	5.78	8.72	
6	497	345	170	274	1286
	1.61	1.12	0.55	0.89	4.16
	38.65	26.83	13.22	21.31	
	5.68	2.32	4.63	7.61	
7	509	391	127	240	1267
	1.65	1.27	0.41	0.78	4.10
	40.17	30.86	10.02	18.94	
	5.81	2.63	3.46	6.67	
8	579	373	73	219	1244
	1.87	1.21	0.24	0.71	4.03
	46.54	29.98	5.87	17.60	
	6.61	2.51	1.99	6.09	

(continued)

Table 2.3 continued

Frequency Percent Row Percent Column Percent	Sample		Non-Sample		
	Completed	Open	Completed	Open	Total
9	513	339	55	197	1104
	1.66	1.10	0.18	0.64	3.57
	46.47	30.71	4.98	17.84	
	5.86	2.28	1.50	5.47	
10	417	405	37	230	1089
	1.35	1.31	0.12	0.74	3.52
	38.29	37.19	3.40	21.12	
	4.76	2.72	1.01	6.39	
11	451	360	15	165	991
	1.46	1.16	0.05	0.53	3.21
	45.51	36.33	1.51	16.65	
	5.15	2.42	0.41	4.58	
12	448	339	9	148	944
	1.45	1.10	0.03	0.48	3.05
	47.46	35.91	0.95	15.68	
	5.12	2.28	0.25	4.11	
13	400	183	0	104	687
	1.29	0.59	0.00	0.34	2.22
	58.22	26.64	0.00	15.14	
	4.57	1.23	0.00	2.89	
14	0	10040	0	0	10040
	0.00	32.49	0.00	0.00	32.49
	0.00	100.00	0.00	0.00	
	0.00	67.48	0.00	0.00	
Total	8755	14879	3670	3599	30903
	28.33	48.15	11.88	11.65	100.00

Note: This table reports the number of sample and non-sample persons with each possible length of continuous participation in the PSID. We distinguish between "completed" and "open" intervals of participation. If an individual has attrited, they are counted in the column labeled "completed"; otherwise they are counted in the column labeled "open". We do not know the final length of participation in the PSID for individuals who have not attrited.

arrangement with a sample person. They dynamics of these two entry mechanisms may be quite different. Many sample persons attrite because of death. On the other hand, ISR's policy of not tracking non-sample persons when they leave their sample household guarantees that the dynamics of exit are very different for the sample and non-sample groups.

Aside from the differences in the shapes of the participation length distributions, the mean participation lengths also differ across groups. Note, for example, that slightly more than half of the non-sample persons have attrited compared to only thirty-seven percent of the sample individuals. Again, we suspect that most, if not all, of this differential is due to ISR's policy of not tracking non-sample persons after they leave the sample household.

We will present evidence below that suggests that non-sample persons do not behave systematically differently from sample persons. If this hypothesis is correct, then ISR may wish to reconsider its policy towards following non-sample persons. Retaining these individuals may be a cost-effective way of maintaining adequately sized samples of individuals who participate long enough to permit dynamic analyses.

Notice that 40.2 percent of all participants in the PSID have attrited. The tapes distributed by ISR contain information only for those individuals participating in the most recent wave of the PSID. Thus, 40.2 percent of all PSID participants, 12,425 persons, do not appear at all on the fourteenth wave tape. Because these individuals have attrited, they represent somewhat less than forty percent of the person-years of PSID data. Nonetheless, an enormous amount of data is inadvertently suppressed by ISR. If attrition is unsystematic, then these data are clearly useful, especially as they contain many long stretches of continuous participation in the PSID. If, on the other hand,

attrition is systematically related to characteristics that influence behavior and economic outcomes, then these data contain precisely the information that researchers need to correct for the biases introduced by attrition. In either case, there seems to be no justification for deleting these records from the public distribution tapes.

III. Characteristics of Those Leaving the PSID

We can offer no structural model for attrition. We seek instead in this section to describe the correlations between attrition and the observable characteristics of the sample individuals. We do this in three ways. First, we use a linear probability model to locate variables that help to predict whether the current wave is the last wave in which an individual responds. This method allows us to relate attrition to the most current information about each individual. Our second approach consists of estimating another linear probability model where we try to predict whether a member of the original 1968 sample still participates by 1969, by 1975, or by 1981. In these regressions, the covariates are all baseline variables, that is, measured as of 1968. Finally, we fit a censored, ordered probit model of the duration in the PSID of male heads of households.

A. The Probability of Exiting Between Adjacent Waves

Table 3.1 displays the estimates of two linear probability regressions — one for sample persons an the other for non-sample persons. In both regressions, the dependent variable is equal to one if this observation comes from the final wave in which this individual appears and is equal to zero otherwise. There can be as many as thirteen observations on each individual. For example, there are four observations on an individual who responds to the 1968 through 1971 interviews. The dependent variable is equal to zero in the first three observations (1968-1970) on such an individual. It is equal to one in

TABLE 3.1

The Probability of No Further Interviews

Variable	Sample Persons	Non-Sample Persons
Intercept	0.049	0.244
	(0.0023)	(0.0197)
SEO Sample	0.003	0.027
	(0.0009)	(0.0045)
Splitoff	-0.015	-0.077
	(0.0031)	(0.0086)
First Interview	0.078	0.067
	(0.0014)	(0.0061)
Mobility		
Might move	0.0005	0.019
	(0.0009)	(0.0044)
Just moved	0.004	0.015
	(0.0011)	(0.0046)
Taxable Income		
Negative	-0.021	-0.016
	(0.0181)	(0.727)
0 to 5,000	0.007	0.044
	(0.0012)	(0.0061)
5,001 to 10,000	0.005	0.017
	(0.0012)	(0.0054)
30,001 to 50,000	-0.004	-0.020
	(0.0025)	(0.0100)
50,001 and above	0.009	0.057
	(0.0045)	(0.0241)
Received Welfare	0.000	-0.002
	(0.0030)	(0.0095)
Demographic Variables		
Female	-0.009	-0.007
	(0.0011)	(0.0067)
Number of Children	-0.012	0.002
	(0.0012)	(0.0045)
Married Male	-0.012	-0.157
	(0.0017)	(0.0160)
Married Female	-0.010	-0.154
	(0.0021)	(0.0182)
Newly Married	0.006	0.092
	(0.0041)	(0.0227)
Newly Single	0.003	-0.037
	(0.0048)	(0.0507)

(continued)

Table 3.1 (continued)

