Technical Description of SIPP Job Identification Number Editing in the 1990-1993 SIPP Panels

Martha H. Stinson
U.S. Census Bureau

July, 2003

1 Introduction

SIPP data collection has made significant quality improvements over the decade of the 1990's. One large improvement was the development of the CAPI system which allowed data collected in previous waves to be readily accessible to field representatives. One of the largest impacts of this innovation was the ability to better identify and track jobs held by SIPP respondents across waves of the survey. When a respondent reported a job, the field representative, using the CAPI software, assigned that job an identification number that was kept constant over time. When a new job was reported, a new identification number was assigned. Thus in the 1996 survey, job transitions can be identified and job tenure can be calculated.

Detailed data on jobs have been collected by the SIPP since the survey's inception. However the early panels had difficulty assigning longitudinally consistent job identification numbers. An investigation of the job id variables in the 1990-1993 panels has shown that accurate calculation of job tenure and identification of job transitions is compromised by longitudinal errors in these variables. There are two major types of problems - improper re-use of job identification numbers and improper assigning of new identification numbers. Tables 1A and 1B give generic examples of these problems. In Table 1A, the SIPP respondent held the same job throughout the first four waves of the survey. However, in wave 3, the job identification number was incorrectly changed, causing it to appear as if there had been a new job transition. In the internal SIPP data, this error is identifiable because the name of the employer stays the same across the waves. However, in the public use SIPP data, there is no way to tell that the jobs described in waves 2 and 3 are the same job. Table 1B shows the second type of problem. In this case, the person changed jobs between waves 3 and 4 but the job identification number was not changed. Thus it appears that the person remained at the same job through all four waves and, consequently, a job transition was missed. Again, the true work history is apparent in the internal SIPP data but not in the public use data.

The goal of this project was to identify the extent of the problems in SIPP job coding and to provide a remedy that could be used by researchers. The project was undertaken by the Longitudinal Employer Household Dynamics (LEHD) Program, a branch of the Demographic Surveys Division, created in part to improve Census demographic surveys through the use of administrative records. These activities were authorized by a February 2001 Treasury Regulation allowing the Census Bureau to receive certain types of administrative data in order to improve Title 13 authorized Census products. This particular project made use of data that provided a secondary source of information about job tenure for SIPP respondents: extracts from the Social Security Administration's Master Earnings File. These extracts contained Detailed Earnings Records (DER) for all SIPP respondents with validated Social Security Numbers in the 1990-1993 panels. These DER data report annual earnings by employer and hence can be used to calculate number of jobs held and frequency of job transitions over the time period covered by the survey. Using these administrative data and the detailed name information contained on the internal SIPP files, LEHD has produced an edited set of job identification numbers for the 1990-1993 SIPP panels for public release. These revised job identifiers are being released in a data file that contains the unique longitudinal public use person identifier, the wave number, the old job identification number, the new job identification number, and a flag indicating whether any edits have been made. The person identifier, wave number, and old job identification number allow

researchers to identify the person-wave-job observation in the already existing SIPP public use data products and to attach the new job identification number. Calculations concerning work histories (i.e. job tenure, time of job transition, etc.) can then be made using the new job codes.

Table 8 gives a summary of the changes made in each of the four SIPP panels. The remainder of this document gives a detailed description of how LEHD produced the revised job identification numbers. Appendix A gives some sample SAS code which can be used to link the revised job identification numbers to the current versions of the public use wave files for the 1990-1993 panels. Appendix B provides a sample data dictionary for the revised job id data set.

2 Description of Job Identification Number Editing Process

2.1 Original SIPP Job counts

The focus of this project is the 1990, 1991, 1992, and 1993 SIPP panels and the specific variables from these panels that identify jobs. These panels were chosen because of the availability of matched administrative records and because they were conducted pre-CAPI instrument and so contained the highest error rates. In the public use wave files, data are stored as one record per person and there are two variables that identify jobs: WS12002 and WS22102. None, one or both of these variables (depending on whether the person had 0, 1 or 2 jobs in a wave) will contain numbers that identify the job. Other variables containing job-specific information are coded to indicate whether the information pertains to the job identified by WS12002 or WS22102. These job ids do not vary across months within a wave, but other job-specific variables such as earnings do vary across months.

