PSID A national study of socioeconomics and health over lifetimes and across generations

Technical Series Paper #00-03

The Quality of PSID Income Data in the 1990's and Beyond

Yong-Seong Kim and Frank P. Stafford
Survey Research Center - Institute for Social Research
University of Michigan

December 2000

This project was supported by funding from the National Science Foundation (SES 9515005).

The Quality of PSID Income Data in the 1990's and Beyond

Yong-Seong Kim and Frank P. Stafford
December 2000

The main goal of the Panel Study of Income Dynamics (PSID) is to provide a data archive for the estimation of a wide array of panel, life course, and intergenerational models. Yet, long experience suggests that by applying cross-sectional weights, the PSID is a good source of information on the distribution of basic economic variables such as income, wealth, poverty status, wages, homeownership, and employment in the larger population.

The sample size is currently projected to rise from about 6,750 families in 2001 to 7,360 in 2005. The sample exhibits this growth because of success in re-interviewing families previously in the study combined with success in following newly formed families as young adults 'split off' to form their own families. Still, with about 7,000 families, this may not be sufficient to provide good measures for specialized subgroups of the population. On the other hand, with a smaller sample it may be possible to do several things not realistic for a large sample study, such as the Current Population Studies (CPS). All studies which collect income data as a core variable may well benefit by pooling resources and sharing ideas on measurement and processing approaches, even if they do not apply equally to each study. The purpose of this paper is to share some of the experiences PSID has had to date in navigating a number of important changes to the study's design and implementation since 1992.

Household (or family) income is a critical economic variable, and has been reported in many national studies, including PSID, the CPS, the Health and Retirement Study (HRS) and the Asset and Health Dynamics of the Oldest Old (AHEAD). How comparable is this variable across these studies? Tables 1 and 2 display the income distribution for the years 1991 and 1992 for two subsets of the population, families headed by an individual age 51-60 and those headed by an individual age 70 or older. The PSID data are from tabulations using family weights.

Table 1

The Distribution of Household Income in 1991: Age 51-60

Percentile	CPS92	PSID92	HRS92
25	19,611	21,178	20,010
50	37,200	40,690	39,000
Mean	44,902	51,085	49,766
75	60,852	64,300	63,000
95	111,329	132,180a	121,200
N	10,407	1,075	5,788

a. the 94th percentile = 119,000

Table 2
The Distribution of Household Income in 1992: Age 70+

Percentile	CPS93	PSID93	AHEAD93
25	9,125	8,228	10,000
50	16,377	15,396	17,000
Mean	24,012	22,041	24,628
75	29,704	26,996	30,000
95	67,600	59,714	67,000
N	10,418	1,093	6,024

The tables show consistent differences between the income measured in the different studies. The reasons for these cross-study differences are not completely clear. A further examination to address the issue of the comparability in household income as measured by the different studies, particularly in the late 1990's, is beyond the scope of this report. The current report is primarily to evaluate the PSID through time to identify any major changes in the income measures which look to be artificial – or the consequence of changing methods. More broadly, the ultimate goal is to make the PSID the 'gold standard' for income information. The ideas learned in improving income measures in the PSID can be shared with other studies.

With this longer term goal in mind, an important Computer Assisted Telephone Interviewing (CATI) redesign issue is improving measures of income by linking income from employment with the employment hours and weeks reports from event history calendar (EHC) methods and by linking income from assets to the asset reports, now that we plan to ask balance sheet information in each wave of interviewing. The Health and Retirement Study (HRS) reports a much greater level of asset income now that the income from assets section has been changed to flow from the balance sheet questions.

A related area needing extensive redesign in PSID (and in other surveys, we believe) is income from unincorporated businesses. Business income is becoming more important and is (and has been) a challenging area for accurate measurement. From the PSID wealth module about 13 percent of our (weighted) sample report owning a business as of 1999 with an average net value, conditional on owning of over $300,000. A preliminary investigation of data from the Health and Retirement Study for wave 2 (1994) indicates that approximately 20 percent of the pre-65 age group reported business income, so a significant share of that pre-retired population has a business income and there is a need to reassess income measurement for older families, too.

How much of the business income is a return to assets invested versus labor income? How much labor income is attributable to which family members, and how many hours of work produced this income? Can sequences be designed to help identify net income flows from the business as distinct from a 'draw' out of prior 'retained earnings'?

