Making Full Use of the Longitudinal Design of the Current Population Survey: Methods for Linking Records Across 16 Months *

Julia A. Rivera Drew
Minnesota Population Center
University of Minnesota

Sarah Flood
Minnesota Population Center
University of Minnesota

John Robert Warren
Department of Sociology
Minnesota Population Center
University of Minnesota

Version: October, 2014

WORKING DRAFT: PLEASE DO NOT CITE OR QUOTE

*Paper prepared for presentation at the 2012 annual meetings of the American Sociological Association, Denver. The research described in this paper was made possible by Grant Number 1R01HD067258 from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD) at the National Institutes of Health. This project also benefitted from support provided by the Minnesota Population Center, which receives core support (5R24HD041023) from NICHD. We thank Steve Ruggles, Trent Alexander, and participants in the Minnesota Population Center's Inequality & Methods Workshop for their guidance and assistance. However, errors and omissions are the responsibility of the authors. Please direct correspondence to Rob Warren, Minnesota Population Center, 50 Willey Hall, 225 – 19th Avenue South, Minneapolis, MN 55455 or email warre046@umn.edu.

Making Full Use of the Longitudinal Design of the Current Population Survey: Methods for Linking Records across 16 Months

ABSTRACT

Data from the Current Population Survey (CPS) are rarely analyzed in a way that takes advantage of the CPS's longitudinal design. This is mainly because of the technical difficulties associated with linking CPS files across months. In this paper, we describe the method we are using to create unique identifiers for all CPS person and household records from 1989 onward. These identifiers—available along with CPS basic and supplemental data as part of the on-line Integrated Public Use Microdata Series (IPUMS)—make it dramatically easier to use CPS data for longitudinal research across any number of substantive domains. To facilitate the use of these new longitudinal IPUMS-CPS data, we also outline seven different ways that researchers may choose to link CPS person records across months, and we describe the sample sizes and sample retention rates associated with these seven designs. Finally, we discuss a number of unique methodological challenges that researchers will confront when analyzing data from linked CPS files.

Making Full Use of the Longitudinal Design of the Current Population Survey: Methods for Linking Records across 16 Months

1. Introduction

The Current Population Survey (CPS) is one of the most widely used data resources in social and economic research. For example, between 2000 and 2013, there were 290 articles that used or cited CPS data in the *Journal of Political Economy*, the *American Sociological Review*, and *Demography*, the leading journals of economics, sociology, and demography, respectively.[1] The reasons for this popularity are simple: The CPS offers a long series of surveys of nationally representative samples of household-based individuals, with large sample sizes, high response rates, and expansive subject coverage.

Since July of 1953, members of each housing unit included in the CPS have been interviewed eight times over a sixteen-month period (U.S. Bureau of Labor Statistics 2006). Despite this longitudinal design, researchers have almost exclusively analyzed CPS data as though it were a cross-sectional survey.[2] There are several reasons for this: CPS records are technically difficult to link across surveys (especially for older files); the CPS's complex sampling design complicates longitudinal analyses; identifying sequences of files containing variables relevant to

[1] These figures are based on a Google Scholar search on September 28, 2014.

[2] Indeed, there is some tendency to deny the CPS its place as a longitudinal survey. In 2002, for example, Burkhauser et al. (2002: 543) wrote, "[a]lthough the CPS is a cross-sectional survey, it does interview respondents over the course of a year." Likewise, O'Connell and Rogers (1983: 369) noted that, "[t]he CPS data do not provide for an analysis of a continuous longitudinal panel of respondents."

a research problem can be laborious; the integration of variables over time is challenging; and data access is awkward, requiring the manipulation of many different files.

The Minnesota Population Center at the University of Minnesota is currently adding extensive new collections of basic and supplemental CPS data to its widely used Integrated Public Use Microdata Series (IPUMS). The IPUMS-CPS data will be fully linked—that is, users will be able to extract longitudinal data on households and individuals from basic and supplemental surveys across as many as 16 months. All measures will be fully integrated and harmonized over time; appropriate longitudinal weights will be available; and relevant metadata and documentation will be provided.

We have three objectives in this paper. First, and most importantly, we describe our techniques for linking all CPS person- and household-level records over time from 1989 onward. This step—which involves the creation of new and unique household and person level identifiers for every CPS record, named CPSID and CPSIDP, respectively—will make longitudinal analysis of CPS data dramatically easier going forward. Consequently, it is important to document our methods for creating these linking keys. Second, we demonstrate several possible research designs based on longitudinally linked CPS records on people. Researchers who have made use of the longitudinal design of the CPS have generally only linked records in a limited number of ways (usually matching records across March supplements); we hope to inspire innovative new research by demonstrating seven different research designs based on linked CPS person-level data. Third, we provide information about the sample sizes and retention rates that researchers can expect when they implement one of these seven research designs based on linked person-level CPS data. That is, for seven research designs likely to be used most frequently by

researchers, we describe how many people those analysts can expect to include in their longitudinal analyses and how much panel attrition they can expect to observe. This information, which previously required considerable effort to obtain, is crucial for researchers seeking to design new longitudinal analyses of CPS data.

2. Overview of the Design of the CPS

The CPS is a monthly U.S. household survey conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistics (BLS). Initiated in the 1940s in the wake of the Great Depression, the survey was initially designed to measure unemployment. A battery of labor force and demographic questions, known as the "basic monthly survey," is asked every month. Over time, supplemental surveys on special topics (e.g., school enrollment, food security) have been added. Among these, the March Annual Social and Economic (ASEC) Supplement—formerly referred to as the Annual Demographic File—is the most widely used by researchers and policymakers. Although some topical supplements are conducted in the same month each year (e.g., the school enrollment supplement has appeared in October since 1968), others have been conducted in different calendar months in different years (e.g., beginning in 2002, the food security supplement has appeared in December, but before that it appeared in April in some years and September in other years).

