Appendix 37: Summary of Data Cleaning Issues

Discrepancies in the Universe Skips

The term "universe skips" refers to the universe of respondents who skip from question to question. For example, CK-HEA-A is a check item that sends respondents who are currently employed to question HEA-2 and all others to question HEA-1. Thus, the universe of respondents in question HEA-1 consists of those who had a value of '0' in CK-HEA-A (not currently employed). Discrepancies in the universe skips occur when the number of respondents appearing in the follow-up question does not match the number of respondents in all lead-in questions (or more precisely, all lead-in categories). For example, a discrepancy of 3 cases exists if a universe of 30 respondents was skipped to a question but this follow-up item has a frequency of only 27 respondents.

After careful examination of the data, CHRR concluded these universe discrepancies were due to two types of problems with the survey program and break-off interviews. The first problem occurred at the close of the interview, when valid "on-path" data values were incorrectly declared off-path and were therefore not written out to the data file. The result is that some cases have no data in a question even though they have valid data in the lead-in and branching questions. The opposite problem also occurred, in which the program wrote out values that should have been defined as "off path." One reason for this is that the CAPI program fills in the data for check items throughout the survey as soon as data is available for satisfying the conditions of the check item. This means, for example, that a check item in section 11 (the income section) can be evaluated while the interviewer is still asking questions in section 2, as long as the data needed for the check item has been obtained. This mainly affected the universe skips among respondents who did not finish the interview. For example, a respondent who broke off the interview in the health section (section 10) might have data for a check item in the other family business section (OFB, section 12). The result is that the universe of respondents in a check item will be greater than the number of respondents leading in to the check item and the universe of respondents in the follow-up item.

"CK-DATA" Items

The first variable in most sections is a check item that identifies cases with complete, incomplete, or irregular data in the section. These items were created by CHRR to help users keep track of changes in the size of the universe of respondents who started and ended each section. For example, these check items could be used to restrict an analysis to respondents who completed the interview up to a certain section. The check items were also intended to identify cases with irregular data, such as a missing lead-in or follow-up question. Since it was not possible to document every irregularity for every case, the check items were intended to isolate the cases within each section where these irregularities were observed.

"Empty" Columns in Rostered Employer Questions

The expected pattern of data in rostered items is for the data to start in the first "column" (the item indexed with 'ARR-01' or '1') and continue to the right until all information has been collected. However, this pattern was not always obtained in the employer name rosters, as shown in panel A of Figure 37.1. (For ease of presentation the -4 and -5 values found in the actual data have been changed to a period.) For example, case #5112 contains employer names in the 2nd and 3rd items of the roster, while the first column is blank. These blanks occurred because of the procedures used by the CAPI program to retrieve, update, and store the contents of the employer name roster. For example, respondents were asked to confirm the name and work history of employer names that were included as part of the input file. If the respondent denied having worked for the employer the name was deleted from the roster. However, the space in the roster occupied by the name remained empty, and new names that were collected in subsequent questions were stored in the next column. Thus, empty columns are not indicative of data problems or irregularities.

Figure 37.1
Employer Names, Work History Information, and Employer Stop Dates

(A: Employer name roster)

YW	EMPL-	EMPL-	EMPL-	EMPL-	EMPL-	EMPL-
ID #	ROST1	ROST2	ROST3	ROST4	ROST5	ROST6
0051	1	2	3	.	.	.
5112	.	2	3	.	.	.
4824	1	2	.	4	.	.
4412	1	.	3	.	.	.
4031	1	2	.	4	5	.

(B: Data in RSP-153)

	RSP-	RSP-	RSP-	RSP-	RSP-	RSP-
ID #	153-01	153-02	153-03	153-04	153-05	153-06
0051	4	.	1	.	.	.
5112	.	2	1	.	.	.
4824	2	.	.	2	.	.
4412	2	.	2	.	.	.
4031	4	.	.	4	2	.

(C: Stop dates)

ID #	DOLI	End 01	End 02	End 03	End 04	End 05	End 06
0051	20-OCT-93	15-FEB-94	15-NOV-91	30-JUN-95	.	.	.
5112	15-OCT-93	.	18-JAN-94	25-JUL-95	.	.	.
4824	02-NOV-93	17-JUN-95	10-APR-93	.	17-JUN-95	.	.
4412	05-OCT-93	28-JUL-95	.	28-JUL-95	.	.	.
4031	15-OCT-93	25-MAY-94	21-SEP-93	.	07-AUG-95	07-AUG-95	.