To avoid the problem of misreported business income in future data collections, one approach might be to reconfigure the CATI instrument to contain on-the-fly checks of income-related responses. This is not a possibility within the DOS-based Survey Craft program slated for use by PSID in 2001, but is a possibility for the 2003 Windows-based CATI (Blaise from the Netherlands statistical bureau, which represents a major data collection transition for HRS and PSID). The CATI could be designed to provide fill answers to checkpoint questions such as, "From what you have told me, I calculate that your income is $X. Does that seem right?" If "No," rather than reroute through the income sequence, which will have severe respondent burden implications, one might activate an open-ended text box in which the respondent is asked to describe

their income situation. That open-ended text can be used in judgmental editing back in Ann Arbor. To illustrate, this would protect against the 1997 case in which $295,000 was entered by keystroke mistake (rather than $29,500).

One general approach which seems promising is to get selected additional information at the point of the interview rather attempting to infer the 'truth' about problematic reports after the field period. A companion approach is to route respondents through a type of business screener, which has questions that are dependent on business type. Traditionally, farms have been treated as a distinct business type with their own (simple) sequence. We are considering the value of these and other options in consultation with the field operations team.

Getting the right inputs at the point of contact with the respondent is one of three critical domains for accurate economic data. The other two are the nature of the sample and its transition trough time and the processing and imputation approach.

The importance of improving the quality of microdata on income, particularly for a sample representing the full U.S. population, cannot be overemphasized. With the rapid changes in the structure of the economy and continual changes in business and government organizations, micro panel data with high individual and item response rates may prove to be the best data source for measuring long-term economic growth from different sources of income and wealth (Jorgenson, 1998; Denison, 1984). And this is quite apart from the needed use of such data in structural models to achieve an understanding of the underlying dynamics and individual-level variables 'explaining' the micro growth process.

As a baseline we believe the following are true about cross-sectional micro data on income:

1. PSID data are among the best available on income, wealth, active savings and average annual hourly earnings. HRS income data are comparable on most of these, and HRS seems better on income from assets (for reasons noted above) and may not have such an effort as PSID in the calculation of work hours and wage rates. The PSID and CPS data have started to diverge since the mid late 1980's with the PSID showing a wider income spread, higher at the top percentiles and lower at the lower percentiles. From the early 1990's on, the PSID and HRS seem to be capturing more income for a comparable (HRS age range) sample than was CPS. This is suggested by Table 1. At each of the percentile points, and especially in the upper percentiles and, then the mean, HRS and PSID report substantially more family income.

2. In the 1993 comparison with AHEAD and CPS for the elderly population (70+) the PSID shows somewhat lower values of the income percentiles. While the PSID sample size is rather modest, the lower income values are consistent with other comparisons at lower income levels. Namely, PSID seems to get more accurate reports of the poor income circumstances of lower-income families. It should also be noted that 1993 was the first year in which a share of the PSID were collected via CATI, and the PSID/AHEAD comparisons *may* be affected by the mode change for the PSID. More on this below.

3. Early work with the income data from the PSID Income Plus files shows that the post-CATI data, 1993-1997, have a higher variance and that this seems to be concentrated in a small number of cases per year (out of 7,000 – 8,500 cases). The question is: are these cases real or just data artifacts?

4. As a part of our migration to a new editing system, PSID extensively reworked our Income Processing Software (IPS) in the spring and summer of 2000. This new version of IPS is meant to work in tandem with our new editing system – more on this below. The IPS calculates income and its components in a straightforward way if all elements are there and there are no item non-responses or other anomalies in the

underlying components. For the cases with unusual features, one can often calculate values from partial information. But this has limits. Also, simple keystroke errors in CATI can give rise to seemingly valid, but extreme values. To identify large, artificial changes the program flags ‘large changes,’ particularly if these are large changes across waves in labor income not accompanied by a change in occupation and/or industry. These flagged cases are looked up to assess the larger context of the record and possible interviewer text fields of notes taken during the interview. If a better judgment edit can be made, a value is assigned. If not, a simple imputation may be used. Cases so modified are recorded as modified for the user. In the judgmental edit phase the approach is not to simply second guess the report of the respondent but to identify real evidence of misreporting. For specific illustrations of such editing, please visit our "Notes on the Income Plus Files" discussion in the Income Plus files at our website.