The CPS sample is representative of the civilian, household-based population of the United States. In recent years, each monthly CPS has included about 140,000 individuals living in about 70,000 households. Upon selection into the CPS sample, household members are surveyed in four consecutive months, left un-enumerated during the subsequent eight months, and then resurveyed in each of another four consecutive months; new rotation groups are brought into

the CPS sample each calendar month. The CPS 4-8-4 rotating panel design guarantees that in any calendar month, about one-eighth of the sample is in its first month of enumeration (month-in-sample 1, or MIS 1), about one-eighth is in its second month (month-in-sample 2, or MIS 2), and so forth.

Table 1 further describes the CPS rotation group design. For each of 16 consecutive months between January of Year X and April of Year X+1, and separately by month in sample (MIS), the table shows the calendar month in which CPS participants first entered the survey. For example, participants in MIS8 in January of Year X first entered the CPS in October of Year X-2. As per Table 1, CPS participants in January of Year X may have begun the CPS in October, November, or December of Year X-2, in January, October, November, or December of Year X-1, or in January of Year X. The shaded boxes in Table 1 represent the calendar months in which participants are in MIS1 through MIS8 among those who first began the CPS in January of Year X. One logical result of this rotation group design is combinations of calendar months for which no longitudinal linkages are possible. For example, no CPS participants are surveyed in both June and October; by design, researchers wishing to link records from the June (immigration) supplement to the October (school enrollment) supplement can never do so.

Of course, in any given month some households and/or individuals within households may refuse or be unavailable to be surveyed; the BLS has generally made sustained efforts to bring non-respondents back into the sample in subsequent survey waves, but this form of non-response complicates longitudinal analysis of CPS data. Furthermore, because the CPS selects a sample of *households*, researchers studying individual *people* must use the data with care. New people can be added to households after MIS1 (e.g., new babies can be born) and people can

leave households prior to MIS8 (e.g., through death, divorce, or migration). More importantly, if the occupants of a residence move out, they are replaced in the sample by the new people who move in. The prior occupants of the residence are no longer included in the CPS. The BLS provides cross-sectional sampling weights for use with the basic monthly and supplemental survey data. Longitudinal weights, which are available only for adults linked between two adjacent samples and are intended for gross flows analysis, are also provided on monthly files from 1989 forward. IPUMS-CPS will provide longitudinal weights appropriate for month-to-month as well as other types of analyses.

Many of the survey items included in the basic monthly survey and of some of the supplemental surveys have remained essentially constant over time. It is possible, for example, to construct long time series of consistent measures of labor force status from the basic monthly surveys and of wage and salary income from the ASEC. However, in many other cases the topics covered by CPS surveys, the way that focal concepts are measured, and/or the universe of individuals who are asked focal questions change over time. The harmonization and integration of measures as part of the IPUMS-CPS collection will save researchers time and effort, but these issues complicate longitudinal analyses. Researchers studying within-person change over time should be aware of changes over time in how questions are asked and in who is asked which questions. Because the BLS frequently imputes missing values that arise from item non-response, researchers conducting longitudinal analyses should also be careful in how they handle imputed values when studying change across surveys.

3. Methods for Creating Unique Household- and Person-Level Identifiers

Despite the long-standing longitudinal design of the CPS and the availability of household- and person-level identifiers on the public-use data, linking CPS records across months is deceptively difficult. Various complications and sources of error make the process more difficult than simple numeric matching based on identifiers, even in the most recent CPS samples. With little guidance from CPS documentation, researchers who want to link records must be aware of many details that complicate the linking process. Among them: The 4-8-4 design constrains the portion of the sample that can be linked in adjacent months and in consecutive years. For several years of CPS data, the household identifiers (which constitute the most obvious basis for record linkage) are not unique across households. Linking is further complicated by changes in the composition of housing units due to migration and mortality, household- and person-level non-response, and data recording errors.

Data from 1962 to 1978 present the most serious linkage challenges. Each housing unit was assigned a unique identifier during most (but not all) years in this period, but person-level identifiers do not reliably identify the same individual in multiple samples. Since the CPS follows housing units from month to month—rather than a particular group of people—researchers must use individuals' demographic characteristics to link people within households over time. Furthermore, because of changes in the numbering scheme for housing units, household-level identifiers cannot be used to link housing units between 1962 and 1963, 1971 and 1972, 1972 and 1973, and 1976 and 1977 (Kelly 1973; Madrian and Lefgren 2000).

In contrast to the earlier samples, data from 1979 to the present contain housing unit identifiers and person identifiers that are (mostly) unique over time and thus useful for

longitudinal linkage. Since 1994, however, housing unit identifiers in the CPS have been re-used once a housing unit has left the CPS after its first four months in sample (Feng 2001). Similarly, many housing units have duplicate identification numbers in the ASEC files from 2001 through 2004 because of the State Children's Health Insurance Program (SCHIP) expansion. The ASEC achieved a sample expansion by administering the March questionnaire to persons in housing units from surrounding months who would otherwise have not received that supplement; these SCHIP expansion cases sometimes have the same housing unit identifiers as "true" March cases. It is possible to distinguish the "true" March cases from the expansion cases and to assign new and unique household identification numbers for linking, but the task is laborious, requiring users to merge the March basic monthly survey and the ASEC files, which poses another layer of complexity. Finally, as is true in earlier years, changes in numbering schemes for housing units prevent linking based on household identification numbers across some pairs of years, including 1984 to 1985, 1985 to 1986, 1994 to 1995, and 1995 to 1996. Furthermore, changes in the identifier schemes requires tedious manipulation to create identifiers that are compatible over time as is the case when linking May 2004 and later data to earlier months.