(D: CK-RSP-I)

		CK-	CK-	CK-	CK-	CK-	CK-
ID #	DOI	RSP-I-01	RSP-I-02	RSP-I-03	RSP-I-04	RSP-I-05	RSP-I-06
0051	30-JUN-95	2	.	1	.	.	.
5112	25-JUL-95	.	2	1	.	.	.
4824	17-JUN-95	1	.	.	1	.	.
4412	28-JUL-95	1	.	1	.	.	.
4031	07-AUG-95	2	.	.	1	1	.

(E: RES-1)

ID #	RES-1-01	RES-1-02	RES-1-03	RES-1-04	RES-1-05	RES-1-06
0051	0	0	0	.	.	.
5112	.	0	0	.	.	.
4824	0	0	.	0	.	.
4412	0	.	0	.	.	.
4031	0	0	.	0	0	.

It is also possible for respondents to have data in the employer roster but not in the corresponding RSP roster item. For example, as panel B of Figure 37.1 shows, case #0051 has no data in the second rostered item in question RSP-153 even though an employer name is listed in RES-EMPL-ROST2 (see panel A). The end date for this employer (see panel C) reveals that the respondent had stopped working at this job prior to the date of last interview, making the employer ineligible for questions in the RSP section. The same is true for other employers for other respondents, such as employer #2 for case #4031 and case #4824. The variables in CK-RSP-I (see panel D) can be used to determine the work history status of each employer name in the job roster. When CK-RSP-I equals 1, the respondent is currently employed with the employer; a value of 2 in CK-RSP-I means the respondent has worked for the employer since the date of last interview but is not currently working; when CK-RSP-I is -4 (blank in panel D), there is no employer name listed or the employer has a stop date preceding the date of last interview.

Finally, it should be noted that employer names that are out of scope for the RSP section may have data in the RES and OJS sections (see panel E). This is because the employer name roster was not finalized until the end of the RES, the point at which the employers' dates and names had been revised and updated.

Section Timings

The section timing variables are unreliable if the interviewer backed up over a section during the interview or had to break off and resume the interview later. More specifically, when an interviewer jumped back to a question in a previous section the start- and end-time counters for that section and all intervening sections were reset to zero. Thus, if an interviewer jumped from the Health section back to the RSP section, the timing variables for the sections between RSP and Health were re-initialized. As a result of this and other problems it was possible for a respondent to end up with negative numbers and zeros for section timings. These negative values were retained in the data so that the researcher could decide how to handle these variables.

Unusual Data Values

Some 1995 data items may contain what appear to be implausible or unreasonable values. For example, according to R18957. (RSP-63-ARR-01), several respondents travel more that 600 miles round-trip to work. While these values may not be incorrect, they seem unusual. In the past, when a respondent had an unusual value for an item, the archivist could refer to the respondent's paper questionnaire to determine if the value was the result of a data entry error. However, this is no longer possible with a CAPI instrument. Instead of blanking these suspicious values they were left in the data so that the individual researcher could decide how to handle them. Researchers using such items may want to look at the rest of the respondent's record for the current year or look at this same item in earlier years.

CK-OJS-Q-ARR

Due to a CAPI programming error the data for these rostered check items were not written out to the data file. However, these check items functioned properly during the course of the survey. The employer information obtained in these questions was used to finalize the employer roster and ultimately contributed to the creation of the CK-RSP-I-ARR variables found near the beginning of the RSP section.

RSP-40: Within-Job Gap Dates

Within-job gaps are defined as periods of one week or more during which respondents did not work for an employer, not counting paid vacations and sick leave. Due to a problem with the CAPI instrument, the data for the start and stop dates of these within-job gaps could not be retrieved.

Income Issues for the 1995 Data Set

Topcoding is done to protect the respondent's confidentiality. The Census Bureau topcodes all income items to a fixed dollar amount (an amount that has increased over the years in response to inflation). All yearly income amounts are topcoded to the same dollar figure and all weekly, monthly, quarterly etc. amounts are topcoded proportionally. Assets are usually topcoded to the mean of the three highest values.

In 1995 there were no soft ranges for dollar amounts programmed in the CAPI questionnaire, and the hard ranges were very broad. The format used in the program for some number and dollar items allowed the interviewer to enter numbers with a decimal in the answer but also accepted numbers without a decimal. Numbers with and without decimals were made uniform by removing the decimals from the data and multiplying the rest of the values by 100. However, there is a possibility that interviewers may have entered a number expecting the program to insert the decimal automatically. For example, an interviewer intending to record $200.00 for monthly income could have entered a value of 20000, assuming that the program would insert a decimal and the value would be converted to 200.00. While the assumption is that most of the interviewers entered the items correctly, there were a number of suspicious values discovered during the checking of the items in question by both the Census Bureau and CHRR. A number of questionable values were left in the data; the individual researcher must decide how to handle them. This format was not used in any of the subsequent rounds of the survey.