5. The IPS output is intended to be then edited within an editing system. This system is described more fully below. In the interim we have effected a simple (but very labor intensive) prototype version of that system by creating Excel files of the input that goes into an income calculation and using that information for editing. This has led to what we regard as excellent quality income data for labor income and total family income. There are still some very minor improvements to the income files, 1994-1997 and 1999 which will be carried out in our new edit system. How does our pre-archive income look?

An overview is provided in Table 3 where we have carried out the application of the extensively revised IPS to process income from its detailed components and have carried out numerous checks for both cross-sectional and cross-wave outliers to identify potentially anomalous cases. It also has also treated the SEO (low income sample from the Survey of Economic Opportunity in 1968) sample differently. In the case of imputations for missing data it has developed within-SEO sample conditional values to avoid imparting a systematic upward bias for those cases. Across all four waves, 1994-1997, there were about 800 cases deemed to need some checking. (Sort of like challenges in the NFL.)

Upon such further review, a total of about 200 cases actually had their values altered based on detailed assessment of each of these 800 cases. If no reasonable judgment was possible, a simple imputation was made if the data were missing. Most of the reviewed cases we not altered at all and only a small share had major alterations. Variables with codes indicating the nature of the processing and assignments are provide in the data files. The review process resembles very closely what we plan to implement with our Graphical User Interface (GUI) PSID Editing System. It is just that the new edit system will make this much more efficient and can also correct confusion where the interviewer inadvertently asks the questions about the head in the wife’s sequence and vice versa. In data editing in the PSID these are referred to as ‘head/wife switches’.

In Table 3 we have the results of reediting labor income with the new system (December, 2000), compared to the data originally released in the Income Plus files (March, 1999). The income components edited are labor income of the husband and wife. Some additional checks will be carried out on Business and Transfer income for the final archive release of the income data. These are far simpler income components in the PSID application and apply to many fewer cases.

Table 3

Income Decile Values Before (March , 1999) and After (December, 2000)
Editing With New IPS and Lookups in 2000

	Top unadjusted	
STAT	FAMINC94	NEW FAMINC94
MIN	1	1
MAX	1,209,136	1,209,443
MEAN	46,664	47,159
STD	64,598	64,263
5%	4,420	4,150
25%	16,565	15,944
50%	32,749	33,321
75%	57,414	59,978
95%	121,350	121,521
99%	259,101	267,679
STAT	FAMINC95	NEW FAMINC95
MIN	1	1
MAX	1,154,365	1,120,782
MEAN	48,112	48,353
STD	62,850	63,155
5%	4,986	4,901
25%	18,355	17,543
50%	34,607	34,654
75%	59,565	60,473
95%	123,462	124,543
99%	249,090	249,087
STAT	FAMINC96	NEW FAMINC96
MIN	1	1
MAX	1,542,499	1,542,499
MEAN	47,867	48,125
STD	60,022	60,182

5%	5,889	5,639
25%	18,582	18,141
50%	35,070	35,173
75%	59,627	60,639
95%	123,219	124,270
99%	235,621	232,744
STAT	FAMINC97	NEW FAMINC97
MIN	1	1
MAX	820,508	829,914
MEAN	48,303	48,156
STD	53,611	53,390
5%	5,941	5,892
25%	19,150	18,877
50%	34,985	34,903
75%	60,354	60,635
95%	127,435	125,449
99%	250,623	245,812

Examination of Table 3 shows a number of points.