Beyond all of this, and even in years in which household- and person-level identifiers are available and useful for linking, most researchers "confirm" their links using demographic information from the linked surveys. That is, they compare the age, sex, race/ethnicity, and other attributes of apparently linked people, and the geography and composition of apparently linked households. Because of migration, mortality, non-response, and recording errors, linkages based solely on housing unit and individual identifiers sometimes result in erroneous links or missed links, even in the most recent samples.

Researchers making or confirming linkages based on demographic and other information typically encounter several obstacles. Demographic variables useful for verifying linked individual-level records are coded differently over time. Race codes were expanded in January 2003 from four to twenty-one categories; the implication is that researchers using race to validate matches between months must bridge the changes in race codes as an additional procedural step. In addition, the ASEC variables are named and sometimes coded in ways that differ from the surrounding months. For IPUMS-CPS, these issues of nonstandard variables across time and supplements will be overcome through data integration and harmonization. More fundamentally, there is no one set of characteristics that researchers agree should be used to check the quality of person-level links over time within housing units, and no consensus on the acceptability of error rates. For instance, Madrian and Lefgren (2000) propose linking individuals within a given housing unit based on sex, race, and age (allowing a tolerance of two years from the expected age). Others use scoring matrices to identify "good" matches across time (Katz, Teuter and Sidel 1984; Pitts 1988). Feng (2001; 2008) suggests using additional variables paired with a Bayesian approach, which minimizes discarded matches and is more forgiving of recording errors.

With these and other issues in mind, we have developed robust linking algorithms that build on the work of Madrian and Lefgren (2000), Feng (2001; 2008), and others. Our algorithms create new household- and person-level identifiers (CPSID and CPSIDP, respectively) that are unique over time. The first month that a household or person is observed in any CPS data file, a new value of CPSID or CPSIDP is created; that value is then assigned to that household or person each time they subsequently appear in the CPS. CPSID and CPSIDP facilitate *mechanical* matches

of households and individuals over time. Extensions to CPSID may include characteristic-based matching and probabilistic matching, though the latter are not the focus here.

The values of CPSID and CPSIDP are based on a combination of four pieces of information: YEAR, MONTH, HHNUM and PNUM. YEAR is a four-digit number that indicates the year in which a household or person appears in the CPS. Likewise, MONTH is a two-digit variable that indicates the month in which a household or person appears in the CPS. These variables come directly from the CPS data. HHNUM and PNUM are created by us during the IPUMS-CPS ingest process. All household and person records are assigned either a household number (HHNUM) or a person number (PNUM) that is unique within a given month but not across months. Household numbers begin at one and increment by one until the last household is numbered. Similarly, every person is assigned a person number that in each household begins at one and increments by one and is thereby unique within households (but not across them, or across months). Household records are assigned a PNUM value of zero; each person with a household shares the same value of HHNUM.

The values of CPSID and CPSIDP in any focal month are assigned in one of four ways, based on the month in sample (MIS) value for that household or person. First, we assign households and persons in MIS1 new values of CPSID and CPSIDP that concatenate YEAR, MONTH, HHNUM, and PNUM. Second, for households and persons in MIS2 through MIS8, we use the original CPS household and person identifiers (State FIPS code, HRHHID, HRHHID2 and, for person records,

PULINENO[3]) to locate records for the household or person in the month in which they should have been in MIS1. If the household or person is not located in the file for the month in which

[3]HRHHID and HRHHID2 are a two-part household identifier that, in combination, is theoretically unique across samples. However, as we note above, there are duplicate household identifiers for several different reasons. Combining the household identifiers with state codes minimizes, but does not eliminate, this problem. HRHHID2 is a variable that CPS makes available on its public use files beginning in the May 2004 sample. Although HRHHID2 was not offered on the CPS public use files prior to May 2004, IPUMS-CPS creates it back to January of 1994 by drawing on three variables (HRSAMPLE, HRSERSUF, and HUHHNUM). The creation of the five-digit HRHHID2 requires transformation and concatenation of pieces of three component variables as follows. Extract the numeric component (second and third digits) of the four-digit alphanumeric variable HRSAMPLE; these become the first two digits of HRHHID2. Convert HRSERSUF from alphabetic to numeric where the letter corresponds to the order in the alphabet (A=01, B=02, etc.). HUHHNUM, trimmed of any leading zeros, is the final digit of HRHHID2. Once these three variables are prepared, the user should concatenate to create HRHHID2 the extracted numeric piece of HRSAMPLE, the alphabetic characters from HRSERSUF converted to numeric, and HUHHNUM. For example, consider the following original values of HRSAMPLE (A72B), HRSERSUF (A), and HUHHNUM (1). To create HRHHID2=72011, use '72' from HRSAMPLE, convert 'A' in HRSERSUF to '01', and use '1' from HUHHNUM. Prior to January 1994, neither HRHHID2 nor the component pieces used to construct it were available; therefore, we use State FIPS, HRHHID, and PULINENO to link records between 1989 and 1993.

they should have appeared in MIS1 (perhaps because of non-response or migration), we attempt to locate corresponding records in the month in which they should have been in MIS2, and so on until we reach the focal month. If we locate records for the household or person in a month prior to the focal month, we use the value of CPSID and CPSIDP from that earlier month and assign it to the household or person in the focal month. Third, if we locate a household record but not a person record in a month prior to the focal month (perhaps because a new person entered the household), we assign the person a value of CPSIDP that is the next available value of CPSIDP within that household. Fourth, if we locate records for neither the household nor the person in months prior to the focal month, we create new values of CPSID and CPSIDP that concatenates the record's values of YEAR, MONTH, HHNUM, and PNUM during the focal month and year when we first observe them. The values of CPSID and CPSIDP are always conditional on the original household- and person-level identifiers and the logic of the CPS rotation pattern.

