Notes on the "SEO" or "Census" Component of the PSID*

Charles Brown

October 21, 1996

The PSID Board of Overseers asked that I prepare a brief report on the initial sampling and subsequent attrition by members of the SEO component of the Panel Study. These notes summarize what I have been able to learn.

The SEO/Census Sample: Origins and Sampling Issues

The original PSID sample was in reality two distinct samples which, with proper weighting, could be combined to form a representative sample that accounted for the purposeful over-representation of low-income families.

One part of the PSID sample is called the "SRC" or "cross-section" sample. Here, selection was at a uniform rate, approximately 1/15,300 (Hill, 1992, p. 19). Response rates were calculated by region (4), by central city and suburb in self-representing areas, and by SMSA and non-SMSA in non-self-representing areas (16 cells), and the probability of being interviewed for each cell was the basic selection rate times the response rate in the cell. Response rates in the SRC component averaged 79 percent (Survey Research Center, 1969, Table 1).

The history of the "SEO" or "Census" portion of the PSID sample is more complicated. First, the Survey of Economic Opportunity was conducted by the Census Bureau in 1966. It was a national sample, drawn from 357 PSUs. There were 8 strata, those living in areas with high minority proportions were oversampled, and there were different sampling rates for minority areas in the different strata. In 1967, the households in the same dwelling units--often but not always the same families--were (re)interviewed. Families that participated in the 1967 survey were asked to participate in PSID. Since the purpose of this portion of PSID was to increase the number of low-income families that could be studied, a SEO family was selected only if the family's income was low (and the head was less than 60 years of age) in 1967. "Low" income meant less than or equal to $2000 + $1000*(number of family members)--i.e., $3000 for a single individual or $6000 for a family of four (Hess, 1985, p. 93). This cutoff amounted to about 175 percent of the poverty line at the time, so the SEO families were low income, but not necessarily poor. Among eligible families asked to participate

---

* My thanks to Irene Hess, Steve Heeringa, Sandy Hofferth, Jim Lepkowski, Robert Moffitt, and Frank Stafford, not just for their help but for its timeliness.

in PSID, 68 percent (separately for whites and non-whites) agreed to allow the Census Bureau to send their name and address to SRC.

While the original SEO sample was drawn from 357 PSUs, SRC used the subset of metropolitan PSUs in which it could field interviews plus a sampling of non-SMSA PSUs in the South. The 90 PSUs used by SRC were sampled from the 357 used by Census. This sampling of PSUs was handled by Joe Waxburg at Census; selection probabilities were adjusted appropriately (see below).

Not all of the names and addresses of families that agreed to be interviewed were actually transmitted to SRC. This sampling was done by a computer consulting company in Washington (Hess, 1985, p. 97). From written records, it is clear that non-white families were more likely to be in the "transmitted" group than white families (55 percent vs. 21 percent: Survey Research Center, 1969, Table 4). There was an apparently minor error in defining income for determining eligibility (car value counted as income, excluding some low- income families with good cars); and "some part of the data file was omitted (in what manner remains unknown)" (Hess 1985, p. 97.)[1] Among families SRC attempted to contact, the response rate was 71 percent (Survey Research Center, 1969, Table 1).[2]

---

[1] According to the PSID Users' Guide (Survey Research Center, 19xx), p. E-4,

> they [the computer consulting firm] apparently omitted some fraction of the names and addresses, hopefully not in any systematic way, resulting in fewer cases than expected. We are forced to consider the area to area variation in this fraction, which is substantial, to be the result of an essentially random process. By the time we realized that not all the addresses of the 'signers' had been forwarded, the Census personnel knowledgeable about the process had moved on to designing the 1970 Census, and OEO personnel were not able to provide us any information. Our repeated efforts to secure more information about the lost cases were not successful.

Overall, my understanding is that the Sampling Section originally expected to subsample the SEO respondents who were eligible and agreed to be interviewed. In the event, this sampling was done by the computer consulting firm retained by OEO. The cases actually transmitted were not quite enough for SRC to obtain its target of 2000 SEO respondents. In effect, the weighting procedure employed by PSID assumed that eligible respondents who agreed to be interviewed were selected randomly within area x race cells.

