Paper PO-07

How to Monitor "Don't Know" and "Refusal" Non-responses in a Large National Survey -- Using Simple SAS® Macros, a Few PROCs, and Data Steps.

Mariah Mantsun Cheng, Ph.D. & Timothy Monbureau
UNC Carolina Population Center, Chapel Hill, NC

ABSTRACT

Monitoring data quality is a critical and sometimes daunting task during any data collection effort. One simple way to assess the quality of typical interview data involves tracking the number of valid and missing responses to survey items. Such tracking may lead to the early detection of problematic questions, enabling researchers to redesign instruments before proceeding further with the potentially costly collection of flawed data. By recoding data into the 4 outcome categories of "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response," we use SAS macros and the common SAS CONTENTS, TRANSPOSE, FREQ, APPEND, and SUMMARY procedures in some uncommon ways to determine response rates for specific questions, larger questionnaire sections, and the survey as a whole. Our technique also facilitates the informative calculation of both relative response rates, which are based on the number of legitimate cases for specific questions, and absolute response rates, which are based on the total number of respondents in the survey sections. We use the well-known Longitudinal Study of Adolescent Health (Add Health) Wave IV interview as an illustration here.

INTRODUCTION

Monitoring data quality is crucial in survey data collection. Project or data managers who are at the receiving end of any sizeable collection endeavor have to oversee a great amount of data streaming in on a regular basis from collection teams. To manage these data systematically, it is necessary to have good programming tools that extract relevant information and generate statistical summaries reporting data quality.

Keeping tabs on item non-response rates is one way to gauge data quality. High percentages of non-response (e.g., "Don't Know," "Refusal") in a certain questionnaire item flag a potential data collection or questionnaire design flaw. Catching and resolving such a problem early in the data collection process help to improve the ultimate data quality.

In this paper, we present simple SAS macros that use common SAS procedures to monitor the item non-response rates of a large national longitudinal survey and ensure its data quality. In general, data quality can be monitored at various levels: at the overall survey level, at the questionnaire section level, or at the level of each item. Here we focus on the non-responses of "Don't Know" and "Refusal" by measuring their percentages at the item level and summarizing them within questionnaire sections. Once item non-response rates are calculated, they can also track individual interviewer's performance or identify potential non-response outliers at the respondent level.

DATA

To illustrate the utility of our quality assurance program, we use the Wave IV National Longitudinal Study of Adolescent Health (Add Health). The Add Health Wave IV data collection was conducted by RTI International. We developed the SAS macro and program codes for monitoring the quality of the in-home personal interview data that we received from them on a weekly basis during 2007-2008.

The in-home interview data were collected using a 90-minute CAPI (Computer-Assisted Personal Interview)/ CASI (Computer-Assisted Self-Interview) instrument. Questionnaire contents were divided into 26 sections, covering a wide spectrum of questions on respondents' socio-demographic characteristics, their psychological and emotional states, social and relational behaviors, as well as many health-related habits and outcomes. A total of 15,701 respondents completed the survey.

Data collected for items under these 26 questionnaire sections were organized by their section code, followed by the question number. For example, H4GH1 was the name of the variable that contained data from the first question item in the questionnaire section H4GH (a prefix that stands for the "General Health and Diet" section). The original data files from RTI were SAS files, with all questionnaire variables stored as numeric type, and the codes -1, -2, and "." represented non-responses of "Don't Know," "Refusal," and "Legitimate Skip," respectively.

Since the Add Health Wave IV data have already been released (Nov. 2009), we re-ran our program on this final data set to demonstrate the SAS macros and other SAS procedures presented in this paper. In order to make the released data more consistent with the original data monitored, we recoded "Don't Know" responses to -1 and "Refusals" to -2. However, for continuity and readability, "Legitimate Skip" values were set to -3. The remaining valid responses were collapsed into a single category equaling 1. Hence, for each variable we had four exhaustive and exclusive categories: -1, -2, -3, and 1, which, again, denoted "Don't Know," "Refusal," "Legitimate Skip," and "all valid

responses." There were a total of 993 variables that contained data collected from all 26 questionnaire sections.