The job data in the internal SIPP files are stored as person-job-wave observations. Job identifiers are contained in a variable called SC2002 which refers to the question number on the survey instrument. Since a respondent reports information on a maximum of two jobs per wave, for each panel the maximum number of person-job-wave observations possible is

$$\text{Total obs} = \text{Number of respondents} \times \text{Number of Waves} \times 2$$

In practice this total is never reached for several reasons. Not all respondents report holding jobs, and those who do may not work every wave and often do not hold 2 jobs at once. Table 2 gives the total number of SIPP respondents, the total number of respondents ever reporting a job, the total number of person-job-wave observations actually observed, and the total number of unique jobs. It is important to understand the concept of a job as used in the context of the SIPP survey instrument. The survey began the Earnings and Employment section by asking respondents who had reported in the Labor Force section of the survey that they were employed whether they worked for an employer or were self-employed. Those who reported having an employer were asked, "What is the name of the employer for whom ... worked during this 4-month period?" This employer was then given an identification number. The identification number was stored on the Control Card along with the name of the employer. At the time of the next interview the person was again asked the name of the employer, and if this employer was the same as in the previous wave, the identification number from the previous wave should have been re-used. Thus a job refers to a respondent-employer match that may continue across waves. The survey instrument was quite explicit about the longitudinal nature of the job or employer id. Check Item E3 (for waves other than the first) instructs field representatives to "Enter employer ID number from cc item 42, or if a new employer, enter the next available ID number." In theory, one should be able to group the person-job-wave observations together using the job id and obtain a count of the number of jobs a respondent held over the course of the survey. However, as shown in the examples in Tables 1A and 1B, either these instructions were not always carried out by the Field Representatives or else re-coding mistakes occurred when the data was put in electronic format. While the reasons for the problems are unclear after the passage of many years, the need to correct these problems is clear.

2.2 Phase 1 of Job ID editing: Name Matching

LEHD began the process of identifying problems and creating new job ids by using the most specific piece of job information available: name of the employer. Because the name of the employer was collected and written down

separately at the time of each interview, the wave-specific names were not always the same for a given employer. Different spellings, use of abbreviations in later waves, and slightly different wording were the most common differences across waves. Hence LEHD employed a name-matching software called Vality. This software parsed the employer names into pieces so that the unique part of the name could be separated from common words such as "company" or "firm." The software then created blocks of records that contained all the job-wave observations for a given individual. Within these blocks, the software attempted to group observations that had the same name using a set of user-supplied weights and cut-off points to define the concept of "same." The result of the process was a new set of job ids that grouped job-wave observations for an individual based on a probabilistic prediction about which names were the same[1].

The results of this exercise are presented in Table 3. The first row gives the total number of person-job-wave observations with valid name information (i.e. cases of blank names, refused names, and "don't know" responses were dropped). The next row shows the new total job count using the set of job ids created by Vality. This row can be compared to the last row of Table 2. For every panel, the new total job count is substantially higher than the original total job count. The third and fourth rows in Table 3 explain this difference. In the 1990 panel, for example, there were 4,073 post-Vality jobs that consisted of job-wave observations that had previously belonged to at least two different SIPP jobs (Table 3, row 3). These are possible cases of the type of error shown in Table 1A: failure by SIPP job ids to accurately link over time. However the number of these cases was small compared to the number of cases where Vality split job-wave observations that had previously belonged to one SIPP job into at least two different jobs. In 1990, 18,386 SIPP jobs were split apart in this manner. These are possible cases of the type of error shown in Table 1B: failure to correctly separate jobs over waves. The large number of this type of case results in a much larger total job count after the Vality name-matching.