Available as part of IPUMS-CPS (https://cps.ipums.org/cps/), these new linking keys (CPSID and CPSIDP) greatly simplify longitudinal use of CPS files. CPSID and CPSIDP will automatically handle major issues that used to make large-scale linking projects less feasible—issues like understanding the rotation pattern well enough to know which records should be linked across which months, how to handle recycled identifiers, how to use geographic information to uniquely identify households, and how to link records that bridge changes to the logic and design of BLS-provided identifiers. Researchers will no longer be required to devote significant time or resources to link records on their own. Because CPSID and CPSIDP build on the "best-practices" linking procedures described by Madrian and Lefgren (2000), Feng (2001; 2008), and others (e.g., Nekarda 2009), researchers will be less likely to introduce errors when

linking on their own. This will increase the accuracy and comparability of substantive CPS-based research in the years ahead.

4. Research Designs Based on Linked Person-Level CPS Data

In this section, we demonstrate seven longitudinal research designs based on longitudinally linked CPS *person*-level records; all can be implemented easily using CPSIDP as described above. In our opinion, the technical difficulties associated with creating linked CPS records have precluded creative uses of those data. Our goal in this section is to facilitate and motivate innovative new research by demonstrating ways that CPS person records might profitably be linked. In each case, we provide substantive examples of the sorts of projects that might be made possible. Our focus on linked CPS *person* records—as opposed to *household* records—is pragmatic. We anticipate that most readers will be interested in research on individuals.

In the tables below, we provide the un-weighted sample sizes and retention rates that researchers can expect to achieve for each of seven research designs; those estimates are derived from linked CPS basic monthly survey person records collected in 1994-1995 and 2009-2010.[4] We provide sample sizes and retention rates before and after omitting linked records that differ with respect to sex, race, or age.[5] We hope that researchers will use this information as a basis

[4]We provide information for only these years because of space constraints. Except where noted, results for years in the interim are similar to those for 1994-1995 and 2009-2010.

[5] When considering the characteristics of linked people, we declare a "mismatch" when sex or race differs or when age declines or increases by more than one year (except among people age

for designing and establishing the feasibility of new research projects. Note, however, that while CPSID and CPSIDP may be used to link supplements, our figures below do not account for CPS supplement non-response; supplement nonresponse rates tend to be somewhat higher than for the basic monthly surveys and therefore linkage rates and numbers of linked records will be lower).

To begin, Table 2 reports the number of people responding to the CPS basic monthly survey, by MIS group, for each calendar month between January 1994 and April 1995 and between January 2009 and April 2010. We selected 1994-1995 and 2009-2010 for demonstration purposes, and in both cases we show January 1994/2009 through April 1995/2010 to show the full progression of people who began the CPS in MIS1 in January of 1994/2009 through the 4-8-4 CPS rotation pattern. Each individual in MIS1 in January 1994/2009 completes their eighth month of participation in the CPS in April of the following year. In general, between 135,000 and 140,000 people respond to the CPS each month. There are typically about 17,000 people in each MIS group in each month. None of these numbers has changed appreciably since the early 1990s.

4.1. Linking Across Two Consecutive Months

The simplest longitudinal research design based on CPS person records involves observing people in just two consecutive calendar months. This application could be powerful for modeling change in labor force, family, and educational statuses—that is, in things observed in each basic monthly survey. Indeed, some labor economists use the CPS in this way for gross-flow analyses of labor

80 or older in 2009-2010, where we allow for an age mismatch of five or fewer years to accommodate top-coding of the age variables).

force transitions (e.g., Frazis et al. 2005). Given the large CPS sample size, many respondents can be expected to transition from "employed" to "unemployed" or from "married" to "separated" (for example). What is more, this design could be used to combine measures collected in adjacent topical supplements (e.g., the October school enrollment supplement and the November voting and registration supplement), or to model outcomes observed in a topical supplement as a function of changes in labor force, family, or educational statuses across the two months.

How many CPS (person) respondents can be linked from one month to the next? As shown in Table 3, people in MIS1-3 or MIS5-7 in January of Year X (the shaded cells of the table) may also be observed in MIS2-4 or MIS6-8 in February of Year X (the outlined cells); the same would be true of linkages between February and March, March and April, and so on. That is, when linking consecutive calendar months, by design, researchers should only expect to retain about 75% of all CPS respondents because respondents in MIS4 and MIS8 are not observed in the next calendar month due to the 4-8-4 rotation pattern.

Using CPSIDP, how often are we able to link people across consecutive calendar months? In the middle and right columns of Table 3, we report results for January to February linkages in 1994 and 2009; results for other calendar months and intervening years are similar. Of the 106,022 people in MIS1-3 or MIS5-7 in January of 1994, we are able to mechanically link to 101,789 (or 96.0%) of them in February of 1994; of these, 100,781 match on age, sex, and race. Similarly, of the 100,750 people in January 2009 who were eligible to participate in the CPS in February of that year, we are able to link records for 96,822 (or 96.1%) of them; 96,049 also match on personal demographic attributes. In both examples, less than 1% of mechanically

linked records are mismatched on age, sex, and/or race. In general, across any two consecutive basic monthly surveys back through at least 1994, researchers can expect about a 5% attrition rate (among those eligible to be surveyed in both months) and between 95,000 and 100,000 linked person records.