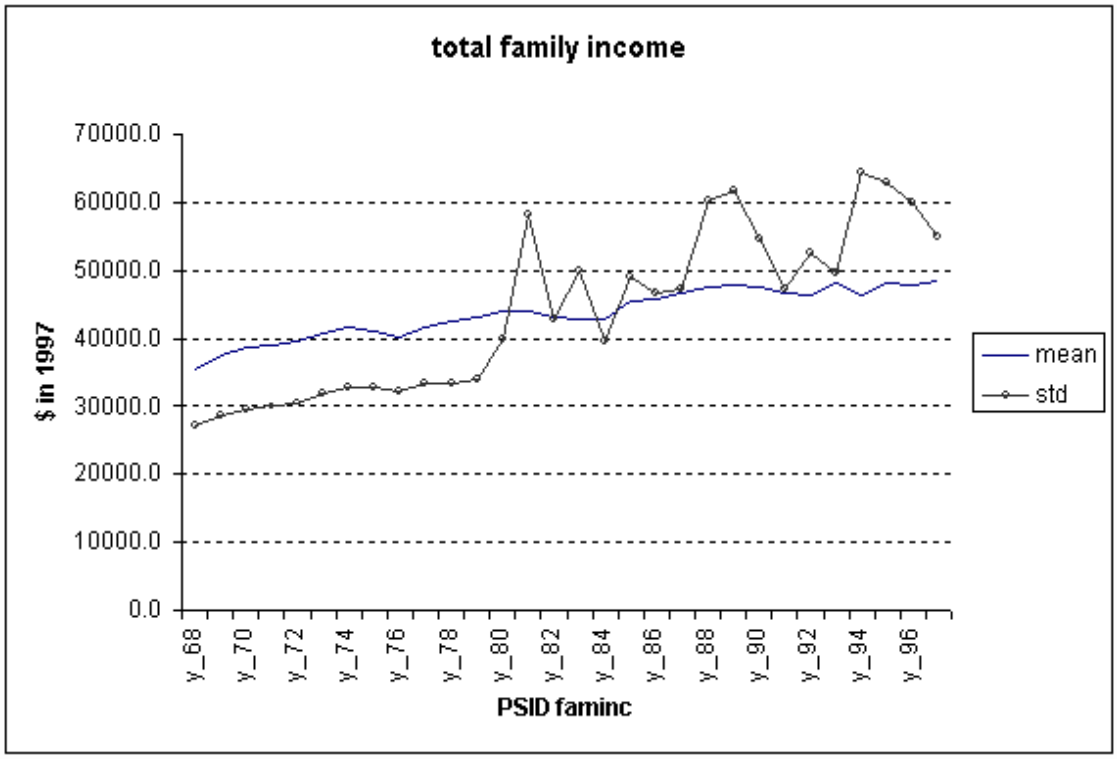
First, for comparison purposes to follow, note that we have set the small number of cases in which total family income is zero or negative to $1. This does not affect any of the discussion to follow, but is a practice for PSID income data prior to 1994. The percentiles (in 1997 dollars and weighted, and recall in 1997 there were important sample changes) show very small changes in values after reprocessing through the new IPS and extensive re-editing checks. At the 95th percentile the 1994 Family Income rises from its initial value of $121,350 (‘faminc94’) to $121,521 (‘new_faminc94’) or by about 0.14 percent. Other changes are also small, and even at the 99th percentile (where we might expect outliers to be more of an issue) the change is only about 3 percent. The percentile changes are both very small and in no obvious direction. The standard deviations, accordingly, are changed by trivial amounts. Similar minor changes in percentiles are observed for Taxable income of the Head and Wife (‘txhw94) and for Labor Income of the Head (‘wghd94’) and Wife (‘wgwf94’).

Moving forward through the years, the standard deviation changes a quite a bit from year to year, falling from 1996 to 1997, at the same time that mean income is very stable. Should this be a concern?

An overview of Total Family Income, 1968-1999 is provided in Figure 1. For the 1994-1999 data, recall that we have carried out the application of the extensively revised IPS to process income from its detailed components and have carried out numerous checks for both cross-sectional and cross-wave outliers to identify potentially anomalous cases. This reprocessing has also treated the Survey of Economic Opportunity

(SEO) sample differently. In the case (the very few) of imputations for missing data, it has developed within-SEO sample conditional values to avoid imparting a systematic upward bias for those cases. Of the approximately 200+ cases which were changed, 1994-1997, more of the changes were for 1994, and fewer going forward to 1997. To a large extent these changes are designed to improve the data quality for panel analysis, but do show up as trivial changes in the cross-sectional variance of income.