At this stage of processing a small edit was performed to handle jobs with missing names. In cases where a job-wave observation was missing an employer name but was reported in a wave immediately after or immediately before a job-wave observation with the same SIPP job id, the two job-wave observations were linked and declared to be the same job. Because Vality could not link jobs with missing names, these observations had been coded as belonging to separate jobs. However for cases where the observations were close in time and had been coded by the SIPP as belonging to the same job, it seemed sensible to keep the SIPP link. This slightly reduced the overall number of jobs, as reported in the fifth row of Table 3.

The difficulty at this stage of the process was to know whether the job id changes made by Vality were correct. Clerical review of the cases where Vality linked two or more separate SIPP jobs into one job revealed that approximately 95% of the time, the names of the SIPP jobs were similar enough that they were almost certainly the same job. However, clerical review of cases where Vality split one SIPP job into multiple jobs revealed a much higher error rate. Because of the large number of cases where Vality had split SIPP jobs, it became critical to bring in outside information to identify respondents with inaccurate job histories post-Vality matching.

2.3 Phase 2 of Job-ID editing: Using SSA administrative data

Since February 2001, the Census Bureau has received administrative earnings records from SSA for SIPP respondents. These data provided information on annual earnings by employer and were used to produce counts of the number of jobs held by an individual over the calendar time covered by the SIPP survey. These individual job counts were then compared to individual job counts calculated from the Vality-generated job ids. In general we expected that there would be at least as many jobs in the administrative data as in the survey data for an individual. This expectation was based on the fact that the administrative data contained earnings reports from all employers while the survey included a maximum of two employers per wave (4 month time period). The survey data also had the disadvantage of beginning and ending in the middle of years while the administrative data was strictly annual. Thus the 1992 panel, for example, would only contain reports on jobs that happened from October 1991 through April 1995 whereas the administrative records would include all jobs from 1991 to 1995. Because of these differences, we decided that job histories where administrative job counts were greater than or equal to Vality job counts were likely to be fairly accurate while job histories with administrative job counts less than Vality job counts were cause for concern. Row

[1]For more technical details about the name matching process, see LEHD Technical Paper No. TP-2002-24 and LEHD Technical Paper No. TP-2002-17 available at http://www.lehd-test.net/papers.

1 in Table 4 shows the number of jobs associated with SIPP respondents for whom the total administrative job count from all the survey years was less than the total Vality job count for the same years. Row 2 shows jobs from people with administrative job counts less than Vality job counts during some shorter time periods: the full survey years (1990, 1991 for the 1990 panel, 1991,1992 for the 1991 panel, 1992, 1993, 1994 for the 1992 panel, and 1993, 1994 for the 1993 panel), the first survey year only, or the second survey year only. Row 3 gives the total of all jobs associated with people who had troubling differences between their DER and Vality job histories.

All of the jobs included in row 3 were sent through another pass of name matching with the Vality software, this time with slightly relaxed name-matching criteria due to the belief that Vality had failed to link some job-wave observations for these individuals in the first pass. Table 5, row 1 gives the number of jobs sent through pass 2 and row 2 shows the new total number of jobs after Phase 2 processing. These numbers are substantially lower than in Table 3, row 5 for all panels, indicating that significant matching happened in this Vality pass. The third row shows how many jobs were associated with people who still have fewer administrative jobs than Vality jobs even after this second pass. While this number was cut in half for every panel, it was still a substantial portion of the overall number of jobs.

2.4 Phase 3 of Job ID editing: Clerical editing

The next stage of the job id editing process was clerical edits. This was accomplished by first flagging people whose total job counts in the DER data were less than their total job count in the SIPP after the second Vality pass. These people who considered to be the largest potential source of false job transitions created by the failure of the name-matching software to link job-wave observations. Thus job-level records were created for each of these individuals using the job ids created by the second Vality name-matching pass. All the job records for an individual with this type of count discrepancy were output and each individual's job history was clerically reviewed. Cases where two jobs had the same name but Vality had failed to link them were marked as belonging to the same job. The first row of table 6 shows the total number of jobs that were clerically reviewed and the second row shows the new overall number of jobs after the review.

