

## **Panel Surveys: Conducting Surveys Over Time**

A Chapter Prepared for The Handbook of Survey Research

Peter V. Marsden and James Wright, Eds.

Academic Press, 2008

**Frank P. Stafford**

*Department of Economics*

*Institute for Social Research*

*University of Michigan*

*Ann Arbor, Michigan 48106*



## **Panels of the Life Course and Across Generations**

### **Introduction**

Panel data are obtained from surveys which maintain a sample of individuals for the collection of repeated and additional measures of interest over time. An important dimension, which can also apply to a cross-sectional survey, is the collection of information on the parents, children and siblings, or in an employment survey, on co-workers. The features of other persons of interest at a point in time and over time and generations are important for understanding the behavior and outcomes of any one person during a lifetime. Panels may be time delimited or they may be of a steady state type, in which a mechanism is set out to refresh the sample to offset attrition or to bring in new cohorts over a future, indefinite period. Panels may have a short data collection periodicity, such as months or quarters of a year, or annual, such as the British Household Panel Study, or biennial, such as the older cohorts of the National Longitudinal Surveys (NLS), the Health and Retirement Study (HRS) or the Panel Study of Income Dynamics (PSID)<sup>1</sup>.

Started in 1968, PSID has followed a representative sample of approximately 4,800 U.S. families and their lineal descendants for almost 40 years. Funded primarily by the National Science Foundation, NIA and NICHD, the resulting data archive now includes more than 8,000 active families and 65,000 individuals who have ever been in the study. Of the current participants almost 98 percent can be described as having no more than two spells of participation, so most of the active respondents have been in the study for much of their life course. With such a long panel it is possible to study both changes over longer periods such as the cessation of smoking, 1986-1999, and transitions over shorter time segments. Specifically, recent findings have revealed the earnings disadvantage of smokers as of 1986 is primarily for

---

<sup>1</sup> Since 1997.

those who we now know were unable to quit, 1986-1999. There has been an increase in income volatility and a widening gap between the rich and the poor. The relative mobility across wealth deciles also appears to be increasing, despite the widening of the deciles (Hurst, Luoh and Stafford, 1998). And health insurance coverage, while about the same in a cross-section between 1970 and 2000, is now more subject to coverage transitions.

Table 1. Transition Matrix for Household Non-Housing Wealth for all Ages, 1999 to 2001

WealthDecile, 1999	Wealth Decile, 2001									
	1	2	3	4	5	6	7	8	9	10
1 -- lowest	46	13	14	10	6	4	2	1	2	1
2	13	44	16	12	6	3	3	1	1	1
3	12	17	29	18	10	6	4	2	1	1
4	11	11	15	21	17	10	7	3	3	2
5	6	6	11	16	22	19	10	3	4	2
6	5	4	7	10	17	24	16	11	4	2
7	3	1	4	6	11	15	24	19	10	6
8	2	2	2	2	6	10	17	30	21	7
9	1	1	1	3	2	6	12	23	34	17
10 – highest	1	0	1	1	2	3	4	7	20	62

Estimates based on 1999 PSID family weights.

Table 1 illustrates some of the unique uses of panel data for the study of economic transitions but can and has been applied in many areas such to the transitions in body mass index, or balanced panel dataset on labor income of individuals through time

<http://psidonline/Guide/tutorials/tutorial3/balancedpanel.html>. The Table 1 data suggest that mobility in non-housing wealth holdings is substantial. Depending on their initial decile, only 21 to 62% of households remained within the same decile across the years 1999 to 2001. Moreover, by 2001, 18-47% of households, again depending on the initial decile, had moved at least two deciles higher or lower than their standing in 1999. With additional waves of data, a more complete investigation of dynamics on wealth holdings over the life course, including

explanations for these observed patterns, can be conducted. Or one can identify outcome variables of interest from the transition table. Which families were persistently in the lower 3 deciles? Which families moved up from below the median to the upper quintile?

An advantage of panels is an option value. Once the panel has been in place for a while, as new areas of research interest are identified, these can be added for a one time data collection or may be added as a regular feature so long as there is a significant synergy in terms of research based on the pre-existing data elements. For new and revised content the input of those with specific expertise is often essential. Of the more than 40,000 variables in the PSID, a large set is those gathered on a repeated basis over the life of the study. Other measures have appeared intermittently – such as vehicle purchases - and others were one time supplements, such as fertility expectations from a demographic module in 1985. What may be called ‘subpanels’ are also possible. For HRS there has been subpanel on consumption and time use (Consumption and Activities Mailout Survey or CAMS). For PSID, a special part of the study focuses on families with children under the age of 18. Funded mainly by the National Institute of Child Health and Human Development, this part of the study includes a wealth of information about how children spend their time and progress through different stages of development. For more information, visit <http://psidonline.isr.umich.edu/>. This supplement (as with CAMS) has panel data on a different periodicity from the core study; for CDS about every 5 years and to date does not have a design for a refresher. But the CDS children will become the adult sample replacement for the current parents – thus providing added intergenerational strength to the overall data collection.

### **The Steady State Panel**

The design element of following the children of sample PSID families as they left to form their own households was intended initially to offset the loss of young families, through ageing and attrition, during a 5-year period. To follow such ‘splitoffs’, it was argued, would offset the

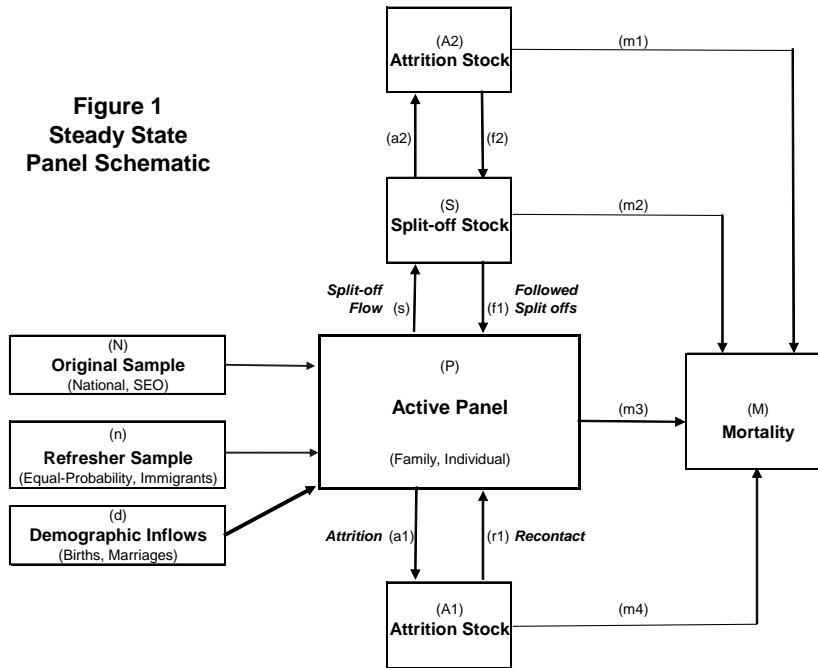
problem of panel attrition and ageing. And by continually adding these young families, a panel study could provide a continuous self-representing sample of the United States population.<sup>2</sup> The HRS has added new cohorts of families entering their pre-retirement years as a way of refreshing their panel. The National Longitudinal Surveys of the U.S. Department of Labor are often 'refreshed' by an entire new sample. In Figure 1 the sample design for PSID is set out, but the main ingredients are, in principle, potential features of all panels. Including the newly formed families of children who left to live on their own provides both continued representation of such young families, and, in addition, supports the study of early adult experiences of children from different economic backgrounds and poverty exposures (Hill 1992). This element of genealogy-based design, which allows the study to maintain the representation of the young (with weights) and to study the effects of family background, was to become a central way in which to create a steady state panel. In effect, the PSID follows a blood line, not a person. Use of this insight opened up the current study of intergenerational connections.