PROGRAM LOGIC

The SAS program codes we developed follow these logical steps:

Step 1. Specify input and output file names. Specify value codes for "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response." Declare length of the variable prefix and individual section codes for each questionnaire section.

Step 2. Read input data and select variables from a list of sections. Recode variable values such that they fall exclusively into one of the four categories -- "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response."

Step 3. Use the CONTENTS procedure to generate a list of variable names in sequential order and the total number of variables to be processed.

Step 4. Use the TRANSPOSE procedure to transpose the variable list for looping through subsequent FREQ procedures and APPEND procedures.

Step 5. Use the FREQ procedure to output the count and percent for each response category per variable/table, attach section and variable names to each record.

Step 6. Use the APPEND procedure to put all records into a data file -- which consists of at least 1 record (and a maximum of 4 depending on the number of response categories present) per variable.

Step 7. A series of DATA STEPs generate records that report frequencies of 0 and percentages of 0 for variables that lack all four of the response categories of interest. Note that the total number of records in this final file should be 4 times the total number of variables. In this case, 4 times 993 = 3,972 records.

Step 8. Use the SUMMARY procedure to compute additional summary statistics for data quality evaluation.

PROGRAM CODES FOR EACH STEP, USING ADD HEALTH WAVE IV AS AN EXAMPLE

The following program codes carry out each of the above logical steps on the Add Health Wave IV In-home Interview Data.

Step 1. Make use of %LET statements to create global macro variables that represent: input and output file names; "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response" values; prefix length, and questionnaire section codes.

```
%LET dataname=addhealth4;  /* Name of input data file                                  */
%LET dataout=freqrcds_ind; /* Name of output data file                                 */

%LET dk=-1;                /* Specify "Don't Know" Code                                */
%LET ref=-2;               /* Specify "Refusal" Code                                   */
%LET skip=-3;              /* Specify "Legitimate Skip" Code                           */
%LET vresponse=1;          /* Specify assigned code for all other "Valid Responses" */
%LET prfx=4;               /* Specify prefix length common to all variables in each */
                           /* section                                                  */
%LET sectlist=H4OD: H4WP: H4WS: H4GH: H4DA: H4PE;
                           /* Specify Section codes to select variables                */
```

Step 2. Read and select variables of interest. Recode variable values such that they fall exclusively into one of the four categories: "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response."

```
*-------------------------------------------------------------------------------*
| Read in the original data file consisting of selected variables. Extract variables|
| of interest and recode responses other than DK, RF, LS as one category --      |
| valid response.                                                                |
*-------------------------------------------------------------------------------- *;
DATA survey (DROP=i);
 SET in.&dataname(KEEP=&sectlist DROP=_CHARACTER_);
  ATTRIB _ALL_ LABEL='';  INFORMAT _ALL_;  FORMAT _ALL_;
  ARRAY recode[*] _ALL_;
        DO i=1 TO DIM(recode);
           IF recode[i] ~IN (&dk,&ref,&skip) THEN recode[i]=1;
           /* Recode responses other than "Don't know," "Refusal," and "Legitimate */
           /* Skip" as "Valid Response"                                             */
```

```
        END;
RUN;
```

Step 3. PROC CONTENTS produces a variable name list and the total number of variables.

```
*-------------------------------------------------------------------------*
| Use PROC CONTENTS to output variable names and their sequential order. |
*-------------------------------------------------------------------------*;
PROC CONTENTS DATA=survey VARNUM OUT=var_list NOPRINT; RUN;

PROC SORT DATA=var_list; BY VARNUM; RUN;
DATA _null_;
  SET var_list;
  CALL SYMPUT('n',VARNUM);
RUN;
%PUT NOTE: A total of &n variables to be processed from this data set;
```