After the clerical edits, a check was done to see whether the editing process to this point had erroneously linked two job-wave observations that were actually in the same wave because the jobs had similar names. The third row shows the number of jobs affected by this problem and the fourth row shows the number of job-wave observations that were clerically checked to solve this problem. The final new overall job total in the fifth row was slightly higher than in the second row because this last correction involved splitting jobs that were falsely linked.

2.5 Phase 4 of Job ID editing: Final edits

Table 7 begins by showing the overall job counts for each panel after the first three phases of job id editing. The next two columns show the types of cases where Vality or hand-editing either joined SIPP jobs (row 2) or split SIPP jobs (row 3). Due to the high number of jobs that were split, some final investigative edits were done to find types of cases where Vality had made incorrect job id assignments. The next row shows cases where a SIPP job had been split into two jobs by Vality and one of the jobs consisted of a single job-wave observation and the other job had at least four job-wave observations. The single job-wave observation fell in a wave either right before or right after the larger group of job-wave observations that made up the second job. In this case, the observations from both groups were clerically reviewed to determine whether the "orphan" job-wave observation was really a part of the larger group and perhaps had been incorrectly excluded because the name of the employer was slightly different.

Row 5 shows cases that were similar to those described in the previous paragraph except the "orphan" job-wave observation fell in the middle of the larger group of job-wave observations. For example if the job with only one wave observation was in the fourth wave of the survey and the job with at least four wave observations was in the second, third, fifth, and sixth waves, then all the wave observations from both jobs were automatically flagged as belonging to the same job. In essence, due to employer name anomalies, Vality had created a false hole in the middle of a job by marking one job-wave observation as a separate job. This edit corrected this type of problem. The last row of Table 7 shows the new total job count after both types of edits had been performed.

2.6 Final Release

Table 8 shows the totals for the final release of the public use data set containing the revised job ids. When the SIPP wave files were initially released, a few people from the internal SIPP files were dropped. In addition, for the 1992 panel, the 10th wave of the survey was never released as a stand-alone public use file. Our data set respects these conventions and includes only people and person-job-wave observations that were part of the original public release. Approximately 200 people per panel were dropped[2] as well as person-job-wave observations from the 10th wave of the 1992 panel. Hence the totals reported in rows 1-4 of Table 8 are slightly lower than the totals in the same rows in Table 2. Row 5 of Table 8 shows the final total number of jobs after all the editing phases were completed. Row 6 shows the number of cases where the editing process linked two separate original SIPP jobs. Row 7 shows the number of cases where the editing process split one original SIPP job into at least 2 jobs.

2.7 Conclusion

This new release of SIPP job ids should allow researchers to calculate job tenure and job transitions with greater accuracy and to make comparisons between the job tenure statistics calculated from early 1990's SIPP data and statistics calculated with more recent SIPP data. This new use of administrative data to improve Census products will greatly benefit researchers and facilitate the usefulness of SIPP data without any additional respondent burden.

3 Appendix A: Sample SAS code for merging revised job ids

The following SAS code will prepare person-job-wave data sets from the public use wave files and will merge these data sets to the revised job id data set.

```
/*Make extract from Public Use Wave File; Keep longitudinal person identifier, jobids, and wave variables
plus any other person/job characteristic variables desired.
These files are person-month level but the jobids do not change across the 4 months of a wave so keep only
the first observation for each person. Output to firstjobs data set if have non-missing jobid in the ws12002
variable and output to secondjobs data set if have non-missing jobid in the ws22102 variables */
data firstjobs_{panelyear}_{wave}(keep=puid jobid1 wave) secondjobs_{panelyear}_{wave}(keep=puid jobid2 wave);
  set SIPP_PUBLIC_USE_WAVE_FILE(keep=suid entry pnum wave ws12002 ws22102 /*plus other job variables
of interest*/);
by suid entry pnum;
length puid $14 jobid1 $2 jobid2 $2;
if first.pnum then do;
  puid=suid||entry||pnum;
  jobid1=ws12002;
  jobid2=ws22102;
 if jobid1 ne ' ' and jobid1 ne '00' then output firstjobs_{panelyear}_{wave};
  if jobid2 ne ' ' and jobid2 ne '00' then output secondjobs_{panelyear}_{wave};
end;

/*Sort both data sets by person identifier*/
proc sort data=firstjobs_{panelyear}_{wave};
   by puid;
proc sort data=secondjobs_{panelyear}_{wave};
   by puid;

/*Repeat the above steps for all waves in the panel*/
```