Variable	Sample Persons	Non-Sample Persons
Age		
1 or less	-0.018	0.093
	(0.0030)	(0.0173)
2 to 16	-0.006	-0.060
	(0.0027)	(0.0119)
17 to 25	0.006	0.035
	(0.0021)	(0.0094)
26 to 44	-0.001	0.023
	(0.0015)	(0.0100)
65 to 98	0.044	-0.068
	(0.0028)	(0.0142)
Unknown	0.284	-0.002
	(0.0269)	(0.0400)
Relationship to the Head		
Husband	0.077	-0.055
	(0.0443)	(0.0451)
Wife	0.000	-0.001
	(0.0025)	(0.0333)
Child	0.001	-0.016
	(0.0019)	(0.0176)
Sibling	0.37	0.103
	(0.0060)	(0.0210)
Parent	0.067	0.071
	(0.0075)	(0.0228)
Grandchild	0.013	0.020
	(0.0033)	(0.0190)
Other relative	0.068	0.113
	(0.0044)	(0.0168)
Unrelated	0.203	0.015
	(0.0122)	(0.0169)
Census Division		
New England	-0.010	-0.089
	(0.0027)	(0.0153)
East North Central	-0.008	-0.066
	(0.0015)	(0.0076)
West North Central	-0.015	-0.077
	(0.0019)	(0.0094)
South Atlantic	-0.011	-0.091
	(0.0015)	(0.0071)
East South Central	-0.017	-0.091
	(0.0018)	(0.0089)
West South Central	-0.003	-0.076
	(0.0020)	(0.0082)
Mountain	-0.011	-0.054
	(0.0029)	(0.0137)
Pacific	-0.003	-0.044
	(0.0018)	(0.0082)

(continued)

Table 3.1 (continued)

Variable	Sample Persons	Non-Sample Persons
Year of First Interview		
1970		-0.024
		(0.0083)
1971		-0.024
		(0.0085)
1972		-0.024
		(0.0082)
1973		-0.014
		(0.0087)
1974		-0.012
		(0.0094)
1975		-0.052
		(0.0096)
1976		-0.009
		(0.0094)
1977		0.000
		(0.0123)
1978		-0.005
		(0.0119)
1979		0.030
		(0.0121)
1980		-0.011
		(0.0147)
R-square	0.028	0.098
F-statistic	155.53	57.32
Number of observations	207,308	26,439

Note: All variables except "Number of Children" are dummy variables. The coefficient on the intercept is an estimate of the probability of leaving the sample before the next wave for a SRC, non-splitoff, single male head of household aged 45 to 64 who has previously answered at least one survey, who does not plan to move in the future, who did not move since the last survey, who has a taxable income between $10,000 and $30,000, who receives no welfare income, and who lives in one of the Mid-Atlantic states. If this male is a sample person, then the probability is 0.049. If he is a non-sample person, the probability is 0.244

The F-statistic is for the hypothesis that all the coefficients except the intercept are equal to zero. The numbers in parentheses are standard errors.

the last observation (1971). Since we do not know whether or not individuals interviewed in 1981 will respond again in 1982, the final wave of the survey is omitted from this regression.

This technique of "stacking" observations for individuals injects substantial serial correlation into the error term. It also exaggerates the apparent degrees of freedom. As a result, it is difficult to interpret the estimated standard errors of the coefficients. Our own attempts at correcting these errors suggest that multiplying the reported standard errors (dividing the t-statistics) by eight provides a conservative estimate of the true significance of each coefficient.[10]

With the exception of the number of children, all the right hand side variables are dummies. The baseline variables are the SEO indicator, the Census division dummies, and, for the non-sample persons, the indicators for the first wave in which the individual appears. All the other variables are measured as of the current wave.

The first point to note in these regressions is their poor fit overall. The regression for sample persons has an R-square of 0.028; the regression for non-sample persons has an R-square of 0.098. Clearly, these regressions do not predict imminent attrition at all well.

With this said, we note that the most significant explanatory variable is the dummy that indicates whether or not this is the individual's first interview. Holding all other factors constant, a sample person is 7.8 percent more likely to attrite after the current interview if this interview is their first. For a non-sample person, the increase in the attrition probability is 6.7 percent. For the sample persons, the likelihood of attriting after the current wave is somewhat higher for individuals who are 65 to 98 years old and much higher for those whose age is unknown or unreported. (These latter

individuals are probably also older than average.) Individuals who are unrelated to the head of household or whose relationship is not one of those listed are likelier to attrite in any period than members of the immediate family.

B. The Probability of Exiting before 1969, 1975, and 1981

The regressions reported in Table 3.1 highlight the characteristics that are correlated with imminent attrition. However, there are problems in determining the significance of coefficients in these regressions because of the way that observations are "stacked". We can overcome this difficulty if we are willing to use only one observation for each individual regardless of the number of waves in which the individual participates. Table 3.2 contains three such regressions. The sample in each regression is composed of the members of the original 1968 wave of the PSID. The dependent variable in each regression is an indicator variable that is equal to one if the individual exits the PSID by the year listed at the top of each column. Thus, for the regression reported in the middle column, the dependent variable is equal to one if the individual attrites by 1975 and is equal to zero otherwise. All the right hand side variables are dummy variables, and they are all measured as of 1968.

As before, all three regressions have very low explanatory power - the R-squares range between 0.033 and 0.057 - even though the right hand side variables are significant. The general patterns of estimated coefficients are similar to those reported in Table 3.1 with the exception of the coefficient on the SEO indicator. In Table 3.2 this coefficient is significantly negative in the regressions explaining exit by 1969 and 1975. However this coefficient is positive in the regression explaining exit by 1981. The SEO sample had been interviewed by the Bureau of the Census for two years prior to the start

TABLE 3.2

Attrition by 1969, 1975, and 1981

Variable	Exit by 1969	Exit by 1975	Exit by 1981
Intercept	0.137	0.299	0.411
	(0.0102)	(0.0135)	(0.0149)
SEO sample	-0.052	-0.020	0.015
	(0.0057)	(0.0077)	(0.0084)
Female	-0.009	-0.096	-0.054
	(0.0056)	(0.0234)	(0.0080)
Taxable Income			
Negative	-0.065	-0.175	-0.302
	(0.2321)	(0.3017)	(0.3356)
0 to 5,000	0.036	0.064	0.097
	(0.0075)	(0.0099)	(0.0110)
5,000 to 10,000	0.036	0.047	0.068
	(0.0071)	(0.0094)	(0.0104)
30,001 to 50,000	0.116	0.126	0.056
	(0.0413)	(0.0545)	(0.0602)
50,001 and above	0.154	0.192	0.205
	(0.0510)	(0.0741)	(0.0745)
Age			
1 or less	-0.033	-0.096	-0.130
	(0.0180)	(0.0234)	(0.0259)
2 to 16	-0.017	-0.101	-0.090
	(0.0131)	(0.0172)	(0.0191)
17 to 25	0.026	0.004	-0.022
	(0.0110)	(0.0129)	(0.0158)
26 to 44	0.001	-0.038	-0.068
	(0.0077)	(0.0110)	(0.0121)
65 to 98	0.014	0.177	0.281
	(0.0131)	(0.0173)	(0.0192)
Unknown	0.333	0.264	0.150
	(0.0760)	(0.1008)	(0.1111)
Relationship to the Head			
Husband	0.089	0.274	0.172
	(0.1894)	(0.2491)	(0.2774)
Wife	-0.006	-0.013	-0.017
	(0.0082)	(0.0111)	(0.0129)
Child	0.014	0.047	0.044
	(0.0112)	(0.0154)	(0.0169)
Sibling	0.077	0.187	0.226
	(0.0278)	(0.0370)	(0.0409)
Parent	0.146	0.306	0.207
	(0.0308)	(0.0407)	(0.0452)