Steps 4, 5, and 6. A simple macro (***table_records***) that makes use of the TRANSPOSE, FREQ, and APPEND procedures generates one-way table frequencies for each variable and then combines them into a complete table record that contains variable and section names, counts, and percentages for the “Don’t know,” “Refusal,” “Legitimate skip,” and “Valid Responses.”

```
*-------------------------------------------------------------------------*
| Transpose the variable list for later use on looping through process.  |
*-------------------------------------------------------------------------*;
PROC TRANSPOSE data=var_list OUT=transpose_vlist; var name; RUN;

*-------------------------------------------------------------------------*
| Write a macro named table_records:                                      |
| To turn one-way table frequency of each variable into a table record   |
| that contains the variable, section names, count, and percentages      |
| corresponding to each 4 exhaustive response categories –               |
| “Don't Know,” “Refusal,” “Legitimate Skip,” and “Valid Response.”      |
*-------------------------------------------------------------------------*;

%MACRO table_records;
%LET k=1;
%DO %WHILE (&k <= &n);
 DATA _NULL_;
 SET TRANSPOSE_vlist(KEEP=col:);
 CALL SYMPUT('var',(STRIP(col&k)));
 CALL SYMPUT('sect',(STRIP(SUBSTR(col&k,1,&prfx))));
 RUN;

 PROC FREQ DATA=survey; TABLE &var/OUT=t_&var NOPRINT; RUN;
 DATA record_&var(DROP=&var);
 SET t_&var;
 LENGTH name $ 8 section $ &prfx response 8;
 Section="&sect"; Name="&var"; response=&var;
 RUN;

 PROC APPEND BASE=all_records DATA=record_&var FORCE; RUN;
 %LET k=%EVAL(&k+1);
%END;
%MEND;

%table_records;
```

Step 7. Write another simple macro (***fill_in***) that generates table records for variables having zero frequency in any of the four response categories.

```
DATA var_list(RENAME=(nobs=sect_nobs));
 SET var_list(KEEP=name nobs VARNUM);
```

```
RUN;
PROC SORT DATA=var_list; BY name; RUN;

*-------------------------------------------------------------------------------*
| Write a macro named fill_in:                                                  |
| To generate table records for variables having 0 frequency in any of          |
| the "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response."         |
*-------------------------------------------------------------------------------*;

%MACRO fill_in(rt);
DATA &rt;
 SET all_records;
 IF response=&&&rt THEN OUTPUT &rt;
RUN;

PROC SORT DATA=&rt; BY name; RUN;

DATA &rt.0;
 MERGE var_list &rt; BY name;
 IF response=. THEN DO;
    response=&&&rt;     count=0;    percent=0;    Section=SUBSTR(name,1,&prfx);
 END;
RUN;
%MEND;

*-------------------------------------------------------------------------------*
| Systematically fill in table records for those variables with 0 frequency in any  |
| of the four response categories -                                             |
| "Don't Know," "Refusal," "Legitimate Skip," and "Valid Response."             |
*-------------------------------------------------------------------------------*;
%fill_in(dk);
%fill_in(ref);
%fill_in(skip);
%fill_in(vresponse);

DATA complete_rcds (RENAME=(percent=abs_percent));
 SET dk0 ref0 skip0 vresponse0;
 table_record=1;
RUN;
PROC SORT DATA=complete_rcds; BY name response; RUN;
```

Step 8. Use PROC SUMMARY to compute additional summary statistics, e.g., relative percentages of "Don't Know" or "Refusal" per questionnaire items, adjusting to the differential number of legitimate cases applicable for each question item.

```
*-------------------------------------------------------------------------*
| Compute relative percentages after removing "Legitimate Skip" counts    |
| from denominator.                                                       |
*-------------------------------------------------------------------------*;
DATA drop_SK;
 SET complete_rcds;  IF response NE &skip;
RUN;
PROC SUMMARY DATA=drop_SK;
 VAR table_record;
 BY name;
 OUTPUT OUT=cum_vcount SUM(count)=sum_validfreq;
RUN;

DATA out.&dataout;
 MERGE complete_rcds cum_vcount(DROP=_type_ _freq_);
       BY name;
 IF response NE &skip THEN DO;
       rel_percent=count/sum_validfreq;
 END;
LABEL sum_validfreq="Sum of responses (DK,RF,VR) in Var"
```

```
       abs_percent="Percent of DK,RF,VR,SK among all cases in Var"
       rel_percent="Percent of DK,RF,VR among legitimate cases in Var"
       count="Absolute count of cases in each DK,RF,SK,VR category in Var";
RUN;
PROC SORT DATA=out.&dataout; BY varnum DESCENDING response; RUN;
PROC PRINT DATA=out.&dataout; RUN;
```

TAKE A LOOK AT THE REAL DATA FILE CREATED

Applying the above program codes to the Add Health Wave IV interview data, which consists of 993 questionnaire items in 26 sections, we obtain a data file containing 3,972 records (= 993 x 4). This file has 4 records per each questionnaire item, reporting the item and section names, the specific response category, frequency count, the absolute percent (by count of all applicable cases in section), and relative percent (by count of only legitimate cases per item).