[2]The exception is the 1992 panel where the number was much higher because anyone entering the panel in wave 10 was dropped.

```
/*Stack all wave extracts together, grouping by person identifier, to make one file containing all the person-job-
wave observations from a panel*/
data alljobs_{panelyear};
  set firstjobs_{panelyear}_1 secondjobs_{panelyear}_1
      .....
      firstjobs_{panelyear}_{lastwave} secondjobs_{panelyear}_{lastwave};
  by puid;
if jobid1 ne ' ' then jobid=jobid1;
else if jobid2 ne ' ' then jobid=jobid2;

/*Sort the resulting data set in preparation for merge with revised jobids*/
proc sort data=alljobs_{panelyear};
   by puid jobid wave;

/*Read in revised jobid file:  convert from ASCII to SAS*/
filename in '.../SIPP_REVISED_JOBID_FILE_{panelyear}';
data alljobs_revised_jobid_{panelyear};
  infile in lrecl=25;
  length suid $9 entry $2 pnum $3 jobid $2 jobid_revised $2;
   input suid $ 1-9 entry $ 10-11 pnum $ 12-14 panel 15-18 wave 19-20 jobid  $ 21-22 jobid_revised $ 23-24
flag_jobid_change 25;
  length puid $14;
  puid=suid||entry||pnum;

proc sort data=alljobs_revised_jobid_{panelyear};
  by puid jobid wave;
/*Merge to edited jobid file by Person identifier, jobid, and wave variables*/
data alljobs_old_and_new_jobids_{panelyear};
  merge alljobs_{panelyear} alljobs_revised_jobid_{panelyear};
by puid jobid wave;
```

4 Appendix B: Sample Data Dictionary for the Revised Jobid Data Set

SIPP 1990 REVISED JOBID DATA DICTIONARY

July, 2003

D SUID 9 1
Sample unit identifier
This identifier is created by scrambling
together the PSU, segment and serial of
the original sample address. It may be
used in matching sample units from
different waves
Range=(000000000:999999999)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D ENTRY 2 10
Edited entry address ID
Address ID of the household that this

person belonged to at the time this
person first became part of the sample
Range=(11:89)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D PNUM 3 12
Edited person number
Range=(101:899)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D PANEL 4 15
Sample code - indicates panel year
Range=(1990:1990)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D WAVE 2 19
Control card item 36A - wave number
associated with the interview status
Range=(1:8)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D JOBID 2 21
Originally Released Employer I.D. number
(Same as WS12002 or WS22102 on public use wave files)
Range=(01:14)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave
D JOBID_REVISED 2 23
Edited Employer I.D. number
Changed to be longitudinally consistent across all waves
Range=(01:10)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave

D FLAG_JOBID_CHANGE 1 25
Indicator for whether a revision was made to the original jobid
Range=(0:1)
U All persons who reported holding at least one job during at least one wave of the panel,
i.e. WS12002 > '00' for at least one wave
V 0 .No revisions made to this jobid in this wave and jobid=jobid_revised
V 1 .Revisions made to this jobid in this wave and jobid $\neq$ jobid_revised

Table 1A: Failure to link job across waves		
Wave	Firm Name	SIPP Jobid
1	AAAA	1
2	AAAA	1
3	AAAA	2
4	AAAA	1

Table 1B Failure to separate jobs across waves		
Wave	Firm Name	SIPP Jobid
1	AAAA	1
2	AAAA	1
3	AAAA	1
4	BBBB	1