[2] SRC attempted to interview the family "head", who was not necessarily the same person who was interviewed by the Census Bureau in 1967 and agreed to the transfer of name and address

Thus, in order for a low-income family from the Census or SEO subsample to be interviewed by PSID, several events were necessary:

(1) the dwelling unit was selected by the original SEO

(2) the dwelling unit was part of the subsample selected by the Census Bureau--all dwelling units in this subsample were revisited by Census interviewers in 1977. (Typically, but not always, the same family was living there. It appears that Census did not return to households where it encountered nonresponse in 1966--see Survey Research Center, 1969, p. 1)

(3) the PSU was part of the 96 PSUs covered by SRC interviewers

(4) the reporting adult agreed to have his/her name and address transferred to SRC

(5) the name and address was transferred to SRC (SRC attempted interviews with all of these families)

(6) the family was successfully interviewed by PSID.

The probability of events (1)-(3) was calculated by PSU, though no correction was made for nonresponse at either wave of Census's interviewing.[3] The joint probability of (4) and (5) conditional on (3)--i.e., the ratio of addresses transmitted to eligible families--was calculated separately by race by area (area=group of PSUs); response rates were calculated for 9 geographic cells (4 regions x self-representing or not self-representing area, with non-self-representing areas in the South divided by SMSA/non-SMSA). The probability of an interview was then the product of these probabilities.

Each observation was assigned a weight equal to the inverse of its probability of being interviewed. Because each low-income family could be in the sample by either as part of the SRC cross-section or the Census SEO sample, weights of low-income families from either sample were adjusted appropriately.

Weighted distributions of families across the 16 geographic cells (region by central city or suburb for self-representing areas, and by SMSA or non-SMSA for other areas) were compared to 1960 Census data; distributions by region by race of head were compared to 1968 CPS data. These distributions lined up well, and provide some reassurance that the weights were calculated appropriately.

Nevertheless, there is reason for concern about the construction of the original Census SEO sample in the PSID:

(Survey Research Center, 1968, p. 6).

[3] Survey Research Center, 1969, p. 1. The sample weighting calculation on p. 4 is consistent with this.

(1) It appears the weights were not corrected for non-response to the 1966 or 1967 SEO surveys;

(2) the procedures used to determine which of the families that agreed to participate were actually passed on to SRC are not well understood;

(3) the probabilities of reaching steps 4-6 were calculated as functions of geography and race, but not other plausible family characteristics. Thus, if the poorest of the poor were less willing to be contacted by SRC, or less willing to be interviewed when contacted, nothing in the weighting procedure would correct for this.

One way of assessing the extent to which these problems make the SEO/Census sample unrepresentative of the low-income population is to compare initial (1968) PSID data to CPS data for the same year. If the SEO/Census is unrepresentative in ways not corrected by the weighting procedure, we might expect that PSID and CPS data (from a single cross-section) would differ appreciably. There are two limitations of such a "test": first, because the SEO/Census sample was limited to roughly the poorest third of families, errors in that component could lead to only relatively subtle errors in the overall PSID means; second, differences in variable definitions or survey procedures could lead to differences between PSID and CPS values for some variables even if the SEO/Census subsample were representative of the low-income population. Fitzgerald, Gottschalk, and Moffitt (FGM) 1996 mention wording of the education question, definition of "headship", and some noncomparabilities in available measures of labor income in the early years of PSID.

FGM (1996, tables 34, 39, 42, 45, and 48) compare PSID and CPS proportions by race, education, marital status, region, employment, and welfare participation, and mean values and dispersion of earnings. For male heads, female heads, and wives, PSID and CPS are very similar whether or not one includes SEO.[4] For other adults, differences are a bit larger--perhaps because sampling error is more pronounced--but including the SEO subsample again has little impact on the comparison.

---

[4] For male heads, PSID earnings are about 5 percent higher and show less dispersion. The earnings concepts are not strictly comparable, and including the SEO marginally reduces these disparities. For wives, PSID tracks CPS well (PSID showing about 4 percentage point lower probabilities of working during the year being the most important difference), and including SEO has no effect on the PSID-CPS comparison. For female heads, including SEO reduces the PSID-CPS difference in earnings (4% vs 13%) but PSID without SEO does a bit better job tracking CPS earnings dispersion. The most important difference here is in probability of welfare receipt (.14 in CPS, .07 in PSID) but including SEO has no impact.