For illustration, the table below lists some of the records with variables: ***Section, Name, Sect_nobs, Sum_validfreq, Response, Count, Abs_percent, and Rel_percent***:

Section	Name	Sect_ nobs	Sum_ valid freq	Response	Count	Abs_ percent	Rel_ percent
S01 OvervDemo	H4OD4	15701	15701	1: VALID	15700	99.994	99.994
S01 OvervDemo	H4OD4	15701	15701	-1: DON'T KNOW	0	0.000	0.000
S01 OvervDemo	H4OD4	15701	15701	-2: REFUSAL	1	0.006	0.006
S01 OvervDemo	H4OD4	15701	15701	-3: LEGT SKIP	0	0.000	.
S01 OvervDemo	H4OD5	15701	978	1: VALID	978	6.229	100.000
S01 OvervDemo	H4OD5	15701	978	-1: DON'T KNOW	0	0.000	0.000
S01 OvervDemo	H4OD5	15701	978	-2: REFUSAL	0	0.000	0.000
S01 OvervDemo	H4OD5	15701	978	-3: LEGT SKIP	14723	93.771	.
S01 OvervDemo	H4OD6M	15701	644	1: VALID	564	3.592	87.578
S01 OvervDemo	H4OD6M	15701	644	-1: DON'T KNOW	77	0.490	11.957
S01 OvervDemo	H4OD6M	15701	644	-2: REFUSAL	3	0.019	0.466
S01 OvervDemo	H4OD6M	15701	644	-3: LEGT SKIP	15057	95.898	.
S01 OvervDemo	H4OD6Y	15701	644	1: VALID	603	3.841	93.634
S01 OvervDemo	H4OD6Y	15701	644	-1: DON'T KNOW	39	0.248	6.056
S01 OvervDemo	H4OD6Y	15701	644	-2: REFUSAL	2	0.013	0.311
S01 OvervDemo	H4OD6Y	15701	644	-3: LEGT SKIP	15057	95.898	.

Here the records belong to Section 1, Overview and Demographics, of the in-home interview questionnaire. The first 4 records each describes the frequency count (***Count***) of the categories: -- "Valid Response," "Don't know," "Refusal," "Legitimate Skip" (***Response***), the corresponding absolute percentage (***Abs_percent*** = ***Count*** divided by all applicable cases in the section (***Sect_nobs***)), and the relative percentage (***Rel_percent*** = ***Count*** divided by all legitimate cases in the specific question item (***Sum_validfreq***)). The variable ***name*** H4OD4 denotes the specific records that contain information from the questionnaire item no. ***4*** from the ***O***verview and ***D***emographic section. Since the calculations for absolute percents of "Don't Know" and "Refusal" categories are based upon all applicable cases per section, they are useful for evaluating the effects of such responses on the overall data quality per section. By the same token, since the relative percents of "Don't Know" or "Refusal" are based on legitimate cases per specific item, they are good indicators of data quality for a particular question item.

In this example, H4OD4 pertains to a question that asks, "Were you born a US citizen?" to all 15,701 respondents in the section. Therefore, the absolute percentages here are the same as the relative percentages. The next question item, H4OD5 ("Have you become a US citizen?"), only applies to those not born a US citizen (i.e., the 978 cases that show up in ***Sum_validfreq*** of the H4OD5 records, corresponding to 100% minus 93.771% of ***Sect_nobs***). The subsequent items, H4OD6M and H4OD6Y, ask in what month and year respondents not born US citizens were naturalized, which further narrows down the cases to 644 (=***Sum_validfreq*** in H4OD6M and H4OD6Y records).