Figure 1



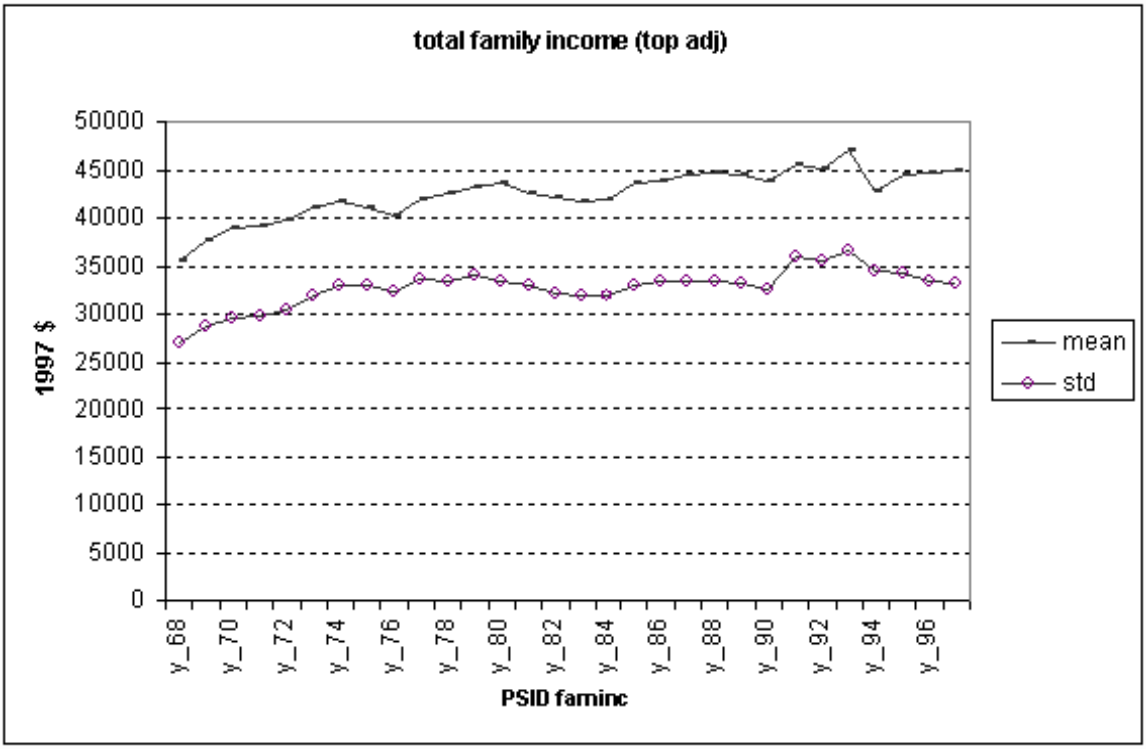
From Figure 1 there are two data 'eras': one 1968 to 1979 and then a post-1980 era. Throughout there is a steady upward rise in (family weighted) mean total family income in 1997 CPI-U dollars, from $35,628 in 1967 (1968 survey year) to $48,156 in 1996 (1997 survey year). The standard deviation of family income appears smooth through time up to 1979 and then from 1980 forward is higher and more variable per year. The explanation for this change is that prior to (survey year) 1979 the topcode value of income was $99,999. In 1980 it was increased to $999,999 and in 1981 it was increased to $9,999,999. The reasons for topcoding at $99,999 prior to 1980 appear to have been limits on logical record length in the early file structures of the PSID and a concern that top values may be dominated by reporting error. The current approach is to provide as many digits as needed to portray the full range of income, to check to be sure that insofar as possible these high values are valid and not keystroke errors, and allow the analyst to deal with issues of robust estimation.

For the period 1968 to 1993, income was also bottom coded at $1. This means that a few families with negative total family income, arising typically from a business loss were set to $1. For 1994-1997 the Income Plus files allow a negative total family income. For the purpose of constructing Figure 1, the negative and 0 values for 1994 to 1997 were set to $1 to maintain comparability with pre-1994 bottom coding. One purpose for bottom coding appears to be simplicity in calculation income-to-needs ratios, a key goal in the study,

particularly in its early years when the central focus was the study of poverty and income needs of the U.S. population.

The data in Figure 1 show a higher standard deviation after the top coding change in 1979-1980. This is not surprising, and Figure 2 shows the total family income series 1968-1997 if the topcoding at $99,999 had been extended throughout. In that case the standard deviation is always below the mean and exhibits relatively little substantial variation year to year throughout. In Figure 2, the standard deviation in total family income in the Income Plus files for 1994-1997 seems quite in line with the long term pattern.

Figure 2



A concern is that even with careful data processing with IPS and individual case editing, the CATI based income could be dramatically different from paper and pencil and individual editing prior to 1993. One simple check is suggested by referring back to Figure 1. Throughout the period when topcoding was extended to $9,999,999 there are much larger year-to-year variations in the cross section standard deviation of total family income. Further, there are three post-1979 ‘episodes’ of notably higher standard deviation values: the early 1980’s recession and recovery of 1981-1983, the late 1980’s expansion of 1987-1989, and the post-Gulf War expansion of 1994-1996. There does not appear to be a pure CATI effect, since 1993 differs from 1994-1996 and 1997 is about in the same range as 1993. Within the CATI years, then 1993 and 1997 are ‘low’ in terms of post-1980 values for the standard deviation, and 1994-1996 are ‘high’ in these same terms.