Table 2: Original SIPP Totals					
Row	SIPP Panel	1990	1991	1992	1993
1	Total SIPP respondents	69,432	44,373	62,412	62,721
2	Total respondents who ever report a job	37,291	23,520	33,920	32,972
3	Person-job-wave observations	216,851	136,693	228,214	208,748
4	Jobs defined by original SIPP jobid (WS12002, WS22102 on public-use wave files SC2002 on internal use wave files)	57,800	35,515	55,453	52,591

Table 3: Phase 1 - Job name matching, Pass 1 through Vality					
Row	SIPP Panel	1990	1991	1992	1993
1	Person-job-wave-obs with non-missing job name	216,156	136,253	227,395	208,062
2	Total job count using job id created by Vality	78,495	46,527	74,405	69,057
3	Jobs where Vality linked 2+ separate SIPP jobs	4,073	2,265	3,811	2,702
4	SIPP jobs that were split into 2+ jobs by Vality	18,386	10,080	16,403	14,129
5	New total job count after edit that linked some jobs with missing names	78,225	46,316	74,078	68,803

Table 4: Phase 2 - Using SSA Admin. Data, Comparing Job Counts					
Row	SIPP Panel	1990	1991	1992	1993
1	Jobs assoc. with people for whom indiv. job count in DER < indiv. job count in Vality	20,983	10,380	16,526	16,229
2	Jobs assoc. with additional people where DER differs from Vality	24,742	13,769	22,226	19,123
3	Total num. of jobs assoc. with people with conflict between Vality and DER job counts	45,725	24,149	38,752	35,352

Table 5: Phase 2 - Using SSA Admin. Data, Pass 2 through Vality					
Row	SIPP Panel	1990	1991	1992	1993
1	Job obs sent through Pass 2	45,725	24,149	38,752	35,352
2	New total job count	69,138	41,814	66,602	62,251
3	Jobs assoc. with people for whom indiv. job count in DER < indiv. job count in Vality	10,011	5,106	8,131	8,330

Table 6: Phase 3 - Clerical edits					
Row	SIPP Panel	1990	1991	1992	1993
1	Job obs sent through clerical review	10,011	5,106	8,131	8,330
2	Total jobs	67,588	41,101	65,676	61,338
3	Jobs with two obs from same wave	628	370	608	601
4	Job-wave obs sent through editing	5,722	3,397	6,520	5,942
5	Total jobs	68,175	41,482	66,328	61,960

Table 7: Phase 4 - Final edits					
Row	SIPP Panel	1990	1991	1992	1993
1	Total jobs	68,175	41,482	66,328	61,960
2	Jobs where editing process linked 2+ separate SIPP jobs	4,570	2,529	4,255	3,134
3	SIPP jobs that were split into 2+ jobs by editing process	11,937	6,776	11,603	9,658
4	Clerically review job-wave obs for cases where SIPP jobid was split into 2 groups; one group has 1 obs and other group has +4 obs; small group falls in wave immed. before first wave or immed. after last wave of large group;	4,057	2,582	4,516	3,687
5	Clerically review job-wave obs for cases where SIPP jobid was split into 2 groups; one group has 1 obs and other group has +4 obs; small group falls in middle wave of larger group and larger group is missing obs for that wave	6,161	3,231	6,272	4,783
6	Total jobs	66,991	40,818	65,278	61,094

Table 8: Public Release - Final Counts					
Row	SIPP Panel	1990	1991	1992	1993
1	Total SIPP respondents	69,101	44,143	61,534	62,346
2	Respondents who ever report a job	37,080	23,362	33,192	32,811
3	Person-job-wave observations	215,799	136,054	210,414	207,909
4	Total jobs with original jobid variable	57,392	35,244	52,813	52,266
5	Total jobs with revised jobid variable	66,329	40,404	61,665	60,549
6	Jobs where editing process linked 2+ separate SIPP jobs	4,534	2,516	3,882	3,125
7	SIPP jobs that were split into 2+ jobs by editing process	10,528	5,988	9,670	8,634

*Differences in totals between Table 8 and Table 2 are due to dropping people who are in Internal SIPP files but not public use SIPP files

*Also all wave 10 observations for the 1992 panel were dropped because this wave file was not publicly released