The relative percentage variable (***Rel_percent***) of the "Don't Know" and "Refusal" records indicates that about 12% of these 644 cases did not know the month and 6% did not know the year in which they became US citizens. The relative percentage also shows that a very small percentage (about 0.31-0.47%) refused to answer the question. Nonetheless, since this question (H4OD6) only applies to a relatively small number of cases (644 out of 15,701), the absolute percentages (0.49% and 0.25% for "Don't know;" 0.02% and 0.01% for "Refusal" in ***Abs_percent***) should accurately gauge the effect that "Don't Know" and "Refusal" responses to these items have on data quality for this section or the overall questionnaire.

SOME INTERESTING RESULTS

1. ANALYSIS AT ITEM LEVEL.

There are many ways to utilize the above table-record file for monitoring data quality in survey research during the data collection process. One simple way is to rank-order the records according to the percentages of "Don't Know" or "Refusal" responses. For example, we can select the "Don't Know" response records and sort them in descending order according to their relative percentages, and set aside those with a large magnitude for closer examination.

As an illustration, the following table shows the question items in Add Health Wave IV that have more than 20% of applicable cases replying "Don't Know" to a specific question. These items warrant closer attention to see if the larger percentages of "Don't Know" responses are due to faulty questionnaire design (e.g., double-barrel wording of a question or unclear instructions to the interviewees) or reflect some reasonable, possible factual "Don't Know" realities.

Section	Name	Sect_ nobs	Sum_ valid freq	Response	Count	Abs_ percent	Rel_ percent
S16 Relatnship	H4TR22	30263	324	-1: DON'T KNOW	166	0.54852	51.2346
S08 HousRostr	H4HR11MN	15701	2	-1: DON'T KNOW	1	0.00637	50.0000
S08 HousRostr	H4HR11MO	15701	2	-1: DON'T KNOW	1	0.00637	50.0000
S15 SuicSexpSTD	H4SE18	15701	32	-1: DON'T KNOW	13	0.08280	40.6250
S03 RelatnSibs	H4WS3C	15701	40	-1: DON'T KNOW	14	0.08917	35.0000
S03 RelatnSibs	H4WS3E	15701	3	-1: DON'T KNOW	1	0.00637	33.3333
S03 RelatnSibs	H4WS3F	15701	3	-1: DON'T KNOW	1	0.00637	33.3333
S08 HousRostr	H4HR7D	15701	12	-1: DON'T KNOW	4	0.02548	33.3333
S02 ParentSupp	H4WP18	15701	29	-1: DON'T KNOW	9	0.05732	31.0345
S18 Pregnancy	H4PG4	21966	202	-1: DON'T KNOW	62	0.28225	30.6891
S03 RelatnSibs	H4WS3D	15701	11	-1: DON'T KNOW	3	0.01911	27.2727
S12 Economics	H4EC3	15701	586	-1: DON'T KNOW	159	1.01267	27.1331
S15 SuicSexpSTD	H4SE24	15701	58	-1: DON'T KNOW	15	0.09554	25.8621
S08 HousRostr	H4HR11ML	15701	8	-1: DON'T KNOW	2	0.01274	25.0000
S02 ParentSupp	H4WP19	15701	29	-1: DON'T KNOW	7	0.04458	24.1379
S16 Relatnship	H4TR8	15701	29	-1: DON'T KNOW	7	0.04458	24.1379
S03 RelatnSibs	H4WS3B	15701	188	-1: DON'T KNOW	42	0.26750	22.3404
S02 ParentSupp	H4WP10	15701	2283	-1: DON'T KNOW	510	3.24820	22.3390
S02 ParentSupp	H4WP12	15701	2283	-1: DON'T KNOW	478	3.04439	20.9374
S02 ParentSupp	H4WP11	15701	2283	-1: DON'T KNOW	458	2.91701	20.0613

H4TR22 belongs to Section 16 where respondents report information on their romantic/sexual relationships. There are more than thirty thousand relationships reported (***Sect_nobs*** = 30,263 relationship records). At first sight, an item with a "Don't Know" response rate of 51% seems alarming. Closer examination, however, reveals that this question is asked only of respondents who do not know the exact age of their partners but know that they are not the same age (324 partners). It inquires about how many years older or younger partners are compared to the respondent. In this light, the results of more than half (51%) of the respondents reporting "Don't Know" do not seem surprising.