Another data check is to compare PSID and CPS data. There was a study comparing the PSID and CPS, 1969 -1991, by Greg Duncan, Timothy Smeeding and Willard Rodgers (Household Income Dynamics in the 1970’s and 1980’s, working paper, April 24, 1995), based on families headed by a 25-54 year old. The study showed that, particularly for the Gulf War Recession year, 1991, the PSID and CPS do line up relative to their values as of 1969. However, in the mid to late 1980's the PSID (with 1969 as the reference year) was

showing substantially lower income at the lower (20 and lower) percentile points. One interpretation is that while the PSID gets a more complete income picture (more components and detail and therefore less underreporting), in more recent years we also get a better response from people with low income and assets. This has come up in the wealth area where, in contrast to evidence from some other studies, we show that African Americans are less likely to have bank account in the 1990's compared to the 1980's (see Erik Hurst, Ming Ching Luoh and Frank Stafford, "Wealth Dynamics of American Families, 1984-1994," Brookings Papers on Economic Activity," (1998:I), p. 296-301.) These low income and low asset ownership respondents, one can hypothesize, actually report how low their income and assets are in the PSID.

In summary, there are several points to bear in mind in CPS/PSID household/family income comparisons. 1. 1991 was a special (vintage) year in terms of CPS/ PSID alignment. 2. Both studies have potentially significant data seam problems in the 1990's. The post 1992 survey year CATI and processing of CATI issues for the PSID have been noted above. 3. At the median there appears to be an approximately a 0-5% PSID/CPS difference in medians between CPS and PSID for 1994(93) through 1997(96).

Table 4

Median Family Income PSID and CPS Comparison

	PSID Median (1997$)		**CPS Median (1997$) ***
	Release of		
	March 1999	**December 20, 2000**	
1993 (1994 Survey)	$32,749	$33,321	$34,432
1994	$34,607	$34,654	$34,807
1995	$35,070	$35,173	$35,807
1996 **	$34,985	$34,903	$35,979

* Source: http://www.census.gov/hhes/income/dinctabs.html

** In 1997, we were forced to drop 2,843 of the 3,967 families in the SEO (low income) sample because we were not successful in our efforts to secure funding for the continued data collection for this part of the core sample. The sample suspension as of 1997 and every other year interviewing changes after 1997 are discussed in our April, 1997 PSID Newsletter and our April, 1998 PSID Newsletter (viewable at the PSID website). The 1997 family weights have been designed to apply to either the remaining pre-1997 PSID core

(allowing for the 2,843 dropped SEO cases) or the 1997 core, which now includes the new sample of post-1968 immigrant (P68I) families.

After all is said and done, the PSID income data still have been produced by a changing system. The resources were simply not there to have done this in as systematic a way as would be ideal. In considering income measures in the PSID from the late 1980's to the early 1990's, there are the following differences, notably between 1992 and 1999-03.

1. Income itself has become more derived from diverse and changing sources. There is greater occupational and industrial mobility for male family heads in each successive 5 – year time interval, 1975 – 1995. This mobility should be expected to give rise to more measurement difficulties – the mobility point is certainly likely to be within a calendar year and reporting on transitional jobs is simply more difficult. Income from assets is likely to now be more important and reported asset values for unincorporated businesses are rising. The last year of paper and pencil data collection, 1992 was for 1991, a year which included the Gulf War recession. Also, from 1968 to 1992 there were a number of important changes, which could lead to differences through time. There were some improvements year by year in the sequence of income and work of the wife in the 1980's, and there were month strings added to employment and receipt of many kinds of income in the mid 1980's.

2. CATI-I the early 1993-1994 version creates a sharp break from pre-CATI. In the early CATI years there was (non-random) use of CATI for some cases and paper and pencil for others. This was separate from the small methodological experiment of 200 cases in CATI and another 200 on paper. This study was to compare recording errors and had a quite limited scope (and budget!). The study demonstrated that error recording rates were quite similar between CATI and paper and pencil interviewing.

3. CATI-II 1995-2001 is much better (fewer errors than CATI-I) -- but different. There are different versions of the CATI software and there was a learning curve.