Becketti et al (1988, pp. 483-484) compare distributions of "age, sex, race, years of schooling, family income, individual labor income, family size, marital status, census region, employment status, and whether or not individuals are in school." They conclude that while PSID-CPS differences are typically statistically significant, given the large samples involved, "[f]or practical purposes differences in the empirical distributions are negligible."

For many analytic purposes, whether PSID reproduces the relationships found in other data sets like CPS or the decennial Census is more important than its ability to reproduce sample means. Problems with the SEO design could lead to PSID regressions (weighted to reflect PSID's oversampling of low-income families) differing from regression models estimated from other data sets where oversampling of poor families is not an issue.

Table 1 presents PSID-CPS differences in coefficients from simple earnings functions, taken from Becketti et al, 1983. The dependent variable is the logarithm of 1967 earnings as reported in 1968. Thus, for example, the upper left entry .030 (.028) means that, using the entire PSID sample but not using sampling weights, the earnings functions for male heads using PSID produced a .030 larger (i.e., less negative) effect of being black than did CPS, and the standard error on this difference was .028. The last line of each column tests the joint hypothesis that race, education, and experience coefficient differences are all due to chance alone.

The most informative comparisons are between the columns that use the full PSID weighted and those that use only the SRC sample. In principle, the addition of the SEO sample should allow us to estimate the effects associated with being in that sample (in particular, effect of being black, and perhaps return to schooling up to high school) with smaller standard errors, but both full PSID and SRC-only results should be similar (and similar to CPS). Problems with the SEO sample would show up in larger PSID-CPS discrepancies when that sample is included.

For male heads, including the SEO sample marginally reduces the PSID-CPS difference for the black coefficient, but otherwise there are few differences. For wives, the SEO reduces the discrepancy for the black coefficient and, more marginally, for the returns to the first 12 years of schooling. For female heads, including the SEO sample leads to a larger difference in the black coefficient, but a somewhat better job of tracking CPS experience profiles. The probability of rejecting the hypothesis that PSID-CPS differences are due to chance alone is lower if one uses only the SEO sample, but this appears to be due to smaller sample sizes reducing the "significance" of the estimated

differences, rather than those differences becoming smaller if the SEO is excluded.

FGM (1996, footnote 39) report that PSID-CPS disparities are similar whether or not they include the SEO sample in the PSID sample. Their regressions include marital status and welfare participation, in addition to earnings, as dependent variables.

On the whole, Table 1 is neither a ringing endorsement nor a compelling indictment of the SEO sample. The standard error estimates for female heads do suggest that analyzing this group is appreciably more difficult with the SRC sample alone.[5] About 60 percent of this group is from the SEO sample.

Attrition in the SRC and SEO/Census Samples

Table 2 presents the basic pattern of mortality-adjusted attrition rates (i.e., attrition for those still alive in 1989).[6] Attrition is somewhat higher among members of the SEO sample (45 vs 39 percent for those who were husbands or wives in 1968, and 49 vs 39 percent for those who were children of heads. To the extent that mortality rates are higher for low-SES individuals in each race-sex-age cell, SEO mortality is underestimated, and attrition likely to be overestimated. But as FGM 1996 note, mortality corrections are relatively unimportant except for the oldest respondents.

The value of the SEO/Census sample depends not only on the way it was originally selected but on subsequent attrition experience. FGM (1994, 1995, 1996) have studied attrition in the PSID in considerable detail. While the SEO sample is not highlighted in their work, several of their findings are revealing.

---

[5] The standard errors in Table 1 are for PSID-CPS differences. The corresponding variance is equal to the sum of the variance for the PSID estimate plus the variance for the CPS estimate. But since the CPS sample is about ten times as large as the PSID sample, the standard error of the PSID-CPS differences is fairly close to the standard error for the PSID coefficients themselves.

[6] The 1989 individual weight is equal to the 1968 weight divided by the mortality-adjusted probability of remaining in the sample. So our estimate of the mortality-adjusted continuation rate for those who remain in the sample in 1989 is the 1968 weight divided by the 1989 weight, and attrition rate is one minus the continuation rate. For the full sample, our overall attrition rate is 43 percent, compared to FGM's (1996, Table 2) 45 percent. (FGM adjust attrition rates with their own life table calculations, but do not report mortality-adjusted rates for SRC and SEO samples separately.)