Low-income families were over-sampled in the original design (the Survey of Economic Opportunity or SEO sample in Figure 1); more than one-quarter of the families were black. With the 1997 - 2001 addition of post-1968 immigrants and their adult children (Refresher Sample in Figure 1), the weighted sample is representative of the U.S. population as a whole. The family is a dynamic unit; hence, as noted, the study has followed not only the original 1968 panel families, but also all members of the 1968 families who left home ('split off'). 'Split-off' families are formed when children leave home, when couples divorce and when other changes cause families to separate. There is a stock of split off families and these are brought into the active panel if they are lineal descendants of the sample and when they are followed successfully. This steady state procedure produces a national sample of families each year, as

---

<sup>2</sup> For a discussion of the issue of attrition and representativeness of long-lived panels, see the Journal of Human Resources, 2002,

**Figure 1  
Steady State  
Panel Schematic**



new families are formed through divorce and by children leaving home, and this mirrors similar changes taking place in the population as a whole. Thus, the panel remains representative with respect to its basic sampling design. We refer to this self-regenerating sample as the ‘steady state’ panel in the Figure 1 schematic.<sup>3</sup>

The steady state design relies on following split offs and efforts to bring back in those who have dropped out of the panel via attrition. For maintenance of long term representation it is important to devote resources both to limiting attrition ( $a_1$ ) and re-contacting those who have dropped out ( $r_1$ ). Through decisions on the extent of effort devoted to limiting outflows ( $a_2$ ,  $a_1$ ) and boosting participation ( $f_1$ ,  $f_2$ ,  $r_1$ ), the panel can be ‘steered’ to some extent. Several panels which have an intergenerational design, such as the British Household Panel Study, begin interviewing the children and collecting more extensive information about the individual as an adult at a given age, such as age 18, and have more inclusive rules governing follow status of individuals who ‘marry in’ to the basic genealogical design. Such early

<sup>3</sup> For a review of the effects of attrition on the long-lived panels see *Journal of Human Resources*, 2000.

coverage of the sample and inclusion of greater individual information asked of each individuals directly – as distinct from proxy reports – add greatly to the cost of a panel study.

The significance of the intergenerational family design is the support of research on intra-generational and intergenerational family influences of one family member on another. In the study of labor income or assets and many other topics there is the possibility of linking across different biological family members residing in different households and at different time points. A substantial research topic has been the intergenerational transmission of wealth – supported by information about gifts and bequests to related family members living in different sample families and at different time points. In a recently completed user tutorial the study of the relationship between the earnings of the Baby Boomers age 40-50 and the earnings of their fathers back in the differing, specific years when they were 40-50 is illustrated. For the national sample the cross-generational earnings elasticity is found to be about .45. Other intergenerational relationships appear to be much weaker. The intergenerational correlation of work hours is quite low; by implication the high labor income relationship is not primarily the result of strongly related hours of market work.

### **Collecting and Keeping Track of the Data Collected**

Long-lived panels give rise to a large number of variables that have been asked year after year in much the same, if not identical, manner; these studies usually have also fielded a diverse array of supplements during the operation of the panel. The greatest detail on employment, background and other individual-level measures in the PSID is available for the head of the family unit and, if applicable, for the wife or long-term cohabitor (termed “wife”) of the head. Throughout the various domains of the study a narrower set of information has been collected for other family members who were neither head, nor wife, nor “wife”. In both panel and cross-sectional studies, information about individuals is collected about them as individuals and also about them in specific roles. In household surveys the individual information may be as primary adult householder, spouse of primary adult householder, children, or any of a wide range of other relations to the reference person. PSID has a set of codes for relation to head with 38 different code categories. In a survey of employees in a firm the parallel would be information on

individual characteristics and information on them as they held roles of manager, foreman, and so on. In addition there may be information on factors which are common to those in the sampled units such as housing characteristics, geospatial location and related neighborhood measures, company sales, worksite safety measures (as distinct from injuries to specific workers). Information at these different levels gives rise to the relational data structures need to archive, document, access, and analyze the resulting variables.

These design features and other tangible and intangible elements of data quality <sup>4</sup>-- response rates and item non-response rates are important features of an overall panel data collection. For PSID, the study had been established to continue for five years and continuance beyond that was just the normal researcher's hope of study longevity. The first interview was a simple, respondent-friendly 32-page questionnaire, and pages 31 and 32 were by interviewer observation only. For long panels an important issue is media and software stability. The data must be processed and documented and set into formats so as to be accessed conveniently. The PSID has experienced dramatic changes in information technology starting from 7-track tapes to web-based delivery of customized subsets and codebooks with descriptive statistics delivered from an Oracle based archive. As in any technology-based environment there are certain to be changes through time. Without attention to this fact a long panel can experience the adverse costs of path-dependency if a new and better technology becomes too costly to adopt, often because of the skills and abilities of the incumbent staff. Retraining of the research staff becomes an on-going enterprise.