4.2. Linking Across Two Non-Consecutive Months

The design above makes sense for labor force, family, or educational outcomes that might be expected to change from one month to the next or for situations in which researchers wish to link records from topical supplements that are administered in adjacent months. In some instances, researchers may wish to allow for more than one month between surveys and/or to link topical supplements that are not administered in adjacent months. How does employment, family, or other status change across three-month windows of time? What are the relationships between educational experiences (observed in the October school enrollment supplement) and patterns of food insecurity (recently observed in the December food security supplements)? Each of these questions involves linking CPS person records across two non-consecutive months.[6]

How many CPS (person) respondents can be linked across non-consecutive calendar months? The answer will depend on the length of time between surveys; for demonstration purposes, we report results for people observed in both October and December. As shown in Table 4, people in MIS1-2 or MIS5-6 in October of Year X (the shaded cells of the table) may also

[6] Of course, researchers would do well to think carefully about the linkages that are not possible based on the CPS rotation group design depicted, for instance, in Table 1. As noted above, for example, no CPS participants are surveyed in both June and October.

be observed in MIS3-4 or MIS7-8 in December of Year X (the outlined cells); the same would be true of linkages between January and March, February and April, and so on. When linking records across a two-month window of time, by design, researchers should only expect to retain about 50% of all CPS respondents.

How often can we link people from (for example) October to December using CPSIDP? In Table 4, we report results for such linkages in 1994 and 2009; again, results for other calendar months and intervening years are similar. Of the 69,650 people in MIS1-2 or MIS5-6 in October of 1994, we link 65,479 (or 94.0%) of them to December of 1994. Similarly, of the 66,435 people in October 2009 who were eligible to participate in the CPS in December of that year, we are able to link records for 62,529 (or 94.1%) of them. As before, a small number of linked records—about 1% of the total—do not match on sex, race, or age.

4.3. Linking to the Same Calendar Month across Two Consecutive Years

The vast majority of research that makes any use of the longitudinal design of the CPS links person records across consecutive years of the ASEC. Most published examples of linked ASEC records feature analyses of earnings dynamics (e.g., Cameron and Tracy 1998; Celik et al. 2012), but linked ASEC records have also been used to study topics like geographic mobility (Geist and McManus 2008; Geist and McManus 2012), trends in the prevalence and correlates of post-retirement employment (Pleau and Shauman Forthcoming), selective emigration of the foreign-born population (Van Hook and Zhang 2011), and movement into and out of labor unions (Zullo 2012). We suggest that linked IPUMS-CPS records will facilitate a new generation of research that considers year-to-year changes in respondent attributes as ascertained in basic monthly surveys and on any number of topical supplements.

Using CPSIDP as described above, how often is a person who is observed in one March CPS (for example) also observed in the following March CPS? As shown in Table 5, only people in MIS1-4 in March of Year X (the shaded cells of the table) may also be observed in MIS5-8 in March of Year X+1 (the outlined cells). In the middle and right columns of Table 5, we report results for March-to-March linkages from 1994 to 1995 and from 2009 to 2010; results for intervening years are similar. Of the 69,409 people in MIS1-4 in March of 1994, we are able to link to 48,140 (or 69.4%) of them in March of 1995; these rates are similar to those reported by Madrian and Lefgren (2000) for the 1980s and 1990s. Similarly, of the 67,884 people in March 2009 who were eligible to participate in the CPS in March of the following year, we are able to link records for 53,486 (or 78.8%) of them; these rates are similar to those reported by Feng (2008) for the 2000s. In recent years, when linking CPS person records from one year to the next, researchers can expect about a 20% attrition rate (among those eligible to be surveyed in both Marches) and between 50,000 and 55,000 linked person records. Again, a small number of apparent links involve records that do not match on sex, race, and/or age. Results for non-March months are similar.

4.4. Linking As Many As Eight Consecutive Records for Single Cohorts of People

Incoming cohorts of CPS respondents in MIS1 may never again appear in MIS2-8, or they may appear in the CPS on as many as seven more occasions. Person-level records with key social and economic attributes as measured at regular intervals over a series of months for a large and representative sample of Americans would seem to be a powerful and underutilized resource for any number of research purposes. The fact that the CPS rotation group design has been in place since 1953—and thus that new time series of person-level attributes are observed for groups of

people beginning the CPS *every month* across *decades*—would seem to multiply the possibilities. How do short-term employment dynamics vary as a function of people's educational attainments and demographic characteristics? How have these relationships changed across decades as labor market and other institutional processes have been transformed? How do short-term labor market responses to the birth of children differ for men and women, and how have these relationships changed over the years as women's labor market opportunities have increased? Questions like these—about long-term trends in short-term processes—can be addressed as never before using a half-century of CPS person records linked over as many as eight surveys.

How often do incoming CPS respondents participate in their first two months in sample? Or, in their first four? How often do they participate in all eight surveys? The top left cell in Table 6 reports the number of people first responding to the CPS in MIS1 in January 1994 (the top panel) and in January 2009 (the bottom panel). The next rows—for February through April of those years and then for January through April of the following years—each report the cumulative number and percentage of those respondents who responded to *every* CPS survey for which they were eligible up through that month's survey. For example, of the 16,942 people first responding in MIS1 in January of 2009, Table 6 shows that 15,142 (or 89.4%) of them responded to every subsequent survey through April of 2009 and that 11,528 (or 68.0%) of them responded to all eight surveys through April of 2010. These numbers dip slightly when we exclude records that do not match on respondents' demographic attributes. Table 6 makes clear that in both years, attrition from the panel is greatest between MIS4 and MIS5.