(continued)

Table 3.2 (continued)

Variable	Exit by 1969	Exit by 1975	Exit by 1981
Relationship to the Head (cont.)			
Grandchild	0.051	0.115	0.102
	(0.0189)	(0.0247)	(0.0273)
Other relative	0.204	0.320	0.250
	(0.0198)	(0.0262)	(0.0290)
Unrelated	0.422	0.430	0.390
	(0.0498)	(0.0066)	(0.0730)
Census Division			
New England	0.015	-0.033	-0.071
	(0.0144)	(0.0196)	(0.0216)
East North Central	-0.023	-0.045	-0.055
	(0.0086)	(0.0113)	(0.0126)
West North Central	-0.082	-0.125	-0.116
	(0.0109)	(0.0142)	(0.0158)
South Atlantic	-0.020	-0.058	-0.076
	(0.0080)	(0.0107)	(0.0117)
East South Central	-0.078	-0.122	-0.149
	(0.0108)	(0.0142)	(0.0157)
West South Central	-0.017	-0.033	-0.017
	(0.0098)	(0.0127)	(0.0139)
Mountain	-0.033	-0.050	-0.084
	(0.0166)	(0.0217)	(0.0241)
Pacific	-0.010	-0.023	-0.013
	(0.0089)	(0.0120)	(0.0135)
R-square	0.033	0.052	0.057
F-statistic	21.23	34.90	38.33
Number of observations	18,387	18,387	18,387

Note: All variables are dummy variables. The coefficient on the intercept is an estimate of the probability of leaving the sample by the specified wave for an SRC, non-splitoff, single male head of household aged 45 to 64 who has a taxable income between $10,000 and $30,000, and who lives in one of the Mid-Atlantic states. The probability that such an individual attrites by 1969 is 0.137. The sample is composed of participants in the original 1968 wave of the PSID.

The F-statistic is for the hypothesis that all the coefficients except the intercept are equal to zero. The numbers in parentheses are standard errors.

of the PSID. This fact may explain the lower initial marginal attrition rates for the SEO sample. The higher eventual marginal attrition rate probably reflects correlations between SEO membership and other characteristics such as income.

C. A Probit Model of Duration

In the regressions above, we searched for variables that are correlated with attrition. Equivalently, we can search for variables that explain duration in the sample. Assume that each individual has an unobservable propensity, call it τ^*, to remain in the PSID and that this propensity is related to observable variables, x, by the equation

$$\tau^* = \beta'x + \varepsilon \tag{1}$$

Assume, in addition, that there are thresholds, τ_j (increasing in j), such that, if the value of τ^* for some individual lies between τ_j and τ_{j+1}, then this individual remains in the sample for exactly j periods.

For those who attrite, the likelihood of remaining in the PSID for exactly j periods is

$$\Phi(\tau_{j+1} - \beta'x) - \Phi(\tau_j - \beta'x) \tag{2}$$

where $\Phi(\ )$ is the standard Normal distribution function. Most individuals do not attrite. For these individuals, we have censored observations of duration: these persons will remain in the PSID for at least as many periods as they already have participated. The likelihood for these observaions is

$$1 - \Phi(\tau_j - \beta'x) \tag{3}$$

where j is the number of periods in which the person has participated as of 1981.

Estimates of this model for male heads of households who are in the labor force are reported in Table 3.3. For the common variables, these estimates

TABLE 3.3

An Ordered Probit Model of Duration

Variable	Sample Persons	Non-Sample Persons
Intercept	-1.351	-0.868
	(0.508)	(0.914)
SEO Sample	0.329	-0.034
	(0.065)	(0.092)
Splitoff	0.405	0.233
	(0.091)	(0.086)
Nonwhite	-0.105	-0.175
	(0.066)	(0.096)
Years of Schooling	-0.003	0.143
	(0.008)	(0.130)
Self Employed	0.176	0.997
	(0.094)	(0.197)
Mobility		
Might move	-0.693	-0.395
	(0.068)	(0.090)
Might move per year	0.055	-1.317
	(0.108)	(0.095)
Taxable Income Splines (coefficients x 1,000)		
0 to 5,000	0.064	-0.168
	(0.062)	(0.119)
5,001 to 10,000	0.079	0.119
	(0.025)	(0.041)
10,000 to 30,000	0.074	0.007
	(0.011)	(0.012)
30,001 to 50,000	0.007	-0.037
	(0.016)	(0.014)
Labor Income Splines (coefficients x 1,000)		
0 to 5,000	0.136	0.302
	(0.063)	(0.096)
5,000 to 10,000	0.157	0.095
	(0.025)	(0.035)
10,001 to 30,000	0.014	0.073
	(0.012)	(0.018)
30,000 and above	-0.049	0.003
	(0.022)	(0.027)

(continued)

Table 3.3 (continued)

Variable	Sample Persons	Non-Sample Persons
Hours of Work Splines (coefficients x 1,000)		
0 to 1500	0.466	0.196
	(0.228)	(0.350)
1501 to 2000	-0.268	-0.296
	(0.235)	(0.329)
2001 to 2500	-0.098	0.008
	(0.161)	(0.260)
2501 and above	-0.587	-0.642
	(0.111)	(0.213)
Hours Unemployed	0.891	0.566
	(0.155)	(0.190)
Demographic Variables		
Married	-0.204	0.333
	(0.083)	(0.555)
Number of Children	0.009	-0.087
	(0.050)	(0.110)
Family Size	0.038	0.056
	(0.045)	(0.102)
Age Splines		
0 to 25	-0.039	-0.037
	(0.007)	(0.022)
26 to 44	-0.010	-0.032
	(0.005)	(0.009)
45 and above	-0.039	-0.013
	(0.007)	(0.021)
Census Division		
New England	0.119	-0.168
	(0.129)	(0.227)
East North Central	0.167	0.052
	(0.083)	(0.144)
West North Central	0.471	0.260
	(0.103)	(0.179)
South Atlantic	0.449	0.280
	(0.080)	(0.134)
East South Central	0.705	0.489
	(0.109)	(0.165)
West South Central	0.299	0.288
	(0.090)	(0.152)
Mountain	0.416	0.365
	(0.184)	(0.216)
Pacific	0.107	-0.034
	(0.092)	(0.148)

(continued)

Table 3.3 (continued)

Variable	Sample Persons	Non-Sample Persons
Threshold Values		
Tau 1	-0.849	-1.487
	(0.033)	(0.065)
Tau 2	-0.591	-1.034
	(0.028)	(0.056)
Tau 3	-0.441	-0.760
	(0.025)	(0.051)
Tau 4	-0.324	-0.520
	(0.021)	(0.045)
Tau 5	-0.199	-0.324
	(0.017)	(0.038)
Tau 6	-0.091	-0.150
	(0.012)	(0.028)
Tau 7	0.0	0.0
Tau 8	0.097	0.096
	(0.012)	(0.024)
Tau 9	0.160	0.188
	(0.016)	(0.036)
Tau 10	0.222	0.298
	(0.018)	(0.050)
Tau 11	0.276	0.383
	(0.020)	(0.063)
Tau 12	0.314	0.502
	(0.021)	(0.096)
Tau 13	0.371	
	(0.023)	
Log likelihood	-3838.15	-1591.30

Note: The sample in each regression is composed of male heads of households older than 17 who have positive labor income and hours of work. Not included are those who are ever retired, disabled, or students. The dependent variable is the number of waves in the longest continuous participation sequence as a male head. The income and hours of work variables are annual averages. All other variables are measured as of 1968.