Even with successful adoption of new technology there will be issues of measurement continuity. In part the new technology will be found to support new measures, some of which may be better, but may be slightly different from earlier measures of the same form. In light of the ancient maxim, "You can't measure change if you change the measures," there is an on-

---

<sup>4</sup><http://psidonline/Guide/Quality/>

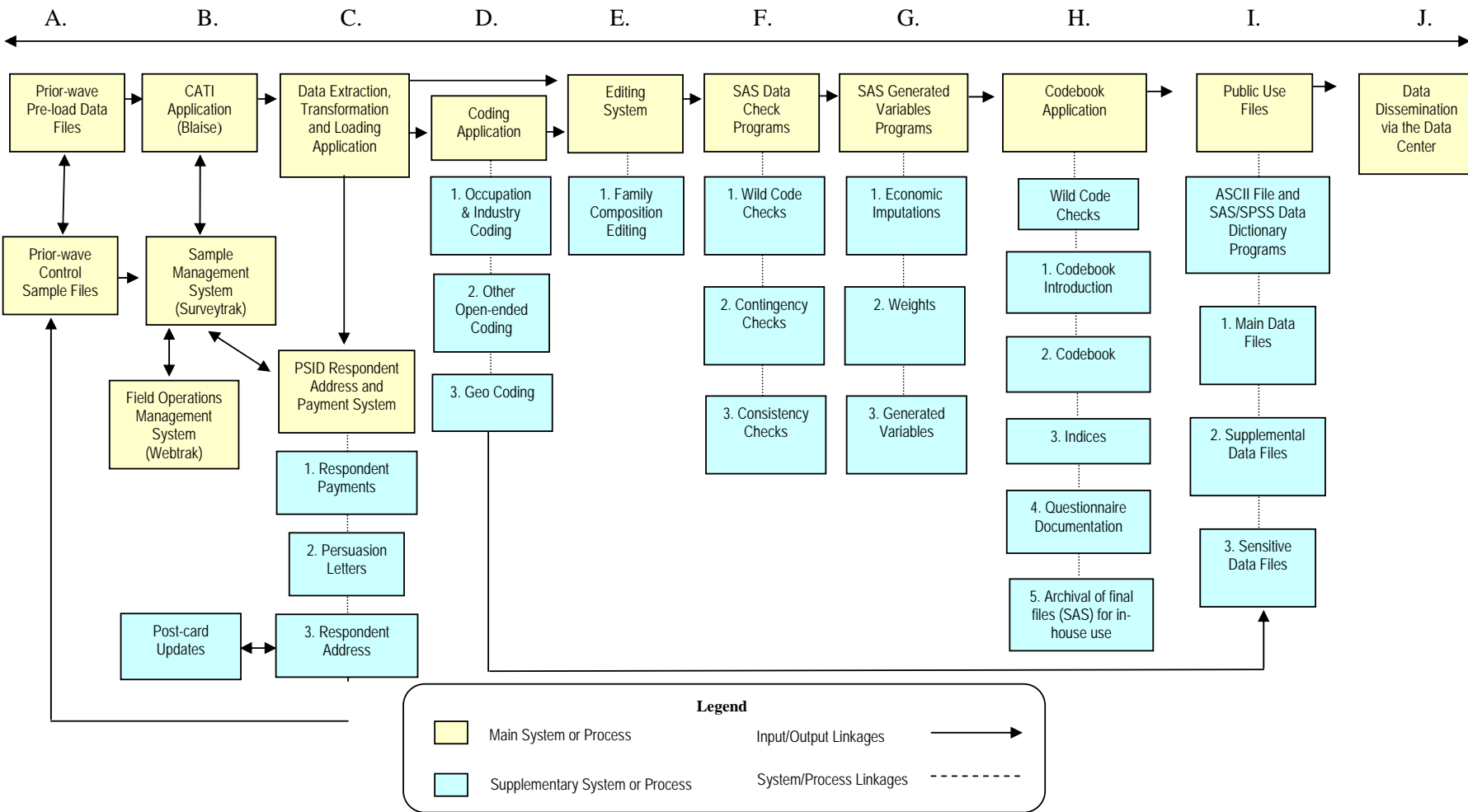
going tension between the new and better versus the consistency through time. A dramatic example is the emergence of new occupations and industries. Here the issue is which occupation codes should be applied to the text field responses. Even if the reports are coded to both old and new codes, the old codes may become less reflective of the reality of current occupations. Further some data collections have shifted to systems of occupation coding through computer assisted menus applied by the interviewer. When such coding and content evolution occurs during the life of a panel there is a need to document when the coding of the variables has changed and under what circumstances the variable can be considered ‘the same’ across time.

A schematic of current PSID data collection, processing and archiving is presented in Figure 2. Any panel must consider these elements of collecting, processing and distributing the data in the initial design phase. The long duration of the PSID has resulted in these functions being handled in very different fashion over the years. In particular, the data were initially collected in person on a ‘paper and pencil’ instrument. The data collection migrated to telephone based data collection – both in a centralized telephone facility and in other years in a decentralized fashion with regional field supervisors managing the interviewers in their region. In 1993 and 1994 the data collection instrument was shifted to computer assisted telephone interviewing or CATI (column B). The fledgling CATI software was quite balky, but in successive years, improvements were made and it became both functional and effective. In more recent years the CATI has shifted to a rather widely used software known as BLAISE.

The initial shift to CATI was quite naive. It was seen as a way to ‘computerize’ and cut the costs of a long-standing function. As the data collection software improved and as research interests in different areas became better formulated, there has been a continual shift to far more complex question contingencies and branching, and to rely on CATI to collect far more information from each respondent and information better targeted to the respondent’s

circumstances. To illustrate, the 2005 CATI application when translated into traditional paper format results in a hefty document of 183 pages, an instrument which could not realistically be administered by traditional modes. Different software has been developed to portray the CATI instrument itself for users as if they were interviewers (H4 in Figure 2). This branching and conditionality of data collection gives rise to a simultaneous increase in the challenge of processing, documenting and using the data. To illustrate, a classic data code for a variable is ('inapplicable'). Good documentation will give the user some of the various conditions which

**Figure 2 Schematic Representation of the PSID Data System**



- A.-B. The development, implementation, and maintenance of pre-production files and production systems is a joint venture between PSID study staff and SRO (SRC's centralized service units).
- B. Computer-assisted data collection application is programmed using Blaise for Windows and Visual Basic is used for the Event History Calendar. The sample management system (Surveytrak) is SRO proprietary software developed in PowerBuilder.
- C. The user interface is an Oracle form that overlays a combination of 1) an ISR Blaise to SAS extraction system, and 2) an in-house hybrid Oracle/SAS based system.
- D. The project uses an in-house Oracle application for geo-coding and occupation and industry coding.
- E. The current DOS/Clipper edit system supports family composition editing; a new Oracle-based editing system is under development
- F.-G. These SAS programs are run independently at present, but will be incorporated into an integrated system linked to the forthcoming editing system, with substantial automation of processing tasks
- F. Contingency checks examine question sequence and skip pattern logic, while consistency checks identify inconsistencies across interrelated variables, and between waves and family versus individual files.
- H.-I. The codebook application is an Oracle/SAS hybrid. Sensitive data files are released only under special contractual arrangements designed to ensure respondent confidentiality and are currently distributed via CD.
- J. The Web-based Data Center is an application that enables users to download a customized subset of variables in various formats along with supporting documentation.

have led to the value being ‘inap.’ But often the full set of paths to ‘inap.’ is now too extensive to effectively enumerate for users without a great deal of resources.