For each incoming cohort of CPS respondents, about 15,000 (or 90%) participate in the first four CPS surveys for which they are eligible. In recent years, more than 11,000 (or about

two-thirds) of respondents participate in all eight CPS surveys; somewhat fewer respondents were as cooperative in earlier years. Of course, these are the most stringent criteria for sample selection that researchers might employ. For any particular application, researchers might be willing (for example) to select people who responded to three of the first four or seven of the first eight surveys; as a result, sample sizes would be higher (and attrition rates lower).

4.5. Linking People in MIS1 to Any Subsequent Survey

For some applications, researchers may simply need to observe CPS respondents more than once. They may not be particularly concerned about whether all respondents are observed in the same calendar months, as long as they are observed at least twice over time. For example, research on the correlates of job loss or union formation may simply require, at least at the outset, multiple observations per respondent.

This design can be implemented in a number of ways, and so we have selected just one of them for expository purposes. In Table 7, we report the un-weighted number and percentage of CPS respondents who are observed at least once in MIS2 through MIS8 among those who first responded in MIS1 in January of 1994 (the top panel) or January of 2009 (the bottom panel). A different design might stipulate that respondents participate in any two surveys in MIS1 through MIS8, regardless of whether they respond in MIS1.

As shown in Table 7, about 16,500 (or about 98% of) CPS person records for people in MIS1 can be linked to *any* subsequent record in MIS2 through MIS8. This should not be surprising since we showed in Tables 3 and 6 that among those eligible to do so, about 95% of people who respond in one month also respond the next month. That is, if a researcher is simply interested

in selecting respondents who are interviewed in MIS1 and then in at least one subsequent survey, they can expect about 16,500 person records and very little panel attrition.

4.6. Linking People in MIS1 to Any Survey in the Subsequent Year

A variant of the design above is to select cases in which CPS respondents participate in at least two surveys that are *separated by at least a year*. This design would facilitate research on relatively rare events like involuntary job loss or the death of spouses, which may not be observed at sufficiently high rates when surveys are separated by shorter time intervals.

In Table 8, we report the un-weighted number and percentage of CPS respondents who are observed at least once in MIS5 through MIS8 among those who first responded in MIS1 in January of 1994 (the top panel) or January of 2009 (the bottom panel). That is, these respondents were first observed in January of 1994 or 2009, and were then observed at least once at some point 12 to 15 months later.

As shown in Table 8, about 12,000 CPS person records for people in MIS1 in January of 1994 can be linked to *any* subsequent person record in MIS5 through MIS8 in 1995. About 13,000 person records for people in MIS1 in 2009 can be linked across a year or more. Attrition rates are higher for this design—as compared to the one above, which did not require that respondents be observed across a full year—because of the attrition during the eight-month gap between MIS4 and MIS5. Note that the linkage rates across one year or more are higher than linkage rates across exactly one year (Table 5) because some of the people who do not respond in MIS5 subsequently do respond in MIS6-8. In general, however, each month more than 12,000 begin the CPS who will eventually be observed across at least a full year.

4.7. Linking across MIS4 and MIS5

As noted above, the greatest rate of attrition from the CPS occurs between MIS4 and MIS5. This is unfortunate, since a powerful research design for some purposes would be to select people who responded in MIS 4 and who then responded nine months later in MIS5. Especially for events so rare that they might not be frequently observed across a single month even in a large sample (e.g., the birth of children, the death of a spouse, involuntary job loss), the extended time interval between MIS4 and MIS5 might be advantageous. We know of no published research that has utilized this design.

How many CPS person records can be linked from MIS4 to MIS5? As shown in the top third of Table 9, people in MIS4 in January through July of Year X (the shaded cells of the table) are observed in MIS5 in October of Year X through April of Year X+1 (the outlined cells). Using CPSIDP, how often are we able to link people across MIS4 and MIS5? In the middle and bottom panels of Table 9, we report results for people in MIS4 in January through July of 1994 and 2009, respectively. Of the 17,500 or so people in MIS4 in each month of 1994, we are able to link to about 12,500 of them nine months later in MIS5. Similarly, of the 17,000 or so people in MIS4 in each month of 2009, we are able to link to about 13,500 of them in MIS5. In recent years, researchers employing this design can expect about a 20% attrition rate across the nine months between MIS4 and MIS5 and about 13,500 linked records.

5. Challenges Associated with the Analysis of Longitudinal CPS Data on People

Fully linked CPS person and household records—along with integrated and harmonized measures—should vastly increase the research uses of this data resource. Researchers using longitudinal designs like those outlined above—or others that we have not thought of—will be able to ask new questions using CPS data and with considerably greater ease. However, with the

great opportunities that the enhanced IPUMS-CPS data offer come new challenges. In this section, we discuss a number of new issues that may arise as scholars transition from thinking about the CPS as mainly a series of cross-sectional surveys to thinking about its full potential as a longitudinal survey.

First, because most researchers have thought about the CPS primarily in cross-sectional terms—or, at most, thought about linking years of ASEC data—those researchers will need to rethink what is possible. To that end, the new IPUMS-CPS data dissemination website will feature a number of data discovery tools that will overview the content of various supplements and make clear what sorts of linkages are possible. However, it will be incumbent on the research community to think creatively as it views the CPS with fresh eyes. With fully linked records, supplements can be linked to other supplements (subject to the constraints described above) and to basic monthly data in novel and potentially fruitful ways. Down the road, we imagine that the capacity to easily link records may inform decisions about when to field topical supplements to maximize their utility for research and policymaking.