The numbers in parentheses are standard errors.

are qualitatively the same as those in the previous tables. Because the sample is narrowed to male heads in the labor force, we are able to include variables on labor income and annual hours of work. Duration in the sample appears to be positively related to labor income and negatively related to hours of work. One oddity is the significant positive coefficient on hours of unemployment. This coefficient may simply reflect the increase in aggregate unemployment in the later years of the survey.

IV. Comparisons with the CPS

We have not yet considered whether the PSID was ever representative, that is, we have not assessed the quality of the initial survey. This section of the paper is devoted to that task. We use the Annual Demographic Files (the March Surveys) of the Current Population Survey (CPS) as a basis for comparison with the PSID. We compare the empirical distributions, in 1968, of various demographic characteristics in each data set for both weighted and unweighted samples. We also compare the results of estimating a simple earnings equation for several subsets in each data set.

It is important to note that comparisons of the PSID to the CPS do not yield conclusive evidence that the PSID is or is not representative of the population of the United States.[11] The CPS may not accurately depict the population. There may be particular variables which the CPS measures differently from the PSID. However, the CPS is widely used by researchers to characterize the population, and it is probably the best available benchmark.

A. Empirical Distributions of Demographic Characteristics

We compared the empirical distributions in both surveys of age, sex, race, years of schooling, family income, labor income of the individual, family size, marital status, Census region, employment status, and whether or not individuals are in school. We compared the distributions of these

characteristics for all persons in each survey, for male heads of households, and for wives. For each characteristic and sample subgroup, we compared the unweighted and weighted empirical distributions where the weights are those distributed with the survey data.[12]

The hypothesis that the PSID and the CPS samples were drawn from the same underlying population is strongly rejected for almost all characteristics using standard chi-square measures of goodness-of-fit.[13] The rejection of this hypothesis at the 0.05 percent significance level (that is, the five ten-thousandths level) applies to both weighted and unweighted samples and to all three subsamples. However, the power of these chi-square tests is very high since the number of observations in both the PSID and the CPS is large. For all practical purposes, the differences in these empirical distributions are generally negligible despite the significance of the statistical differences.

There are a few characteristics for which the PSID and CPS samples are similar even statistically. For example, the unweighted sex ratios are insignificantly different across the data sets. Somewhat surprisingly, the weighted sex ratios are different at the seven percent significance level, but the magnitude of the difference is small: for the weighted samples, 49.2 percent of the PSID is male compared to 48.5 of the CPS.

Table 4.1 through 4.6 display the weighted and unweighted empirical distributions of six characteristics for male heads of households, the most commonly analyzed subgroup. Table 4.1 shows that the PSID male heads are younger on average, in 1968, than their CPS counterparts. The PSID has a substantially larger proportion of male heads between the ages of nineteen and twenty-four inclusive.

Table 4.2 displays the race distributions in the two samples. Low income and minority families were deliberately oversampled in the PSID, hence the

TABLE 4.1

Age in 1968, Male Heads of Household

Age	PSID %	CPS %	Difference
Unweighted			
0 - 18	0.260	0.247	0.013
19 - 24	9.422	6.382	3.040
25 - 34	21.012	19.486	1.526
35 - 44	24.509	21.884	2.625
45 - 54	20.665	20.901	-0.237
55 - 64	13.902	16.421	-2.519
65 +	10.231	14.679	-4.448
Weighted			
0 - 18	0.267	0.244	0.022
19 - 24	8.989	6.698	2.291
25 - 34	19.012	19.743	-0.731
35 - 44	23.411	21.621	1.790
45 - 54	19.408	20.766	-1.357
55 - 64	15.057	16.204	-1.147
65 +	13.855	14.724	-0.869
Number of observations	3,460	37,653	

TABLE 4.2

Race in 1968, Male Heads of Household

Race	PSID %	CPS %	Difference
	Unweighted		
White	71.114	90.851	-19.737
Black	25.725	8.236	17.489
Other	3.161	0.914	2.248
	Weighted		
White	88.447	90.897	-2.450
Black	8.889	8.159	0.730
Other	2.664	0.944	1.720
Number of observations	3,448	37,653	

large fraction of PSID blacks in the unweighted distribution is not surprising. When the data are weighted, the proportion of blacks is roughly equal to the proportion in the CPS. There is no reason to expect comparability for the white and other non-white categories because they are defined differently by PSID and CPS. In particular, white and Spanish American are mutually exclusive categories in the PSID.

An interesting difference appears in the comparison of the education levels of the male heads (Table 4.3). In both the weighted and unweighted samples, a smaller fraction of the PSID reports completing exactly twelve years of school. At the same time, a greater proportion of the PSID claims to have completed from nine to eleven years of school. These differences raise the possibility that the PSID may be obtaining more accurate reports of educational attainment. That is, CPS individuals with some years of high school may simply report that they completed high school rather than reporting their actual years of schooling. A similar pattern is observed in the comparisons of wives' education. However, in the wives' data, this discrepancy is more pronounced at the eighth grade level. Relatively fewer PSID wives report exactly eight years of school, and relatively more PSID wives report from one to seven years of school.

Some of the difference in educational attainment between the PSID and CPS individuals may arise from the differences in the age distributions of the two samples. When the education attainment of male heads is compared within age groups, a significant difference (five percent level) is found for the weighted samples only in the 45-54 age group. However, the greater proportion of CPS male heads reporting exactly twelve years of school occurs in all age groups.

TABLE 4.3

Years of Schooling Completed in 1968, Male Heads of Household

Schooling	PSID %	CPS %	Difference
	Unweighted		
Age < 14	0.152	0.000	0.152
Zero	1.155	1.171	-0.016
01 - 07	20.432	13.980	6.451
08	12.983	13.484	-0.501
09 - 11	20.736	17.114	3.622
12	24.567	30.186	-5.619
13 - 15	9.912	10.591	-0.680
16	5.990	7.492	-1.502
17 +	4.074	5.981	-1.907
	Weighted		
Age < 14	0.174	0.000	0.174
Zero	0.975	1.161	-0.185
01 - 07	14.465	13.718	0.747
08	13.718	13.372	0.346
09 - 11	18.392	17.126	1.266
12	27.499	30.354	-2.855
13 - 15	11.521	10.717	0.805
16	7.869	7.490	0.379
17 +	5.387	6.064	-0.677
Number of observations	3,289	37,653	

Tables 4.4 and 4.5 report the distributions among male heads of family and labor income respectively. The larger fraction of low income male heads in the PSID in the unweighted comparisons reflects the deliberate oversampling of low income groups in the PSID. There are two important points to notice in these two tables. First, a smaller fraction of the PSID reports no family or labor income despite the higher proportion of low income persons in the PSID. This discrepancy is consistent with the hypothesis, advanced by Minarik (1975) and others, that the PSID obtains more accurate income reports than does the CPS. Second, note that weighting does not force the income distributions to be the same across the PSID and the CPS. Conditioning on the age distribution does not eliminate these differences, although the weighted distributions of labor income (conditional on age) are significantly different only for the 35-44 age group.