In any survey there will be missing data – some questions are simply not answered in the detail the researcher would like. This is particularly true of economic measures such as values for donations to charity, the market value of owner occupied housing and components of income and wealth. In some cases the prior or subsequent panel data can provide a good estimate. If the family hasn’t changed residence and reported the house value a regional adjustment for house price growth to the last wave report can provide a reasonable estimate of the current value. PSID’s wealth module is also characterized by its pioneering methodological innovation in 1984 of “unfolding brackets,” which collects at least some useful data when respondents refuse or are for some reason unable to answer questions about the exact dollar amount of their asset holdings. For example, respondents are asked for the value of real estate other than their main home: “If you sold all that and paid off any debts on it, how much would you realize on it?” Respondents unable or unwilling to give a dollar figure are first asked: “Would it amount to \$50,000 or more?” If they respond affirmatively, they are asked “\$150,000 or more?”; if negatively, “\$5,000 or more?” The number of such questions and the dollar cutoffs they employ varies from one wealth component to another.

PSID continues to experience very low item non-response rates on wealth and, in fact, very few respondents go through the unfolding brackets sequence, giving instead numeric values (Hurst, Luoh and Stafford, 1998) which are used in generating wealth variables (a component of G3 in Figure 2). Unfolding brackets are now commonly used in most national economic surveys. Using these partial answers are far better than an imputation (G1 in Figure 2) based on a set of predictors in the processing of economic measures, but are they good enough to support panel analysis (Hamermesh, 1992)? Perhaps not. In the housing example a carry forward of the last report of house value with an adjustment may be quite good as an estimates of the current level but will be a very poor measure of house value change in the study of housing wealth effects, for example. So panels create both the opportunity to apply prior year information to reduce the extent of item non-response and provide better alternatives to imputation. Yet an important

use of panel data, analysis of change, is compromised when the change is contaminated with measurement error.

As the number of variables accumulates through time and as various wording changes are needed, a proliferation of codes gives rise to the need to create and manage metadata ('data-about-data') files – files which contain the rapidly accumulating text, codes and other user information about all the variables. Creating and maintaining such files is a central task, and a 'codebooking' software (H2 in Figure 2) is needed to connect and edit the metadata files to the actual variables in a given data collection round or supplement. Which of these variables at different time points or modules have the same or slightly different value ranges and question wordings gives rise to cross-year or cross-module indices (H3 in Figure 2 and Figure 3). These indices identify variables that are fully the same on the same line and those that are nearby, but not quite the same. Users can view the full set of codes and descriptive statistics on-line before deciding if and how to include them in the file for their research project.

The long operation of a complex panel leads to a data archive in the sense that almost no rational user could possibly want the full dataset for a research project. What's there, and when, and where? Most panels have systems that allow users to search through the variables in the growing archive (<http://simba.isr.umich.edu/VS/s.aspx>). The PSID user can search by year and by type of file with the cross-year index. Multiple search modes are the order of the day. Users can search on the questions themselves and related explanation text, the label or name of the variable. The search can be connected by 'or', 'and,' and/or can be a phrase. The variables from an interactive search and selection system are presented on-screen and the selected variables of interest are then added to the subsetted file (or 'data shopping cart' within Figure 2, column)

### The Cross Year Index

*This page works best with Netscape 6+, Opera 7+, and Internet Explorer 5+.*

Click a node label to view variables and codebooks. To add a tree node's variables to your data cart, check the box next to the node then select "Add To Data Cart".

[Expand All](#) | [Collapse All](#) | [Uncheck All](#)

Add Selections To Data Cart

- Individual
  - ACCURACY CODES (indications of assignments)
  - ADDITIONS AND REPAIRS TO DWELLING, WHETHER DID:72
  - AGE:01 99 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80 79 78 77 76 75 74 73 72 71 70 69 68
  - BIRTH WEIGHT:01
  - CHILD DEVELOPMENT SUPPLEMENT
  - CHILDREN
  - CRIMINAL PROBLEMS
  - EDUCATION
  - EMPLOYMENT STATUS:01 99 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81 80 79
  - FAMILY COMPOSITION
  - FERTILITY AND MARITAL HISTORY
  - FOOD STAMPS
  - HEALTH
  - HOURS
  - HOUSEWORK, WEEKLY HOURS:86 85 84 83 81 80 79 78 77 76 74 73 72 71 70 69

**Figure 3. Screen Shot of Mock-Up of Data Center Cross-Year Index Variable Selection**

### Keeping Track of the People, Over Generations, Active and Otherwise

It is now widely appreciated at the design phase: the enormous long-run challenge of keeping track of all the family and individual histories which will to accumulate from this genealogical or steady state design, or even a time delimited panel. Accurate id files are essential for both continuing the panel and exploiting the research opportunities afforded by having measures on individuals with particular relations to one another, either biological or through marriage. These id systems must give each individual a 'primary key' or unique identifier for linking to others and the diverse data elements, such as measures for families at a point in time or across time. For the PSID the primary key is the pair, 1968 family id and person number of the individual in the family tree which is included in the study. The data for a given year can be

subset at the level of the individual or the family. For example one could be interested in panel data on the 2001 - 2005 families' wealth holdings. In the PSID and most other panels (The Survey of Income and Program Participation or SIPP has been an exception), wealth components are collected at the family level. On the other hand employment is recorded for individuals – in the PSID it is employment of the wife and husband in the most detail – as these individuals assume roles within the family. So, to know the wealth holdings of an employed individual the user has individual employment and wealth of the family in which the individual resided for the same year. If the wealth data is for a different year the individual may have lived in a different family, but to know this a accurate map of the individual and the family or residence in different years is essential.

For changes in wealth most researchers would work with data at the family level. But in a panel what about divorce and other family transitions? To help users get started, the online Data Center (Column J in figure 1) has a default definition of the family in a pair of years as the condition that the family has the same head in each year. Of course other wider or narrower definitions of the 'same' family are possible. For research on variables that apply to individuals – such as wage growth, a panel of individuals who report wage income in successive time periods is needed. To create a file at the level of the individual earner a different perspective is needed. This is discussed in more detail in one of our user tutorials – Tutorial #3 to be specific <http://psidonline.isr.umich.edu/Guide/tutorials/tutorial3/balancedpanel.html>.