Second, because it will be easy to link records over time and for people entering the CPS across many years, researchers will face new challenges associated with the consistency of measures. In many cases, the way that key questions are asked has changed over time. Even when question wording has remained the same, the universe of respondents who are asked certain questions has changed over time; indeed the longitudinal dimension of the CPS means that people can age into or out of the universe of people who are asked certain survey items. Some items that appear in every monthly CPS data file are not actually asked every calendar month. In the case of educational attainment, for example, respondents are only asked questions

about that concept in February, July, October, or in MIS1 and MIS5; data in all other months are carried forward from earlier surveys. These sorts of measurement complications will be documented in the new IPUMS-CPS data dissemination website, but it will be important for researchers to understand them going forward.

Third, for many applications, the fact that the CPS samples households (and not people) will pose new challenges as researchers design longitudinal projects. For example, it would seem possible to use linked CPS records to study the impact of marital disruptions on women's labor force participation (and perhaps how that varies over time and geography). However, one consequence of marital disruption is that the parties involved often move out of their residences. This mobility, which represents a non-ignorable form of selective attrition of people from the CPS, might plague such a project. In general, researchers will have to think carefully about how their research design may be impacted by this design element of the CPS.

Fourth, problems associated with weighting and variance estimation will be greatly complicated by using linked CPS records. The BLS provides cross-sectional weights for use with single CPS basic monthly or supplemental data files. We are developing longitudinal person and household weights for dissemination as part of the enhanced IPUMS-CPS, but the weights we produce may not be perfect for all purposes. For example, it is not clear to us that any of the seven research designs described above would utilize the same longitudinal person weights since each is subject to different types and volumes of selective panel attrition; this is to say nothing of the need for longitudinal *household* weights for various research designs. We suspect that there may be need for as many sets of weights as there are possible ways to link CPS records.

Unless there is consensus that a single weight can feasibly support all research designs, this issue should concern researchers who seek to produce externally valid results.

Finally, by virtue of the short time intervals between basic monthly surveys, CPS data may be especially prone to panel conditioning (Warren and Halpern-Manners 2012). This form of bias arises as respondents to longitudinal surveys change their attitudes, behaviors, or statuses (or at least their survey reports of those things) because of being interviewed on multiple occasions. The BLS has long warned that panel conditioning or "time in sample" effects may influence CPS-based estimates of unemployment and labor force participation rates. Indeed, the issue has made its way into documentation about the design of the CPS (e.g., U.S. Bureau of Labor Statistics 2006: Pp. 16-17). A number of observers have noted that unemployment rates, in particular, are considerably higher among respondents who are participating for the first time as compared to those who are experienced CPS respondents (e.g., Bailar 1975; Bailar 1989; Hansen et al. 1955; Shack-Marquez 1986; Shockey 1988; Solon 1986; Williams and Mallows 1970). Halpern-Manners and Warren (2012) recently showed that these apparent biases cannot be attributed to panel attrition or mode effects. In general, researchers—especially those using items collected on basic monthly surveys—will need to think carefully about whether their inferences about changes across months can be influenced by panel conditioning.

6. Discussion

With support from the National Institute for Child Health and Human Development, we have recently begun a project to develop integrated data, dissemination software, and associated metadata that will make longitudinal analyses of CPS data radically easier. We will freely disseminate the data as part of the IPUMS project via an innovative user interface that will

dramatically simplify and improve search, discovery, research design, and data access. We will provide researchers with flexible access to integrated and well-documented longitudinal data across all CPS surveys, including all surviving basic monthly surveys and all topical supplements. The resulting data will serve the scientific enterprise by reducing wasteful duplication of effort (e.g., in linking files and harmonizing variables), eliminating common technical errors (e.g., in linking and variance estimation), making findings easier to replicate, and encouraging and facilitating sophisticated and powerful new longitudinal analyses in many research domains.

Longitudinally linked and comprehensive CPS basic monthly and supplemental survey data will be valuable to multiple research communities for at least two reasons. First, the capacity to link data across CPS supplements on different topics will multiply the substantive topics that may readily be studied. For example, detailed questions about educational enrollment do not appear on the same supplement as detailed questions about veterans' issues; thus, linking data across supplements will facilitate new research on relationships between education and military service. Second—and even more exciting—linked samples will make it dramatically easier for researchers to use the CPS as a longitudinal data resource.

Many widely used longitudinal surveys—such as the Panel Study of Income Dynamics (PSID), the National Longitudinal Survey of Youth (NLSY), and the Health and Retirement Survey (HRS)—focus on particular content domains and follow specific cohorts of Americans over long periods. We do not pretend that the longitudinal IPUMS-CPS—that observes respondents for just 16 months—can replace these important resources. We do suggest, however, that IPUMS-CPS data will be valuable for studying processes of change. Moreover, the relatively limited sample sizes of surveys like PSID, NLSY, or the HRS sharply restrict researchers' capacity to study smaller

population subgroups; the large sample sizes of the CPS accommodate such subgroup analysis. The longitudinal IPUMS-CPS will also serve as an important complement to the Survey of Income and Program Participation (SIPP), which is now the main data resource for studying short-term income dynamics, program participation, and poverty. In particular, IPUMS-CPS will have far broader subject coverage, larger sample sizes, and substantially greater chronological depth than SIPP. CPS data have been collected and released annually for nearly 50 years, while there have been just four non-overlapping SIPP panels since 1993.