Table 4.6 reveals another interesting difference: the high proportion of very large families in the PSID. This difference does not disappear after the samples are weighted. Neither does it reflect differences in the age distributions. This difference appears just as strongly in the comparisons of the wives.

B. Earnings Regressions

Even if the joint distribution of characteristics in the PSID were nothing like the joint distribution in the CPS sample, the behavior of individuals conditional on their characteristics might be the same across samples. More formally, the estimated coefficients in important behavioral relationships might be the same whether the estimates were calculated from the PSID or the CPS. If the PSID individuals behave in the same fashion, conditional on their characteristics, as their CPS counterparts, then the PSID may provide a sound basis for a variety of analyses regardless of the within-

TABLE 4.4

Previous Year's Family Income Reported in 1968,

Male Heads of Household

Family Income	PSID %	CPS %	Difference
	Unweighted		
$1 - $2499	10.838	9.861	0.977
$2500 - $4999	20.260	14.966	5.295
$5000 - $9999	40.173	40.504	-0.331
$10000 - $14999	18.960	22.840	-3.881
$15000 - $19999	6.098	6.969	-0.871
$20000 - $24999	1.850	2.273	-0.424
$25000 - $29999	0.867	0.953	-0.086
$30000 - $34999	0.318	0.425	-0.107
$35000 - $39999	0.145	0.287	-0.142
$40000 - $49999	0.145	0.210	-0.065
$50000 - or more	0.260	0.236	0.024
None	0.087	0.475	-0.389
	Weighted		
$1 - $2499	8.537	9.896	-1.360
$2500 - $4999	14.800	14.964	-0.164
$5000 - $9999	38.671	40.631	-1.960
$10000 - $14999	24.371	22.743	1.628
$15000 - $19999	8.440	6.898	1.542
$20000 - $24999	2.699	2.302	0.397
$25000 - $29999	1.219	0.943	0.276
$30000 - $34999	0.463	0.425	0.039
$35000 - $39999	0.199	0.293	-0.094
$40000 - $49999	0.201	0.210	-0.009
$50000 or more	0.350	0.226	0.124
None	0.049	0.468	-0.418
Number of observations	3,460	37,653	

TABLE 4.5

Previous Year's Labor Income Reported in 1968,

Male Heads of Household

Labor Income	PSID %	CPS %	Difference
Unweighted			
$1 - $2499	13.256	10.047	3.209
$2500 - $4999	22.200	15.215	6.984
$5000 - $9999	38.842	44.302	-5.460
$10000 - $14999	11.693	12.883	-1.190
$15000 - $19999	2.171	3.030	-0.860
$20000 - $24999	0.695	1.137	-0.442
$25000 - $29999	0.347	0.595	-0.248
$30000 - $34999	0.232	0.252	-0.021
$35000 - $39999	0.029	0.178	-0.149
$40000 - $49999	0.087	0.090	-0.003
$50000 or more	0.145	0.151	-0.007
None	10.304	12.119	-1.815
Weighted			
$1 - $2499	10.302	9.933	0.369
$2500 - $4999	15.586	15.149	0.437
$5000 - $9999	41.696	44.394	-2.697
$10000 - $14999	15.560	12.848	2.712
$15000 - $19999	3.156	3.036	0.120
$20000 - $24999	1.019	1.142	-0.123
$25000 - $29999	0.492	0.594	-0.102
$30000 - $34999	0.327	0.253	0.074
$35000 - $39999	0.039	0.180	-0.141
$40000 - $49999	0.114	0.094	0.021
$50000 or more	0.207	0.146	0.061
None	11.500	12.232	-0.732
Number of observations	3,455	37,653	

TABLE 4.6

Family Size in 1968, Male Heads of Household

Family Size	PSID %	CPS %	Difference
	Unweighted		
1	8.439	7.434	1.006
2	25.405	30.914	-5.509
3	15.780	19.154	-3.374
4	15.723	18.060	-2.337
5	13.035	11.789	1.245
6	7.775	6.531	1.244
7 +	13.844	6.119	7.725
	Weighted		
1	8.199	7.634	0.565
2	30.776	30.986	-0.209
3	17.418	19.184	-1.766
4	16.779	18.030	-1.251
5	12.356	11.698	0.658
6	6.687	6.476	0.211
7 +	7.785	5.993	1.792
Number of observations	3,460	37,653	

sample distribution of characteristics.

We estimated a standard earnings equation for subgroups of the PSID and the CPS and compared the estimates. There are many relationships for which such a comparison could be made. We chose to use an earnings equation for two reasons. First, one of the goals of the PSID is to provide data with which to analyze the determinants of family and individual income. Many of the economic studies that have used the PSID have concentrated on labor supply and income formation. Second, many of the biases that some researchers fear may occur in the PSID are exacerbated when some form of income is the variable to be explained.[14]

We analyzed earnings for three subsamples: male heads of household, female heads of household, and wives. A single specification of the earnings equation was used throughout. The dependent variable is the log of the individual's annual labor income earned in 1967 (reported in 1968). The independent variables include a constant, dummy variables to indicate race (black and other nonwhite -- white is the omitted group), a spline for years of schooling that is split at 12 years, the number of years of job experience (age - years of schooling - 6), the square of the number of years of experience, and dummies for the Census region (the Middle Atlantic region is the omitted group). Other specifications of the equation that included various interactions of the explanatory variables were estimated. None of the substantive results were altered.

Individuals were excluded from the regressions if they did not work at all, if they were full time students, if they were less than 18 years old or more than 64 years old, if they were self-employed, or if their weekly wage (calculated as their reported annual labor earnings divided by their reported annual weeks of work) was less than ten dollars. Additionally, individuals

were excluded from the regressions if any of these data were missing.

Some of the regressions were weighted. In these regressions, the "person weight" was used for the CPS individuals. For the PSID individuals, the 1968 family weight was used. (There was no individual weight calculated for PSID individuals until 1972.) In each weighted regression, the weights were normalized so that their sum within each subgroup was equal to the number of individuals (with positive weights) in that subgroup.