In Figure 2 there are feedbacks to added data collection in the A, B and C columns. Partly this is driven by the practical needs of making payments and re-contacting respondents. Here too, connections are necessary to maintain accurate records on names and addresses and to update these files. For reluctant respondents a series of individualized persuasion letters is used. Prompt and accurate payment of respondents is also likely to be important. To expedite payments a

specialized software has been developed. This keeps track of the date the payment requests were made, when the payment was sent out. The software helps sort out and avoid the complications which arise when there are different respondents in the same household – as with the Child Development Supplement. There, individual children may receive money of gifts and the caregivers may get paid for the CDS interview as well as for responding to the core survey.

Panel re-interviews can be aided by dependent interviewing. Dependent interviewing is the use of (preloading of) information from the prior data collections to focus and simplify the data to be collected. One form of needed preload is information needed to re-contact that sample, but here we consider information about measures reported from the prior wave. To illustrate, the prior report of occupation and industry may be preloaded into the CATI system. The need for an effective ID system is evident. If the (correctly identified!) respondent confirms that there has been no change in employer or job, the time consuming process of recording and coding occupation and industry (Column D in Figure 2) can be spared.

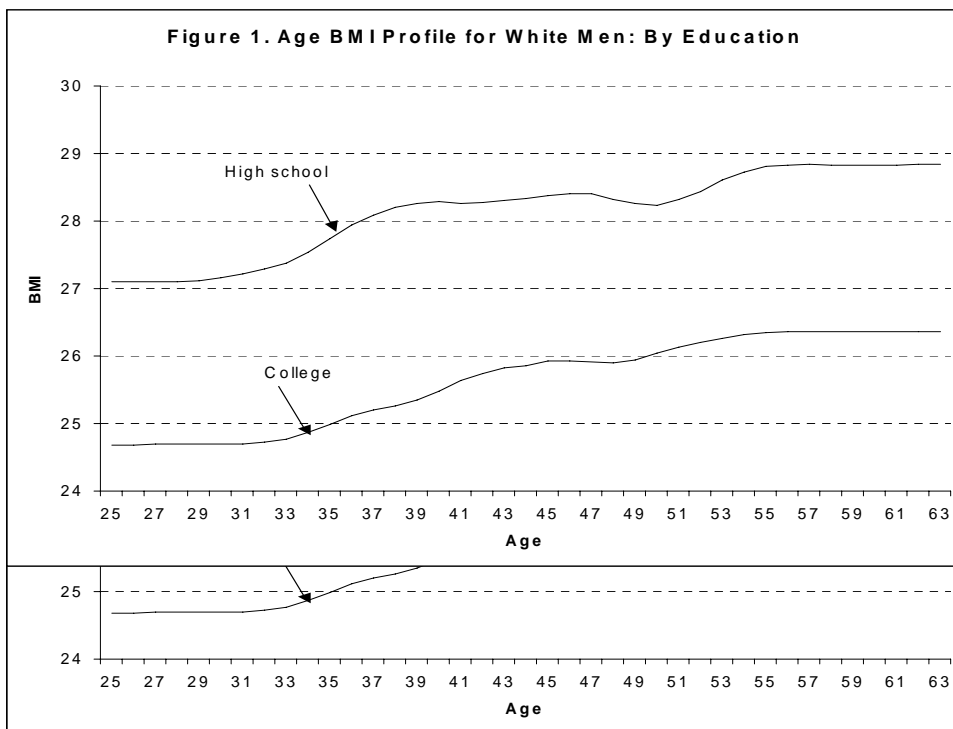
The limitations to dependent interviewing include the possibility that the prior individual respondent or proxy respondent is no longer available – suppose there has been a remarriage. Then the application needs to have a smooth shift to a replacement respondent and a question sequence that will be appropriate in the absence of preload information. If the target of the panel is a specific individual (as distinct from a lineage) then dependent interviewing can work in the sense that there is no need for a system to handle a replacement if the target respondent has left. Other issues with dependent interviewing are possible bias of the respondent in reporting ‘no change’ simply to move the interview along. Since a main purpose of panels is to measure transitions, this is a problem!

Intergenerational analysis represents an important area of research for social scientists. Researchers have discovered that, upon reaching adulthood, many children have outcomes that

are similar to their parents'. For example, existing research on earnings reveals that the elasticity of sons' earnings with respect to fathers'--often called the "intergenerational income elasticity"--is about 0.40 in the United States (Solon 1992, 1999; Lee and Solon 2006). This result implies that, on average, the earnings of a son whose father's income was \$50,000 a year would be expected to be about 40 percent higher than the earnings of someone else whose father earned only \$25,000 a year. Hence, a son's earnings are positively correlated with his parent's, suggesting that high-income parents will have high-income children, even though one might think it more natural to expect earnings to be determined solely by individual characteristics (not by family or class background). Researchers have found similar correlations across generations in other labor market outcomes, such as occupation, union membership, hours worked, and participation in pension plans (Treiman and Robinson 1981; Blanden and Machin 2003); Gouskova, Stafford and Chiteji, 2006). In the areas of health we know that there is a substantial correlation in rates of obesity between children and their parents and grandparents (Kim McGonagle and Stafford, 2001). An intergenerational panel with data search and tools needed to map individuals from one generation with prior or subsequent generations can support such work.

In the PSID the 'mapfiles' have been used in the Child Development Supplement (CDS). Many users have research interests which can be supported by merging measures on the child with measures about the primary caregiver (PCG). To illustrate, to what extent is smoking cigarettes by teens related to smoking by their mother? The primary key (68ID-PN) of the child can be matched to the primary key of the PCG. If the PCG was the head or wife in the years of the CDS then the smoking behavior of the PCG would be reported in his/her role as head or wife. Further if the PCG was in the active panel in 1986 – when smoking behavior was also asked - it is possible to merge in 'youthful smoking' along with contemporaneous smoking of the PCG as it relates to the teen's smoking.

In addition to differences by gender, race and education, data on obesity rates as measured by Body Mass Index (BMI) also have a life cycle signature. The pattern is better represented in a semi-parametric model than the quadratic often used in earnings. Cross-sectional BMI rates are rather flat over the 20's then rise quite steadily to age 40 and have more of a plateau but some rise from age 40 onward as shown in Figure 4 for white males by education level ([Kim, McGonagle and Stafford, 2001](#)). In this case too, the study of intergenerational BMI could benefit from the observation of BMI at specified ages rather than the ages of the different generations at a given time point. In the case of the PSID, BMI was measured in 1986 and as recently as 2005. This allows the researcher to align the ages of the different generations to some extent. Labor earnings have been measured throughout in the PSID, 1968 to present. This allows better life cycle alignment across the generations and, as we will show, this better alignment is of great importance for studying the intergenerational correlation in earnings.