FGM 1994 report that economically disadvantaged children and parents had higher attrition rates than other PSID sample members Tables B2a and B2b). SEO children have higher attrition rates even after controlling for measures of socioeconomic disadvantage, though this difference is not enormous (roughly .04 on a base attrition rate of about .5 (FGM, 1994, Table B3)); for SEO adults there is no evidence of higher attrition of SEO members after controlling for socioeconomic differences (FGM, 1996, Tables 18-24).

Whether patterns of attrition are different for SEO sample members than for other low-income PSID members is unclear. FGM 1996 present means for those age 25-64 in 1989, from both CPS and PSID. The PSID data are, alternatively, (i) weighted using 1989 weights; (ii) weighted using only 1968 weights; (iii) unweighted, using only the SRC subsample. They emphasize the relatively good correspondence between the PSID means using 1989 weights and comparable CPS means. Using the other two sets of PSID means makes no attempt to correct for attrition, and so might be expected to diverge from the CPS means (or the PSID means using 1989 weights) if attrition were non-random.

Table 3 summarizes the mean absolute CPS-PSID discrepancies for each of the three groups (male heads, wives, female heads). My reading of their tables is that, with very few exceptions, the three sets of PSID means all do about equally well in reproducing the CPS means. Using 1989 weights that account for attrition tracks CPS slightly better than using 1968 weights that do not account for attrition. While, as noted above, attrition is not random, it is not sufficiently correlated with these variables to make much difference for overall means (FGM, 1996, p. 35).[7,8]

In general, combining the SRC and SEO samples (using 1968 weights) or just using the SRC sample (unweighted, and again making no correction for subsequent attrition) makes little difference. At the level of these aggregate means, it's hard to see evidence that the larger sample gives more accurate estimates, or that deficiencies in the SEO sample are throwing the full-sample estimates off track.

There is some evidence that including the SEO allows one to track the lower tail of the earnings distribution more accurately (See the line for the variance of ln earnings, which is quite

---

[7] Hill, 1992, p. 23 notes that "the attrition adjustments are a small component of the weights because differential attrition in the PSID is small."

[8] PSID weights do not include post-stratification adjustments that would force PSID distributions and means to mirror those of the CPS. See Survey Research Center, 19xx, p. E-10.

sensitive to low-end values, and the ratio of the 20th percentile to median earnings. Including SEO gives one a slightly smaller discrepancy for welfare participation and individual earnings, for female heads.)

Conclusions

The birth of the SEO sample was affected by non-cooperation or non-response at several stages, and cumulatively these problems raise the possibility that it did not represent the low-income population as well as was intended. Simple PSID-based analyses that include the SEO line up reasonably well with comparable CPS-based analyses, and one does no better on this score by excluding SEO observations.

This "test" is not very powerful--we are asking whether any problems in the SEO, which represents part of the lower third of the income distribution, are severe enough to noticeably distort full-sample means and regressions, and we find they are not. A much more limited set of analyses focused at the low end of the education and earnings distribution are also consistent with the SEO landing roughly on its feet.

Subsequent attrition is somewhat higher for low-SES respondents than for others in the PSID, and this is true for SEO respondents in particular. On balance, however, attrition seems less an issue than initial selection problems in evaluating the SEO sample.

Table 1
Discrepancies Between PSID and CPS Earnings Functions, 1968

	Male Heads			Wives			Female Heads		
Sample	PSID	PSID	SRC	PSID	PSID	SRC	PSID	PSID	SRC
Weights	None	1968	None	None	1968	None	None	1968	None
Black	.030 (.028)	.087 (.037)	.107 (.047)	-.238 (.073)	-.081 (.092)	-.185 (.112)	-.121 (.083)	-.122 (.092)	-.035 (.148)
Not Black or White	-.067 (.067)	.014 (.071)	-.012 (.091)	-.283 (.184)	-.109 (.206)	-.144 (.271)	-.155 (.257)	-.113 (.243)	-.119 (.352)
Education for Ed≤12	-.005 (.005)	-.005 (.006)	-.004 (.007)	-.002 (.017)	-.000 (.019)	-.018 (.023)	-.040 (.018)	-.032 (.021)	-.038 (.033)
Education for Ed≥13	.011 (.007)	.008 (.006)	.009 (.007)	.040 (.022)	.037 (.020)	.034 (.024)	-.026 (.026)	-.019 (.021)	-.019 (.031)
Experience	.007 (.003)	.007 (.003)	.007 (.004)	.002 (.008)	.001 (.008)	.002 (.010)	.007 (.010)	.008 (.009)	.028 (.014)
Experience-squared/100	-.013 (.007)	-.010 (.006)	-.012 (.008)	-.015 (.018)	-.015 (.018)	-.020 (.021)	-.017 (.021)	-.019 (.019)	-.062 (.030)
F-prob	.00	.00	.13	.00	.09	.10	.26	.34	.49