We had three objectives in this paper. First, we described our techniques for linking all CPS person- and household-level records over time from 1989 forward. This step—which involves the creation of new and unique household and person level identifiers for every CPS record, named CPSID and CPSIDP, respectively—builds on methods described by Madrian and Lefgren (2000), Feng (2001; 2008), and others, but implements them on an entirely different scale. Second, we demonstrated seven possible research designs based on longitudinally linked CPS records on *people*; a similar diversity of designs can be implemented for research on *households*. Our goal in this section was to inspire and inform innovative new research by demonstrating these various research designs based on linked CPS person-level data. Third, we provided information about the sample sizes and retention rates that researchers can expect when they implement one of these seven research designs. Information of this sort is foundational for researchers seeking to design new longitudinal analyses of CPS data.

7. References

Bailar, B. A. 1975. "Effects of Rotation Group Bias on Estimates from Panel Surveys." *Journal of the American Statistical Association* 70(349):23-30.

Bailar, Barbara A. 1989. "Information Needs, Surveys, and Measurement Errors." Pp. 1-24 in *Panel Surveys*, edited by Daniel Kasprzyk, Greg J. Duncan, Graham Kalton, and M.P. Singh. New York: Wiley.

Burkhauser, Richard V., Mary C. Daly, Andrew J. Houtenville, and Nigar Nargis. 2002. "Self-Reported Work-Limitation Data: What They Can and Cannot Tell Us." *Demography* 39:541-55.

Cameron, Stephen, and Joseph Tracy. 1998. "Earnings Variability in the United States: An Examination Using Matched-CPS Data." Unpublished Paper, Federal Reserce Bank of New York.

Celik, Sule, Chinhui Juhn, Kristin McCue, and Jesse Thompson. 2012. "Recent Trends in Earnings Volatility: Evidence from Survey and Administrative Data." *The B.E. Journal of Economic Analysis & Policy* 12, 2 (Contributions):Article 1.

Feng, Shuaizhang. 2001. "The Longitudinal Matching of Current Population Surveys: A Proposed Algorithm." *Journal of Economic and Social Measurement* 27:71-91.

—. 2008. "Longitudinal Matching of Recent Current Population Surveys: Methods, Non-matches and Mismatches." *Journal of Economic and Social Measurement* 33:241-52.

Frazis, H. J., E. L. Robison, T. D. Evans, and M. A. Duff. 2005. "Estimating gross flows consistent with stocks in the CPS." *Monthly Labor Review* 128(9):1-7.

Geist, Claudia, and Patricia A. McManus. 2008. "Geographical Mobility over the Life Course: Motivations and Implications." *Population, Space and Place* 14(4):283–303.

—. 2012. "Different Reasons, Different Results: Implications of Migration by Gender and Family Status." *Demography* 49(1):197-217.

Halpern-Manners, Andrew, and John Robert Warren. 2012. "Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey." *Demography* 49:1499-519.

Hansen, Morris H., William N. Hurwitz, Harold Nisselson, and Joseph Steinberg. 1955. "The Redesign of the Census Current Population Survey." *Journal of the American Statistical Association* 50:701-19.

Jennifer Van Hook, and Weiwei Zhang. 2011. "Who Stays? Who Goes? Selective Emigration Among the Foreign-Born." *Population Research and Policy Review* 30(1):1-24.

Katz, Arnold, Klaus Teuter, and Philip Sidel. 1984. "Comparison of Alternative Ways of Deriving Panel Data from the Annual Demographic Files of the Current Population Survey." *Review of Public Data Use* 12:35-44.

Kelly, Terence F. 1973. "The Creation of Longitudinal Data From Cross-Section Surveys: An Illustration from the Current Population Survey." Pp. 206-11 in *Annals of Economic and Social Measurement*, edited by National Bureau of Economic Research. Washington, D.C.: National Bureau of Economic Research.

Madrian, Brigitte C., and Lars John Lefgren. 2000. "An Approach to Longitudinally Matching Current Population Survey (CPS) Respondents." *Journal of Economic and Social Measurement* 26(1):31-62.

Nekarda, Christopher J. 2009. "A Longitudinal Analysis of the Current Population Survey: Assessing the Cyclical Bias of Geographic Mobility." Unpublished paper, Federal Reserve Board of Governors.

O'Connell, Martin, and Carolyn C. Rogers. 1983. "Assessing Cohort Birth Expectations Data from the Current Population Survey, 1971-1981." *Demography* 20:369-84.

Pitts, Alan. 1988. "Matching Adjacent Years of the Current Population Survey." Unpublished manuscript. Los Angeles, CA: Unicon Corporation.

Pleau, Robin, and Kimberlee Shauman. Forthcoming. "Trends and Correlates of Postretirement Employment, 1977–2009." *Human Relations*.

Shack-Marquez, J. 1986. "Effects of Repeated Interviewing on Estimation of Labor-Force Status." *Journal of Economic and Social Measurement* 14(4):379-98.

Shockey, James W. 1988. "Adjusting for Response Error in Panel Surveys: A Latent Class Approach " *Sociological Methods & Research* 17:65-92.

Solon, G. 1986. "Effects of Rotation Group Bias on Estimation of Unemployment." *Journal of Business & Economic Statistics* 4(1):105-09.

U.S. Bureau of Labor Statistics. 2006. *Design and Methodology: Current Population Survey*. Technical Paper 66. Washington, D.C.: U.S. Department of Labor, Bureau of the Census.

Warren, John Robert, and Andrew Halpern-Manners. 2012. "Panel Conditioning Effects in Longitudinal Social Science Surveys." *Sociological Methods & Research* 41(4):491-534.

Williams, W. H., and C. L. Mallows. 1970. "Systematic Biases in Panel Surveys Due to Differential Nonresponse." *Journal of the American Statistical Association* 65:1338-49.

Zullo, Roland. 2012. "The Evolving Demographics of the Union Movement." *Labor Studies Journal* 37(2):145-62.