The other two questions that have 50% "Don't Know" responses are H4HR11MN and H4HR11MO, which belong to a series of questions that ask respondents to report their residence across-state moves since their last interview –

specifically it requests the states' names and when the moves occurred. The questionnaire asks H4TR11MN and H4TR11MO to respondents who had made their 13th or their 14th moves, respectively. Specifically these items measure the months that these respondents made such moves. Considering the contents and contexts of these questions, again it is not surprising that only 2 respondents have valid data for these questions (***Sum_validfreq***=2), and of whom only 1 remembers the exact month of the move (***Count***=1, ***Rel_percent***=50.0). Since the legitimate cases are so few, the absolute percentages for the "Don't Know" responses are really very small (0.00637%). Consequently, the effects of these null responses on the overall data quality of the section, or to the whole survey, are very minimal and non-consequential.

Similar methodology can be used to examine question items with a large percentage of "Refusal" responses. We provide a table below that shows questionnaire items in the Add Health Wave IV interview that have refusal rates of more than 4%.

Section	Name	Sect_ nobs	Sum_ validfreq	Response	Count	Abs_ percent	Rel_ percent
S14 MentlHealth	H4MH11B	15701	101	-2: REFUSAL	32	0.20381	31.6832
S16 Relatnship	H4TR8	15701	29	-2: REFUSAL	7	0.04458	24.1379
S16 Relatnship	H4TR21	30263	660	-2: REFUSAL	144	0.47583	21.8181
S23 TobacoDrug	H4TO93	15701	767	-2: REFUSAL	77	0.49041	10.0391
S16 Relatnship	H4TR22	30263	324	-2: REFUSAL	21	0.06939	6.4417
S19 LiveBirth	H4LB2Y	14749	829	-2: REFUSAL	42	0.28477	5.0663
S19 LiveBirth	H4LB2M	14749	829	-2: REFUSAL	38	0.25764	4.5838
S15 SuicSexpSTD	H4SE37M	15701	23	-2: REFUSAL	1	0.00637	4.3478

H4MH11B belongs to a battery of mental ability tests that asks respondents to repeat some strings of numbers in reverse order. This particular question item is posed to respondents who could not or refused to repeat a prior string of numbers (***Sum_validfreq***=101). Among these 101 respondents, 32 (i.e., 32%) also refuse to repeat in reverse order the string of numbers H4MH11B poses. Hence, the large refusal percentage for this question is as expected.

The question that has the second highest percent of refusals is H4TR8, which is a question addressed to female respondents who refuse a prior question on current pregnancy (***Sum_validfreq***=29). Seven of these female respondents also refuse to answer a subsequent question asking, "Do you think that you are probably pregnant, or not?" Again, it is not surprising to have a high refusal rate for a question directed to a group of respondents who had already refused a prior, similar or related question. Comparable situations are found in H4TR21 and H4TR22. The former applies to respondents who do not know or refuse to report their partner's exact age (H4TR20), asking, "Is your partner younger, older, or the same age as you?" The latter applies to those who only know their partner is younger or older, asking specifically how many years older or younger. Other items that have a high refusal rate tend to be sensitive questions that respondents might not feel comfortable answering. For example, H4TO93 asks respondents to select the one illegal drug s/he uses most often, and H4SE37M asks respondents, who were told they had HIV before, if a health care professional recently informed them again that they had the disease.

2. ANALYSIS AT SECTION LEVEL

Besides using the table-record file for analyzing individual question items, we can also compute summary statistics (e.g., using SUMMARY or MEANS procedures) on these items per section to evaluate survey data quality at the questionnaire section-level. Using Add Health Wave IV interview data as examples, the following table shows the distributions of relative percentages of "Don't Know" responses among items in the 26 questionnaire sections.