One approach to conducting intergenerational analysis relies on a measure of interest taken from a fixed calendar year for each generation. Another approach constructs measures of a variable of interest incorporating information from many calendar years of data in order to match the different generations at the same life course point – in the case of earnings age 40-50 is often regarded as the life cycle peak – or in principle, with an intergenerational panel of sufficient length, the full lifetime of each generation. In addition, the PSID data provide many sibling pairs for researchers to address the long-standing issue of unobserved heterogeneity. This body of literature helps to understand how early family events, poverty, welfare receipt, and early human capital investment affect achievements in adulthood after controlling for a wide range of observed and unobserved family and neighborhood characteristics (Haveman et al. 1997; Solon et al, 2000).

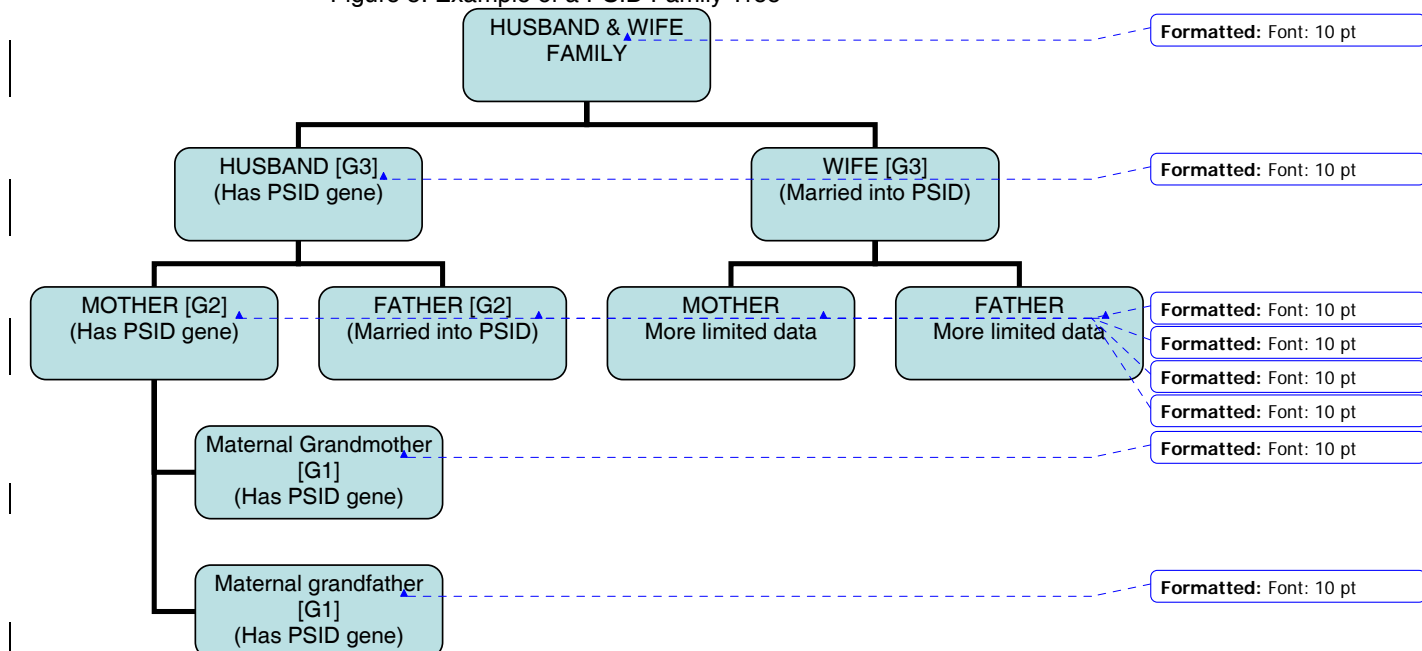
Cross-generational connections in labor earnings serves as a case study in order to highlight essential features of the Panel Study of Income Dynamics (PSID). Tools are needed to construct a file by merging data from two generations when the objective is to see the measures at the *same age or life cycle point*. The smoking example above is simplified by the fact that,

prior to 1999, smoking was measured only in 1986. In the case of earnings it is well-known that earnings peak later in the life cycle. “Since earnings reach a plateau at later ages in the most highly educated groups, both dollar and relative annual earnings differentials among schooling groups grow with age until 45-50, and later still for weekly earnings.” (Mincer, 1974, p. 70). And since labor earnings are a staple feature of the PSID, it becomes possible to measure earnings across generations at a similar life cycle point rather than an arbitrarily chosen year or two. This can be accomplished with the use of the PSID Family Id Mapping System (FIMS) and companion SAS processing programs. The specifics are quite detailed and are set out in a user tutorial <http://psidonline.isr.umich.edu/Guide/tutorials/IG/IG.aspx>.

Research using long intergenerational panels is really just beginning, and there are intriguing issues related to the obvious selectivity which arises, since the intergenerational observations on families with larger numbers of children born earlier in the life course will be overrepresented. Another aspect of intergenerational work is that going back there will be four grandparents and eight great grandparents. As set out in Figure 5 the branching problem is evident.

The point of Figure 5 is to demonstrate that even with the rich genealogical design of the PSID, the cross-generational information is selectively available. In terms of the full set of measures in the PSID adult core, the sample goes down one branch of an adult couple’s genealogy one generation back, and then down a single branch two generations back, and so on. As a result, the extensive set of characteristics of grandparents is oftentimes from a single pair or a single one of four biological grandparents. The impact of long generational processes is limited unless the marriage patterns are of similar individuals. At the same time, the PSID investigators have been mindful of the value of generational information from both sides of the family tree. A

Figure 5. Example of a PSID Family Tree



regular data collection in every wave is an extensive set of background questions for ‘new heads’ and ‘new wives’ who marry a person with the ‘PSID gene’. These questions are part of the interview each wave and include early work history, education, religious preference, experience growing up (urban rural, parental occupation industry and education) and number of brothers and sisters of the new head or wife. Also included is a question about whether the family when growing up was ‘poor’, ‘average’ or ‘pretty well off’, and this has been asked of all new PSID sample, including the original participants as of 1968.

To emphasize, all of the functionalities discussed above -- working with family or individual data, knowing who should be re-interviewed and paying them for their participation, supporting intergenerational work and accuracy of any panel measure of an individual – depend on accurate id files and the capacity to link information on research variables and connect the measures of different families in a research data file.