4. CATI-III (Blaise) will be better (but at least somewhat different).

5. CATI-III will include EHC's, which will improve timeline domains - a topic critical for Macro Micro modeling.

6. From 1997 on interviewing is every other year.

7. From 1997 on there are new weights.

8. From 1997 on there is a new sample, including post 1968 immigrants and children born to such immigrants. The income of post-1968 immigrants and their children seems to be sufficiently similar to the rest of the population to have not markedly altered the income distribution in the PSID between 1996 and 1997 - and in comparison with CPS.

9. From 1997 on there is a reduced SEO sample of low income families.

10. From 1993-1999 we have used a DOS-based editing system.

11. From 1993 on there were a series of shifts out of companion processing system programs as the mainframe and a variety of DOS-based software programs disappeared.

12. From 2000 on we have a GUI editing system which is vastly better. We do not have the resources (people!) to go back and re-edit all data from 1990-1997).

One critical tool, which can both improve the post-paper data, particularly for 1999 and forward is the use of our new IPS in conjunction with our new graphical user interface (GUI) editing system (EDITSYS). The EDITSYS can carry out in far more efficient fashion the time consuming work needed to generate the Income Plus files described in Table 3.

In a complex economic panel, and as illustrated with our discussion of income, there is a need to process underlying components from the initial data collection into variables representing useful aggregates for the analyst. The PSID has many such aggregate or 'generated' variables. Notably, these variable include components of family income, each of which may be built up from numerous underlying elements; wealth, built from numerous balance sheet components; active savings, calculated from numerous flow components each with complex skip patterns; annual hours of work; weeks of unemployment; and many other 'calculated' variables.

Each such aggregate may be dependent on numerous components, and, in turn, each of those can be subject to item non-response, specialized reporting procedures (such as unfolding brackets for dollar fields which would otherwise be totally refused), or complex loop structures (such as multiple jobs held within the year) and string indicators (for months worked during a calendar year, for example). Still, the main volume of cases for a given variable can often be directly calculated or estimated - as in the case of valid responses to a sequence of unfolding brackets or simple job loop patterns.

What should be done with the 'other' cases? One temptation is to assign the values by simple or complex imputation systems. As an alternative there is a route which may be called *informed calculation* – as distinct from imputation. There is a wide set of such cases for which values can usually be informed by partial information, including interviewer notes – entered as text fields. The problem for complex aggregates based on a wide array of potentially incomplete elements is that the almost unbounded set of combinations overwhelms the possibility of pre-specifying rules in a software routine. While the first line of attack is to define combinations of conditions and specify a rule for assigning a value, this has limits.

At some point it is more effective to engage in judgmental editing. In this case an editor looks at the partial inputs and possibly contextual material on a case-by-case basis. For example: If for this year there is only a fragmentary report of labor income, what was the respondent's labor income last year? Did the respondent change industry or occupation? Is the extreme value of labor income just the result of a keystroke error? Are there marginal notes (text fields) which the interviewer added during the interview which, in combination with other reported information, could be the basis for a plausible income value?

From 1993 to 1999 PSID has had a computerized editing system to accomplish the task of editing from CATI input. This was based on DOS. It had severely limited flexibility in comparison with the new system. It relied on companion software, which is no longer adequately supported. A new system has been a vital need.

We are pleased to now have a functioning system and companion programs for income editing, performing consistency checks, and identifying and correcting wild codes and 'short codes'. Only in this way can we expect long run improvements in processing productivity and data quality. Another chapter to this story is more effective data collection software. Some of this was seen in Table 3, where improvements in the DOS-based CATI appear to have been achieved cumulatively, 1994-1997 and where extensive editing has preserved the main elements of continuity. Further on this front, we have invested a great deal in validity

studies on the use of semi-flexible event history calendars to achieve better recall over an 18-30 month interviewing cycle. There are plans for a major migration and change to our CATI software for 2003.

The new edit system allows the creation of 'editing screens' which, through the use of 'tabs', can effectively and flexibly display for an editor input information which could provide far better value of the target variable than a simple (or complex) assignment from a multivariate imputation program or by direct calculation. These screens can be customized and saved. These screens assemble the input and other information decided to be of likely contextual value so that this judgment editing can be expedited. The software provides an 'audit trail' of who made the editing decision, when, and why. This is recorded (along with the original inputs). When many possible inputs could inform an editing judgment, including values of variables from prior years, the value of graphical tabs, each of which presents a particular input domain can keep a rich set of information at hand without overwhelming the editor.