Nine pooled regressions were estimated: three comparing the CPS to all of the PSID unweighted, three comparing the CPS to all of the PSID weighted, and three comparing the CPS to only the SRC subsample of the PSID unweighted.[15]

Table 4.7 reports the marginal significance levels for three different F-tests calculated for each regression. The first F-test (labeled "all coefficients") is calculated for the null hypothesis that none of the coefficients differ across the two data sets. The second F-test (labeled "all but constant") permits the constant to differ across the data sets. Since the PSID members report higher incomes than members of the CPS, other things being equal, the first F-test may be too stringent. This second F-test allows the average income reported to differ across the samples and tests only whether the marginal effects of the explanatory variables are the same. The third F-test (labeled "all but constant and region") allows the effects of location (the coefficients of the Census division dummies) to differ in addition to the average income.[16]

As in the comparisons of the empirical distributions, the large number of combined CPS and PSID observations make the power of these F-tests very high. For example, the null hypothesis (that the coefficients are the same across the two surveys) is never rejected for the female heads even though the

TABLE 4.7

Probability that the Coefficients of an Earnings Equation are Different from those Estimated for the CPS

	Subgroup		
	Male Heads	Female Heads	Wives
PSID, unweighted			
all coefficients	0.00	0.46	0.00
all but constant	0.00	0.46	0.00
all but constant and region	0.00	0.26	0.00
PSID, weighted			
all coefficients	0.00	0.14	0.01
all but constant	0.00	0.14	0.01
all but constant and region	0.00	0.34	0.09
SRC, unweighted			
all coefficients	0.00	0.12	0.07
all but constant	0.22	0.79	0.19
all but constant and region	0.13	0.49	0.10

Note: The entries in this table are the marginal significance levels for F-tests on the null hypothesis that selected coefficients in a common labor income equation are different for CPS and PSID participants. The "all coefficients" hypothesis restricts all coefficients to be the same. The "all but constant" hypothesis restricts all coefficients except the constant to be the same. The "all but constant and region" hypothesis restricts all the coefficients except the constant and the coefficients of the Census region dummies to be the same.

In the common equation, the dependent variable is the log of the individual's annual labor income earned in 1967 (reported 1968). The independent variables include a constant, dummy variables to indicate race (black and other non-white — white is the omitted group), a spline for years of schooling that is split at 12 years, the number of years of job experience (age - years of schooling - 6), and the square of the number of years of experience, and dummies for the Census region (the Middle Atlantic region is the omitted group).

absolute differences between the coefficients are of the same order of magnitude as the differences found in the regressions for male heads. The abundance of observations on male heads guarantees that the null is rejected often. The point to note in Table 4.7 is the failure to reject the null hypothesis for the male heads and the wives when only the SRC subsample of the PSID is included in the regression. This result suggests that the SRC subsample may provide a very good reflection of the underlying population.

Tables 4.8 and 4.9 report two estimates of the earnings equation that compare the SRC subsample to the CPS. Table 4.8 reports estimates calculated using data on male heads of household from the 1968 surveys (the F-tests from this equation are included in Table 4.7) while Table 4.9 reports estimates calculated using data from the 1981 surveys. The format of both tables is the same. The first column of numbers in the table lists the coefficients estimated from the CPS data. The numbers in parentheses are the standard errors of the coefficients. (In order to save space and to highlight the coefficients of interest, the estimated coefficients of the Census region dummies are not listed.) The second column of numbers lists the differences between the coefficients estimated from the PSID data and the coefficients estimated from the CPS data. The sum of the numbers in the two columns is the coefficient derived from estimating the equation using only the PSID data. Thus, a positive number in the second column indicates that the estimated coefficient for the PSID individuals is higher than the estimated coefficient for the CPS individuals. The numbers in parentheses in the second column are the standard errors of the differences between the coefficients. Following the lists of coefficients and differences are the R-square for the equation and the number of observations (CPS in the left column, PSID in the right column) used to estimate the equation.

TABLE 4.8

1967 Labor Income Regression, Male Heads of Household, Unweighted

CPS - SRC Comparison

	CPS	SRC Difference
Constant	7.712 (0.021)	0.066 (0.098)
Black	-0.327 (0.011)	0.107 (0.047)
Other Nonwhite	-0.185 (0.032)	-0.012 (0.091)
Schooling Spline ($\leq$ 12 years)	0.072 (0.002)	-0.004 (0.007)
Schooling Spline (> 12 years)	0.078 (0.002)	0.009 (0.007)
Experience	0.033 (0.001)	0.007 (0.004)
Experience squared	-0.00056 (0.00002)	-0.00012 (0.00008)
R-square	0.305	
Number of Observations	25593	1457
F-tests	F-value	Prob > F
All coefficients	4.015	0.00
All but constant	1.262	0.22
All but constant and region	1.638	0.13

TABLE 4.9

1981 Labor Income Regression, Male Heads of Household, Unweighted,

CPS - SRC Comparison

	CPS	SRC Difference
Constant	8.049 (0.028)	0.224 (0.158)
Black	-0.293 (0.014)	0.043 (0.061)
Other Nonwhite	-0.152 (0.022)	0.103 (0.102)
Schooling spline ($\leq$ 12 years)	0.090 (0.002)	-0.015 (0.013)
Schooling spline (> 12 years)	0.078 (0.002)	0.013 (0.008)
Experience	0.048 (0.001)	0.006 (0.005)
Experience squared	-0.00076 (0.00002)	-0.00010 (0.00011)
R-square	0.206	
Number of observations	31423	1896
F-tests	F-value	Prob > F
All coefficients	7.522	0.00
All but constant	0.883	0.58
All but constant and region	1.459	0.19

At the bottom of each table, we report the same set of F-tests that we reported in Table 4.7. From the results of these F-tests, we see that, if anything, the PSID has grown more like the CPS over time. It may be that the non-sample persons included in the latter regression enter the PSID purely randomly, thus diluting any initial design problems with the PSID.

V. Attrition and Behavioral Relationships

In the previous section, we considered whether the initial draw of the PSID is representative of the population of the United States. The potential biases for which we searched could make it dangerous to infer the size of particular groups, such as the number of Americans in poverty, from the PSID. We closed the previous section by searching for evidence of a different kind of bias, the kind that affects coefficient estimates of behavorial equations. In this section, we continue the search for this second kind of bias.

A. More Earnings Regressions

We re-estimated the same equation for 1967 labor income that was estimated in the previous section. In this case we estimated the equation only for PSID members, but we distinguished between three groups of PSID participants — all 1968 wave participants, those still in the sample in 1975, and those still in the sample in 1981. (Clearly one can be a member of all three groups.) Table 5.1 reports the marginal significance levels for the same set of hypotheses that we used previously. When we were comparing the CPS and the PSID, the last hypothesis, that all coefficients except the constant and the coefficients on the Census region dummies are the same, was of most interest. In comparing one group of PSID members to another, the reasons for calculating all three F-tests are less compelling. The final hypothesis does have the virtue of restricting attention to the coefficients most important to economists. In any event, there is no harm in reporting all

TABLE 5.1

Probability that the Coefficients of an Earnings Equation

Are Different for Those who Attrite from the PSID

	Subgroup		
	Male Heads	Female Heads	Wives
Unweighted			
all coefficients	0.29	0.70	0.35
all but constant	0.42	0.69	0.31
all but constant and region	0.14	0.87	0.48
Weighted			
all coefficients	0.09	0.01	0.09
all but constant	0.18	0.01	0.06
all but constant and region	0.06	0.31	0.35

Note: The entries in this table are the marginal significance levels for F-tests on the null hypothesis that selected coefficients in a common labor income equation are different for those members of the 1968 wave of the PSID who remain in the sample through 1975 and for those who remain in the sample through 1981. The "all coefficients" hypothesis restricts all coefficients to be the same. The "all but constant" hypothesis restricts all coefficients except the constant to be the same. The "all but constant and region" hypothesis restricts all the coefficients except the constant and the coefficients of the Census region dummies to be the same.