## **A Proliferation of Data Collection Modes and Designs**

Within a panel a wide array of data collection modes may be applied in different parts of the data collection. Besides traditional paper and pencil and computer assisted modes, many panels have considered and adopted web-based components. Many panels are seeking to link administrative records from health care providers, employers, or government records. These record-based data can be considered a data collection mode. Limits to external records include extensive efforts to ensure confidentiality, accuracy of matching via record based id's, and the lack of design of the records for the intended research purposes.

In the PSID an event history calendar has been designed to collect a two year history of employment activity and residential location. As with web-based modes, EHC's are rather new and experimental in panels. The adoption of EHC methods by the PSID was motivated by the shift to biennial interviewing after 1997. Methodology work indicated that event history calendars (EHC's) could be effective when administered on the phone (Belli Shay and Stafford, 2001) and recover information two years back. An intriguing extension of EHC's would be to combine them with dependent interviewing. There the endpoint of the prior wave would be offered as a starting point for the current data collection. In principle this would build up a lengthy string and the periodicity within that string would be months – or as in the PSID EHC, thirds (beginning, middle and end) of a month. Combined with other information reported on a monthly basis the resulting data could support event history calendar analysis.

Another quite different type of data collected in the PSID is time diaries. These have a long design history (Harvey, 2006) and have been collected for a school day and a weekend day in the CDS. Also, some were collected in personal interviews and others via phone interviews. These data have activity records which are diary based time segments and companion descriptors as the basic unit of analysis. In principle one could think of each of the several hundred thousand

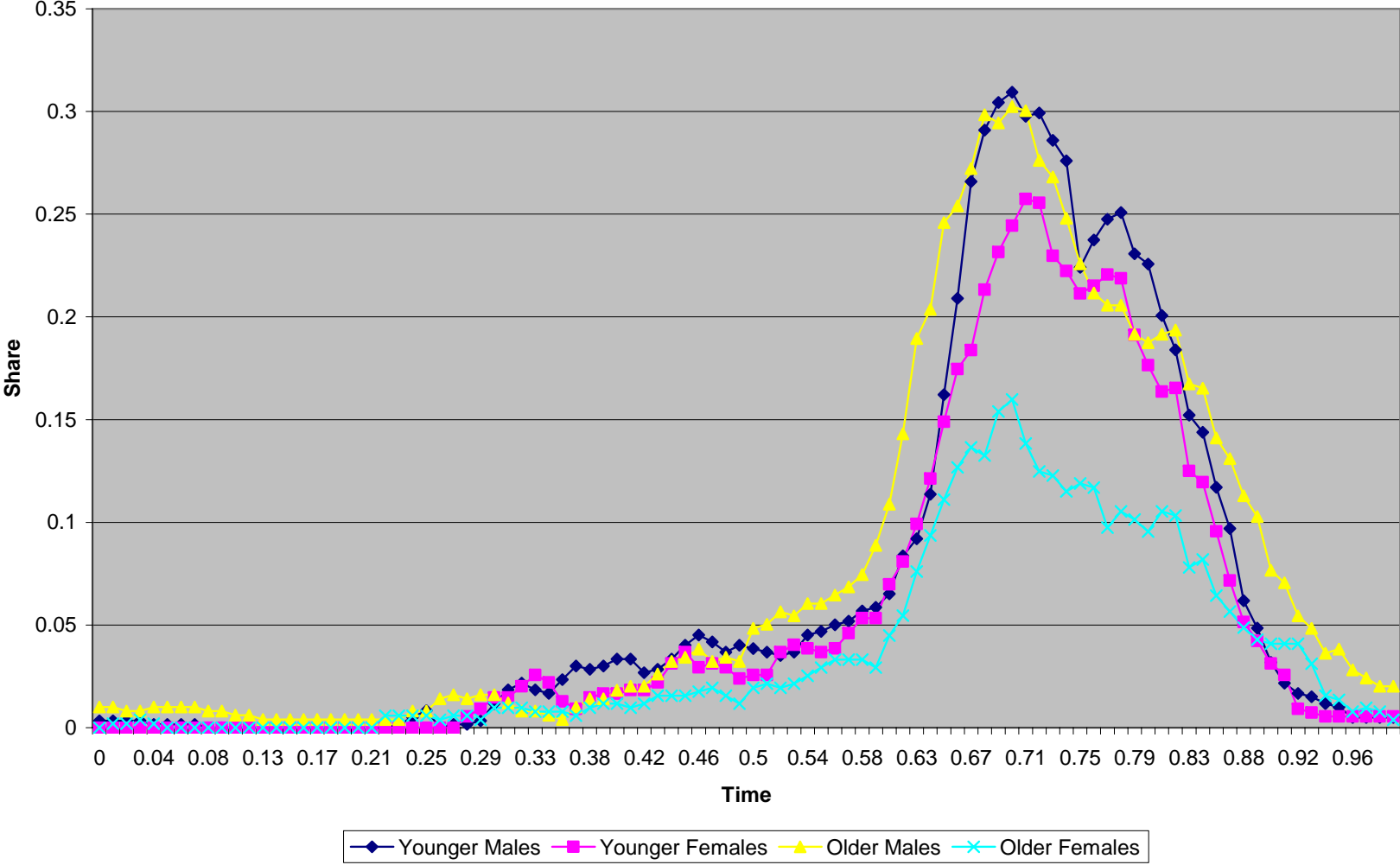
diary episodes in the PSID Data Center as a unit of analysis, via relational data structures one could attach related measures – characteristics of the child, time during the day and so on. So diary activity records provide a nice thought experiment for relational data structures. In principle, one could work with the separate activity record as the unit of analysis and merge in the variables for the individual – and – in principle – the individual’s spouse or parents or even grandparents. In this hypothetical one could ask to what extent is the probability that the primary dairy activity at 10:00 A.M. on a weekday is market work as a function of variables describing the individual at the time of the diary, variables describing the individual in prior years, variables describing the market work of the parents and so on.

EHC’s and time diaries have a common characteristic: they produce very fine grained information. Beyond the basic aggregation into specific time use codes the activity records and the various descriptors leave great, or perhaps excessive, latitude to the different users. This opportunity for reworking the underlying measures in different ways can lead to frustration. To create all or even a good number of the potentially interesting measures becomes time consuming and costly for the project, yet the user may not be in a position to realize the underlying research opportunities. Here we illustrate the point with an organization of time diary records in a temporal fashion, or tempogram of time use by boys’ and girls’ sports activity during a weekday (Figure 4). In a traditional portrayal of the data one would simply have a difference of average sports time from the aggregated files. Here we can see that the difference is occurring ‘after school,’ and is most pronounced in terms of a lower and flatter time pattern for older girls.