In the case of the PSID we have, over the last several years, developed a set of companion software programs , notably the IPS described above, which will provide a directly calculated value for the straightforward cases and which will flag cases deserving customized scrutiny. For the case of income editing, for example, the 'flagged cases' are those with fragmentary inputs not handled by the algorithm, cases with 'extreme' values or with extreme *changes between years*. Then these suspicious cases are processed through the editing system as described above. The editing system allows for the screen to portray values of variables from previous data collection waves – if that is deemed informative. The saving of the screens constitutes a documentation of the editing 'procedures' and the audit trail provides a record of the actual decisions.

Only as a 'last resort' are values imputed by a statistical procedure ('hot deck' or multivariate imputation, for example.) This is a PSID tradition. In pre-CATI days this judgment or 'Gestalt' editing was effected by editing paper worksheets for complex variables and then the essentially 'correct' final values were entered into the nearly final data files. Editors could look at values given elsewhere in the paper questionnaire, at last year's questionnaire and true 'marginal notes' on the perimeter of the questionnaire page in judgmental editing. With direct entry of data and notes in CATI, this process can only be replaced by the type of editing system we have recently developed. In the last years of pre-CATI interviewing, PSID was, for budget reasons, unable to carry out editing based on cross year responses. This new flexible system can restore this desirable feature.

From an estimation perspective, this judgmental editing is much preferred to imputation on a cross-sectional basis. An approach which makes extensive use of cross-sectional imputations invites huge errors in *change* measures and will surely be the downfall of such popular methods as 'fixed effects' panel models. The rush to migrate the PSID to CATI in the early 1990's was done before an effective editing system of the current sort could be written. In fact, it can be argued that even the CATI data collection software of that bygone 'era' was not adequate to the task. The software tools have since improved enormously. The new PSID EDITSYSTEM is written in SAS. Using high level features of this system made it possible to develop the current edit system this year.

The PSID EDITSYSTEM allows editing of the complex relationship files in the PSID. This is known as 'Family Composition Editing'. The PSID follows a blood line which means that essentially individuals and their lineal descendents are followed. This gives rise to potentially great power for intergenerational analysis – but only if these relationships are recorded accurately with ID's and the right persons are followed. In addition, the CATI application collects data for generic entities, such as 'head of household' or 'wife of head of household'. Yet an individual's status as head or wife can change from year to year, through such events as divorce, marriage, and death of a spouse, for example.

The PSID keeps track of individuals and families with a complex ID system. For measuring change in many economic variables, even over shorter time periods of a year or two, one needs to allow for changing family composition. Wealth and income of a family are known to respond to changing family composition in major ways.

To maintain the integrity of this ID system, we have to keep accurate records of who is in a given family in each year, and their relationship to the designated head for that year. Errors can be made in the field. From confidential information on names and other field information we keep track of these relationships from year to year and in addition conduct 'family composition editing'. With the new editing system one can view these family history trees for individuals in the study. Essentially, this amounts to correcting on-the-fly errors which an interviewer may make about who is in the family and their relationship to the head that year. These errors are understandable, given the stressful conditions of a long, complex interview. Here again the editing system is critical since contextual information and information from fragmentary sources can be used to achieve high levels of accuracy. Family composition editing is crucial. What could be a larger error than attribution of the reported income to the wrong member of the family? And then in future years interviewing the wrong person – corrupting the entire sample design!

In brief summary, the PSID has undergone a great number of changes 1992 – 2000 and will be going through many more significant changes in the next few years. During this change process a large number of potential data seams could have arisen.

The analysis in these notes seems to indicate that by recreating the pre-CATI editing process and careful migration to CATI (despite technical problems, especially in 1993-1994) the basic continuity of the income series has been preserved. In future versions of the data collection we hope to improve the measures of income, notably from assets and business. We also believe that error rates from labor income can be reduced by connecting the labor income questions to prompting information from employment timelines from an EHC mode. We have also done work, not reported here, suggesting very high quality of our household wealth measures both pre CATI – in 1984 and 1989 - and post CATI – in 1994 and 1999 (in comparison to HRS and the Survey of Consumer Finances). We believe the wage rate measures (average annual hourly earnings of head and of wife) also to be of high quality.