Obs	Section Name	Num_ Vars	Mean_ DK_rel_ percent	SD_ DK_rel_ percent	Min_ DK_rel_ percent	Max_ DK_rel_ percent
1	S03 Relationships w/ Siblings	12	13.3894	15.4345	0.00000	35.0000
2	S02 Parental Support/ Relation	44	6.4137	8.2875	0.00000	31.0345
3	S24 Mistreatment by Adults	6	5.1704	4.9541	0.41399	9.8496
4	S16 Relationships	31	3.9207	10.2151	0.00000	51.2346
5	S12 Economics	19	3.5263	6.3665	0.05732	27.1331
6	S19 Live Births	15	3.4133	2.8366	0.16234	9.2220
7	S18 Pregnancy Table	17	3.2303	7.2144	0.06742	30.6981
8	S08 Household Roster/ Residence	106	3.1028	8.2397	0.00000	50.0000
9	S20 Children and Parenting	48	2.0508	1.7488	0.06345	8.2274
10	S04 General Health/ Diet	21	1.7841	4.0274	0.00000	15.7428
11	S15 Suicide/ Sex Experience/ STD	86	1.7081	5.5902	0.00000	40.6250

12	S01 Overview and Demographics	14	1.3080	3.4616	0.00000	11.9565
13	S22 Invlvemt w/ CrmJustc Systm	69	0.8724	0.4858	0.05759	2.1963
14	S17 Relationship in Detail	34	0.6734	0.8810	0.08807	4.0747
15	S23 Tobacco, Alcohol, Drugs	135	0.5125	1.0247	0.00000	10.2999
16	S13 Religion and Spirituality	11	0.4675	0.6689	0.00000	1.9952
17	S25 Daily Activities	34	0.3251	0.9561	0.00000	5.5796
18	S10 Military	30	0.2993	0.4764	0.00000	1.9711
19	S11 Labor Market	33	0.2811	0.5864	0.00000	2.9747
20	S07 Sleep patterns	15	0.1834	0.0811	0.01911	0.2739
21	S05 Health Services/ Insurance	12	0.1699	0.1649	0.01274	0.5773
22	S06 Illnss/ Medicatn/ Disability	65	0.1672	0.3785	0.00000	1.8265
23	S26 Personality	41	0.1249	0.1532	0.04459	0.8535
24	S14 Soc Psychology/MentlHealth	36	0.0790	0.1542	0.00000	0.8990
25	S21 Criminal Offend/ Victmizatn	20	0.0767	0.0432	0.04458	0.2481
26	S09 Education	30	0.0491	0.1495	0.00000	0.7445

Here ***Num_Vars*** indicates the numbers of question items in each section; ***Mean_DK_rel_percent*** is the mean relative percentage of "Don't Know" responses in the specific section; ***SD_DK_rel_percent*** shows the standard deviation of the mean statistics; ***MIN_DK_rel_percent*** and ***MAX_DK_rel_percent*** denote the minimum and the maximum of the distribution, respectively.

Periodic computation of summary statistics such as these alerts us to sections with a relatively high mean or high maximum percent of "Don't Know" responses. This allows us to further examine the items comprising these sections to make sure that we understand the contributing factors behind the response patterns. For example, the high percent of "Don't Know" responses in the top 3 ranking sections here is due to the abundance of questions that ask for date information on events that occur during a respondent's childhood (e.g., month/year when a certain sibling died; when a biological mom/dad died; dates of parental incarceration; when various types of mistreatment or abuses by adults first occurred; etc.). The respondents might truly not know or not recall the exact time when these events happened.

Similar analysis can be done on the mean relative percentages of "Refusal" responses per questionnaire section. Here, as an illustration, we show those sections with mean statistics exceeding 1 percent and those with less than 0.02 percent from the Add Health Wave IV survey.

Obs	Section Name	Num_ Vars	Mean_ RF_rel_ percent	SD_ RF_rel_ percent	Min_ RF_rel_ percent	Max_ RF_rel_ percent
1	S16 Relationships	31	2.12226	5.69077	0.00000	24.1379
2	S20 Children and Parenting	48	1.57101	1.43531	0.10156	3.9059
3	S19 Live Births	15	1.55515	1.67758	0.16234	5.0663
4	S14 Soc Psychology/ Mentl Health	36	1.16201	5.25433	0.00000	31.6832
5	S22 Invlvemt w/ CrmJustc Systm	69	1.10308	0.51216	0.08955	1.8357
6	S18 Pregnancy Table	17	1.04055	0.99536	0.04494	3.9648
22	S09 Education	30	0.01863	0.06442	0.00000	0.3460
23	S07 Sleep patterns	15	0.01826	0.00989	0.00000	0.0255
24	S06 Illnss/ Medicatn/ Disability	65	0.01161	0.02232	0.00000	0.1508
25	S05 Health Services/ Insurance	12	0.00731	0.01051	0.00000	0.0304
26	S03 Relationships w/ Siblings	12	0.00595	0.00939	0.00000	0.0263