In the common equation, the dependent variable is the log of the individual's annual labor income earned in 1967 (reported in 1968). The independent variables include a constant, dummy variables to indicate race (black and other nonwhite — white is the omitted group), a spline for years of schooling that is split at 12 years, the number of years of job experience (age - years of schooling - 6), and the square of the number of years of experience, and dummies for the Census region (the Middle Atlantic region is the omitted group).

three, and, for the sake of consistency, we do so.

For the most part, these results suggest that labor income is determined in the same way for attriters and "survivors". Actually, the absolute differences in the coefficients between groups are of roughly the same order of magnitude as the differences between the CPS and the PSID coefficients. The changes in the marginal significance levels mainly reflects the much smaller sample sizes in these regressions than in the CPS-PSID comparisons.

A curious fact revealed by Table 5.1 is the uniform increase in the marginal significance level of each F-test when the regressions are weighted. We have no explanation for this phenomenon. It does suggest that researchers should be cautious before using the weights supplied by ISR in a regression analysis.

Tables 5.2 and 5.3 report the unweighted and weighted, respectively estimates of this equation for the male heads of household. There are three columns in these two tables. The leftmost column contains the coefficients obtained by estimating the equation for all the male heads in the 1968 wave of the PSID. The middle column displays the differences in the coefficients for those individuals who remain in the sample in 1975. The rightmost column shows the additional differences in the coefficients for those individuals who remain in the sample in 1981. (The coefficient for an individual who remains in the sample in 1981 is the sum of the numbers in all three columns.)

ISR assigns a weight of zero to all non-sample persons, thus these individuals are excluded from any weighted regressions. ISR's decision reflects the difficulty of determining the <u>a priori</u> probability that any particular non-sample person appears in the PSID. However, in a properly specified model, there appears to be no reason to ignore the substantial information available on the non-sample persons.

TABLE 5.2

1968 Labor Income Regression, Male Heads of Household, Unweighted

	1968	1975 Difference	1981 Difference
Constant	8.007	0.120	-0.497
	(0.138)	(0.234)	(0.158)
Black	-0.317	-0.033	0.078
	(0.045)	(0.079)	(0.158)
Other Nonwhite	-0.267	-0.036	0.104
	(0.093)	(0.164)	(0.159)
Schooling spline ($\leq$ 12 years)	0.052	-0.003	0.028
	(0.010)	(0.016)	(0.159)
Schooling spline (> 12 years)	0.101	-0.042	0.030
	(0.016)	(0.026)	(0.159)
Experience	0.033	-0.003	0.014
	(0.006)	(0.011)	(0.159)
Experience squared	-0.00061	0.00005	-0.00016
	(0.00011)	(0.00021)	(0.00019)
R-square	0.428		
Number of observations	2279	1720	1457
F-tests		F-value	Prob > F
All coefficients		1.125	0.29
All but constant		1.031	0.42
All but constant and region		1.433	0.14

TABLE 5.3

1968 Labor Income Regression, Male Heads of Household, Weighted

	1968	1975 Difference	1981 Difference
Constant	8.013 (0.127)	0.026 (0.224)	-0.403 (0.176)
Black	-0.270 (0.057)	-0.030 (0.098)	0.088 (0.176)
Other Nonwhite	-0.177 (0.088)	-0.124 (0.179)	0.173 (0.176)
Schooling spline ($\leq$ 12 years	0.05479 (0.00919)	0.00000 (0.01589)	0.02030 (0.01444)
Schooling spline (> 12 years)	0.094 (0.012)	-0.045 (0.022)	0.038 (0.176)
Experience	0.028 (0.005)	0.008 (0.010)	0.009 (0.176)
Experience squared	-0.00047 (0.00011)	-0.00019 (0.00020)	-0.00006 (0.00018)
R-square	0.373		
Number of observations	2278	1720	1457
F-tests		F-value	Prob > F
All coefficients		1.366	0.09
All but constant		1.236	0.18
All but constant and region		1.728	0.06

We reran the earnings regressions to see if the coefficients for non-sample persons are the same as for the sample individuals. In these regressions, the dependent variable is the log of 1980 labor income (reported in 1981). The marginal significance levels for the F-tests are displayed in Table 5.4. (Since most non-sample persons enter the PSID by marriage, there are not enough non-sample female heads of household to estimate an earnings equation for this subgroup.) We compare the coefficients for all PSID individuals in each subgroup and separately for the SRC and SEO subsamples. With the exception of the SRC male heads of household, there is no evidence that the coefficients are different for the sample and non-sample persons.

As a further test for bias due to attrition, we estimated this earnings equation including dummy variables for the time of attrition as explanatory variables. We experimented with a number of different specifications. In none of these experiments did attrition exhibit any substantial explanatory power.[17]

VI. Concluding Observations

In this paper we examined the dynamics of participation in the PSID and considered whether attrition has affected the representativeness of the PSID. We found some observable variables that are correlated with attrition, but these variables explain only a negligible portion of the attrition in the PSID. We found no evidence that attrition (or entry) has any effect on estimates of the parameters of behavioral equations.

The 1968 PSID sample is quite unlike the population of the United States if we can use the CPS as a benchmark. Weighting the PSID with the weights supplied by ISR goes a long way towards making the PSID sample resemble the CPS sample. While there are still statistically significant differences in the empirical distributions of observable characteristics, most of these

TABLE 5.4

Probability that the Coefficients of an Earnings Equation are Different for Sample and Non-Sample Persons

Sample	Subgroup: Male Heads	Subgroup: Wives
All PSID		
all coefficients	0.65	0.86
all but constant	0.59	0.83
all but constant and region	0.15	0.97
SRC only		
all coefficients	0.18	0.51
all but constant	0.24	0.47
all but constant and region	0.09	0.60
SEO only		
all coefficients	0.97	0.97
all but constant	0.99	0.95
all but constant and region	0.85	0.57

Note: The entries in this table are the marginal significance levels for F-tests on the null hypothesis that selected coeffieients in a common labor income equation are different for sample and nonsample persons. The "all coefficients" hypothesis restricts all coefficients to be the same. The "all but constant" hypothesis restricts all coefficients except the constant to be the same. The "all but constant and region" hypothesis restricts all the coefficients except the constant and the coefficients of the Census region dummies to be the same.