Notes:
All regressions for both CPS and PSID included intercept and region dummies, not shown separately
F-prob tests the joint hypothesis that the PSID-CPS differences are zero for all coefficients shown (i.e., all except constant and region dummies)
All other entries in table are the PSID-CPS coefficient difference, with the standard error of that difference in parentheses.

Source: Becketti et al, 1983.

Table 2
Mortality-Adjusted Attrition Rates
in SRC and SEO Samples

	Attrition Rate	
Demographic Group	SRC	SEO
All	.39	.48
Heads and Wives in 1968	.39	.45
Children of Heads, 1968	.39	.49

Table 3
Discrepancies between PSID and CPS Sample Means, 1989
Heads and Wives age 25-64, 1989

Group:	Male		Heads	Wives			Female		Heads
Sample:	PSID	PSID	SRC	PSID	PSID	SRC	PSID	PSID	SRC
Weights:	1989	1968	none	1989	1968	none	1989	1968	none
Age	.0	.0	.2	.3	.3	.5	-.6	-.3	-.6
Race (2)	.01	.02	.03	.01	.02	.03	.07	.02	.02
Education (4)	.03	.04	.04	.02	.02	.02	.03	.02	.02
Marital Stat (4)	.01	.01	.01				.02	.02	.02
Region (4)	.03	.03	.03	.03	.03	.03	.02	.02	.03
Owns Home (1)	.02	.03	.04	.03	.04	.05	.01	.03	.04
Work Last Yr (1)	.04	.04	.05	.04	.04	.04	.05	.06	.07
Weeks Worked (+)	-2.4	-2.4	-2.3	-2.0	-1.9	-1.9	-2.6	-2.5	-2.1
Hours Worked (+)	7	11	34	-60	-58	-55	-8	-4	22
W+S Earnings (+)	.02	.03	.04				.03	.06	.08
Var ln Earnings	-.123	-.133	-.172	-.172	-.167	-.156	-.115	-.125	-.275
20%ile/Med Earn	.016	.005	.023	-.014	-.023	-.014	.037	.029	.063
Welfare Partic	-.01	-.01	-.01	.00	-.01	-.01	.00	-.02	-.03

Notes:
A number in parentheses indicates the number of categories for that variable. When that number is greater than one, the mean absolute discrepancy in proportions.
For other rows, numbers in the table are mean (PSID-CPS) discrepancies.
A (+) indicates the means are conditional, with only positive values included. Earnings discrepancies are as a proportion of CPS.

Source: FGM, 1996, Tables 36,41,44.

References

Becketti, Sean; Gould, William; Lillard, Lee; and Welch, Finis. "Attrition from the PSID," Unicon Research Corporation, November 1983.

Becketti, Sean; Gould, William; Lillard, Lee; and Welch, Finis. "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," Journal of Labor Economics, vol. 6, no. 5, 1988, pp. 472-492.

Fitzgerald, John; Gottschalk, Peter; and Moffitt, Robert. "A Study of Sample Attrition in the Michigan Panel Study of Income Dynamics," February 1994; revised, October 1996.

Fitzgerald, John; Gottschalk, Peter; and Moffitt, Robert. "The Impact of Attrition in the PSID on Intergenerational Analysis," August 1995.

Hess, Irene. Sampling for Social Research Surveys, 1947-1980 (Ann Arbor, MI: Survey Research Center, University of Michigan, 1985).

Hill, Martha. The Panel Study of Income Dynamics, A User's Guide (Newbury Park CA: Sage Publications, 1992)

Survey Research Center, "Panel Study of Income Dynamics User Guide," 19xx.

Survey Research Center, Sampling Section. "The Reinterview Sample for the Panel Study of Family Economics," May 1968.

Survey Research Center, Sampling Section. "The Weighting Procedures for the 1968 Study of Family Economics," November 1969.