While such timeline data could be created from a single cross-section (Michelson, 2005), the attraction of a panel is that a wide array of measures can be brought in to investigate what may cause such patterns. A time diary requires about 20 minutes of respondent time and if two

diaries are collected for each respondent, then 40 minutes is taken up, leaving little room for extensive companion measures of interest in a one time data collection. The 1975-76 National Science Foundation study of time use (Juster and Stafford, 1985) was in fact set up as a panel of four quarterly waves, and this allowed a quite extensive set of non-diary measures needed to create a meaningful set of measures for analysis.

Sports and Active Leisure, by Sex



## Panel Data Within a Panel

A major innovation introduced in the 2003 wave of PSID was the computerized Event History Calendar (EHC) designed for application over the telephone<sup>5</sup>, which provides 2-year long timelines of employment, residence, and features of employment across job transitions. The layout of the EHC from the 2003 CATI application is presented in Figure 5<sup>6</sup>. Having 2-year data in these content areas has helped fill the gap of data caused by moving the study to a biennial data collection. The fine-grained EHC timeline data can be used to support the construction of traditional measures – such as weeks of employment, unemployment, and time out of the labor force. Methodological research has shown that the telephone based EHC interviewing methodology leads to consistently higher quality retrospective reports in comparison to traditional standardized question-asking methods (Belli, Shay and Stafford, 2001; Belli et. al., 2004). In addition, these timeline data can be used to analyze interrelated events such as the timing of auto purchases, residential moves, and employment transitions. In a sense EHC's can be incorporated in a panel to obtain more accurate measures of activities over the re-interview window in total, but can also provide a much shorter periodicity for selected timeline measures of interest – even when it is logistically impossible to re-interview respondents on a very short periodicity. In an initial wave of a panel EHC's can be used to help obtain measures of interest which happened in the past and are seen as important in defining the initial conditions for the respondents in the panel.

In 2007 the PSID is using an early childhood health history. Even though many of the respondents have been in the panel for many years, there are some whose childhood predates

---

<sup>5</sup> The concept behind a telephone based calendar is that a primary cognitive cueing arises across domains and is not dependent on a visual calendar for the respondent. Recalling a residential history places the timing of major employment transitions, for example.

<sup>6</sup> <http://psidonline.isr.umich.edu/Data/documentation/ehc/ehc-demo.html>

their entry into the active panel (as portrayed in Figure 1). Moreover the research value of early health experiences has become more evident, so an early health history calendar can be seen as a way to partially substitute of prospective health measures in a long-lived panel.

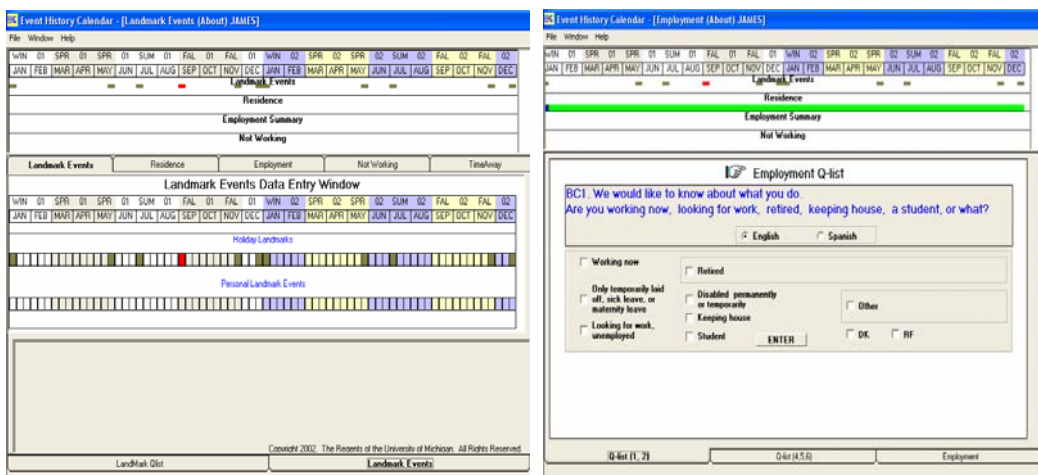


Figure 6.

Example of Employment and Residential CATI-Based Event History Calendar, 2003 PSID

## Concluding Remarks

The data collected in complex panels is to economists, sociologists, and other social scientists what the telescope is to the astronomer. Several specific features of panels allow more precisely the illumination of human behavior in multiple domains, through the measurement of social, economic, and psychological content and a unique sample designs. For PSID these features include longitudinal data collection, a sample that is nationally representative and genealogically-based, content domains that are broad and recurring, and innovative supplements. An area of recent interest is the inclusion of biomarkers – as witnessed in ADHealth, HRS and

PSID. Collection and storage of biological samples (such as blood spots) has several challenges: collection and storage of such measures requires fully informed consent by respondents, may burden response rates, has special storage requirements for samples that are to be assayed at future dates, and requires active interdisciplinary work across biological and social science. Lacking these requirements, and especially a joint research effort the danger is that the research synergies will not be realized and potential discoveries will remain only as possibilities.

Long lived panels cannot describe and determine why individuals and families make the set of complex social and economic decisions that they do. But the panel data does provide the lens through which we observe choices and outcomes. Having a long panel of data improves the precision of the measurement as multiple measures are collected within the same families as well as from multiple family members over a period of many decades. This lens, which becomes more precise over time, supports numerous and expanding scientific advances.

Clearly the history of the large panels is rich: nearly 2,100 published works including 400 dissertations are based on the PSID and there can be an economy of scale to a large data collection. As the panel incorporates additional measures and repeated measures, the types of research which can be supported keeps growing rather than diminishing (Stafford, 1984). At the same time, capitalizing on information technologies, dramatic improvements in the distribution of data and documentation are becoming a reality to support an extended network of scholars. These innovations will shrink the time it takes for new and experienced users to create analytical extracts that meet their highly specialized needs.

Nonetheless, complex panels often give rise to serendipitous findings from long forgotten measures in early waves.

In 1975 the PSID included questions asking if the respondent won enough money to live comfortably would they still keep working. Preliminary investigation suggests that answers to

these questions have power in predicting retirement 25-30 years later. A more recent example is the inverse relation between childhood obesity and consumption of dairy products. The measures of diet were not designed with the hypothesis that dairy products and milk consumption are a substitute for soft drinks and can thereby lead to lower BMI (Bray, Neilsen and Popkin, 2004).