In the common equation, the dependent variable is the log of the individual's annual labor income earned in 1967 (reported 1968). The independent variables include a constant, dummy variables to indicate race (black and other non-white — white is the omitted group), a spline for years of schooling that is split at 12 years, the number of years of job experience (age - years of schooling - 6), and the square of the number of years of experience, and dummies for the Census region (the Middle Atlantic region is the omitted group).

differences are of no practical significance or can be explained by known differences in the coding of answers across the two surveys. For some variables, particularly income and education, there is some reason to believe that the reports in the PSID may be more accurate than those in the CPS. At any rates, the PSID participants behave almost identically, conditional on their observed characteristics, to participants in the CPS.

In a regression setting, the weights supplied by ISR appear at times to accentuate differences between the PSID and the CPS and between subgroups in the PSID. We have no explanation for this phenomenon. Because of the complexity of the weighting algorithms employed by ISR and the number of different weights they supply, we are unable to explore this issue in this paper.

The most striking results of this study is the finding that a substantial percentage of all the data ever collected in the PSID is omitted from the public distribution tapes. These are the observations on individuals who did not respond to the most recent survey. Many of these individuals participated for many years in a row and thus could provide precisely the kind of data needed in studies of behavior over time. In addition, we find no difference between the behavior of the sample persons and the non-sample persons. This finding suggests that it may be cost effective for ISR to try to retain non-sample persons when they leave the sample household.

FOOTNOTES

[1]In the space of this section, we can only give a brief description of the PSID. For a detailed description of the PSID consult the documentation supplied by ISR (Survey Research Center, 1972, Volumes I and II).

[2]For a description of the SRC master sampling frame, see Kish and Hess (1964).

[3]For a detailed description of the design of the 1968 SRC sample, see Survey Research Center (1972, Volume I).

[4]Again, see Survey Research Center (1972, Volume I) for details.

[5]In order to carry out this study of attrition from the PSID, we needed a data set that contained exactly one record for every individual who ever answered a questionnaire between 1968 and 1981. The construction of this data set turned out to be a herculean task. This task was complicated by ISR's policy of reassigning individual identification numbers every year. In addition, information on an individual sometimes appears in the ISR tapes for waves in which the individual did not actually participate.

To overcome thse difficulties, we were forced to design a complicated algorithm that requires data from the merged family-individual tapes for waves 5 through 14 and the single year family-individual tapes for waves 1 through 4. A complete description of our algorithm can be found in Appendix A of Becketti, Gould, Lillard, and Welch (1983). A description of our procedures for eliminating spurious records appears in Chapter 2, Section 2 of the same study.

[6]A central problem whose magnitude was not fully appreciated in advance by ISR is that the "family" or "household" is not a well-defined object in a

longitudinal data set. In this study, we largely ignore the fact that the PSID began as a sample of families. Throughout, we measure the attrition of individuals.

[7]The sheer volume of data and the number of different ways in which it might be examined make it difficult to document our findings in a reasonable amount of space. Throughout this paper, we have been forced to select those tables and regressions that we feel reflect our findings most accurately. Readers who require greater detail should consult Becketti, Gould, Lillard, and Welch (1983).

[8]Both in their documentation and in private conversations, the staff at ISR have noted that they were unprepared for the high attrition rate following the initial wave of the PSID. In response to this experience, ISR revamped its procedures for tracking down sample members and for persuading them to continue to participate in the survey.

[9]In this table, we ignore the possibility that there may be gaps in an individual's participation in the PSID. This omission does not affect any of our conclusions. Out of 30,903 total participants through the fourteenth wave, 4,944 respond to only one survey, 24,103 respond to more than one survey with no gaps, and only 1,856 respond to more than one survey with some dicontinuity.

[10]See Duan (1981) for information on making these corrections.

[11]Minarik (1975) found that the poverty count in the PSID is lower than the count in the CPS. The documentation prepared by the Survey Research Center for the PSID contains several comparisons of the PSID and the CPS.

[12]As we noted above, the PSID is a deliberately non-random sample. A random sample would not have obtained enough data on minorities and individuals in poverty. Thus there is no reason to expect the unweighted

empirical distributions of the PSID and CPS data to be at all similar. We calculated the distributions both ways for two reasons. First, some researchers have neglected to weight the PSID data. Our comparisons give us an indication of the seriousness of this error. Second, by calculating both the weighted and the unweighted empirical distributions, we obtain some notion of the effectiveness of the weights supplies by ISR in eliminating the systematic sampling bias.

ISR has made complicated adjustments over time to the weights. Even if the original weights are correct in some sense, these adjustments may have introduced errors. By limiting our comparisons to the base year of the PSID, we can determine the adequacy of the original weights and we avoid any controversy over which of the many weights supplied by ISR is the appropriate one. Another advantage of comparing the 1968 samples is that there are no zero weight, non-sample persons in the 1968 wave of the PSID.

[13]We treated the PSID as a random sample rather than a probability sample in the chi-square tests. If we had taken account of the sample design effects, the marginal significance level of these chi-square tests would have been higher. Given our observations below on the practical unimportance of the differences in the empirical distributions, we feel justified in overlooking this statistical nicety.

[14]Of course, the results of analyzing one kind of behavioral relationship do not necessarily predict the results for some other relationship. For example, the PSID individuals may mimic the CPS members in their earning behavior, but their divorce and fertility behavior (conditional on observed characteristics) may be quite different.

[15]The weights supplied by ISR apply only to the combined SRC-SEO sample. In their documentation, ISR explicitly warns against using any of their

weights with one or the other of the subsamples alone.

[16]There are slight differences in the definition of Census divisions between the samples which may contribute to a difference in the coefficients.

[17]We estimated a variety of other behavioral equations. In the majority of the equations, attrition played no significant role. The volume of the results prevents us from including them here, but readers interested in the details are invited to consult Becketti, Gould, Lillard, and Welch (1983).

References

Becketti, Sean, William Gould, Lee Lillard, and Finis Welch. "Attrition from the PSID." Unpublished manuscript. November 1983.

Duan, N. "Correcting for Intra-Cluster Correlation in Probit Regression Models." RAND Corporation working paper, 1981.

Kish, L. and I. Hess. "The Survey Research Center's National Sample of Dwellings." Institute for Social Research, University of Michigan, 1964. ISR No. 2315.

Minarik, Joseph J. "New Evidence on the Poverty Count." American Statistical Association, Proceedings of the Social Statistics Section. 1975: pp. 554-65.

Survey Research Center. A Panel Study of Income Dynamics. Institute for Social Research, University of Michigan, Vol. I, 1972, ISR No. 3455; Vol. II, 1972, ISR No. 3456; Vol. III, 1973, ISR No. 3567; Vol. IV, 1974, ISR No. 4013; Vol. V, 1975, ISR No. 3867; Vol. VI, 1976, ISR No. 3917; Vol. VII, 1978, ISR No. 3997; Vol. VIII, 1979, ISR No. 4150; Vol. IX, 1980, ISR No. 4360; Vol. X, 1982, ISR No. 4551.