The higher percentages of "Refusal" responses are expected among items in the sections that ask private and sensitive questions regarding one's sexual/romantic relationships, pregnancies, live birth and fertility histories, or one's involvements with crime or legal offenses. The higher percentages of "Refusal" responses in the Social Psychology/Mental Health section are, to a large extent, due to questionnaire design. In order to estimate respondents' mental ability, the questionnaire includes a battery of test items designed to be progressively more difficult. Respondents who refuse, do not know, or fail to answer correctly a prior item are asked another test item that has the same degree of difficulty. Hence, it is not surprising that these items have a consistently higher refusal rate than others.

The advantage of examining *relative* percentages of "Don't Know" or "Refusal" responses is that these percentages are based on the legitimate cases *per individual question item*, and they may indicate the extent of difficulty or ease

with which respondents answer specific questions. However, if we want to gauge the overall data quality of the whole questionnaire or per questionnaire section, we might want to factor in the number of applicable cases per item. In other words, we want to give differential weights to questions proportional to the number of legitimate cases applied to them (thus their response patterns). Hence, we can use the *absolute* percentages (based on legitimate cases per section) for the summary statistics computation. We can compare the average *absolute* percentages of "Don't Know" or "Refusal" separately across sections. We can also combine the "Don't Know" and "Refusal" percentages and evaluate data quality across sections by the average percentages of these invalid or null responses.

From the table below, we can see that the section on Economics (asking respondents their amounts of: household income, personal earnings, mortgages, and financial loans/gifts etc.) has the highest absolute percentages of "Don't Know" and "Refusal" among all questionnaire sections -- even higher than the sections asking respondents sensitive and private questions about romantic/sexual relationships, pregnancies, live births, children, and their childhood abusive experiences. In contrast, we can see that respondents are more willing and/or found questions easier to answer regarding their social and household demographics, their physical health and medical accessibility, and their military or educational experiences. The average absolute percentages for these questionnaire sections are less than 0.1 percent, which indicate the good quality of data reported.

Obs	Section Name	Num_ Vars	Mean_ DKRF_abs percent	SD_ DKRF_abs percent	Min_ DKRF_abs percent	Max_ DKRF_abs percent
1	S12 Economics	19	2.05150	3.09721	0.08917	10.3178
2	S19 Live Births	15	1.88849	2.30540	0.01356	7.2276
3	S24 Mistreatment by Adults	6	1.88205	1.61184	0.56684	5.0379
4	S18 Pregnancy Table	17	1.34453	1.94205	0.04552	6.4779
5	S20 Children and Parenting	48	1.34043	0.67808	0.10994	2.0202
6	S16 Relationships	31	1.27832	2.23025	0.00000	8.3852
21	S08 Household Roster/ Residence	106	0.08484	0.33717	0.00000	2.6304
22	S05 Health Services/ Insurance	12	0.08014	0.06045	0.01274	0.2038
23	S01 Overview and Demographics	14	0.05596	0.14782	0.00000	0.5095
24	S09 Education	30	0.03567	0.12230	0.00000	0.6433
25	S10 Military	30	0.02208	0.02871	0.00000	0.1083
26	S06 Illnss/ Medicatn/ Disability	65	0.01930	0.02229	0.00000	0.1274

CONCLUSION

Utilizing simple SAS macros and the CONTENTS, TRANSPOSE, FREQ, APPEND, and SUMMARY procedures, we have successfully created a powerful means of assessing data quality by tracking rates of "Don't Know," "Refusal," "Legitimate Skip," and "Valid" response in the Add Health Wave IV interview. With minimal input from its users, one run of our basic program produces the data set from which statistics describing data quality will be generated . In this basic program, PROC CONTENTS provides the variables of interest, while PROC TRANSPOSE rearranges the output data generated by the CONTENTS procedure to facilitate passing these variables to PROC FREQ one at a time. PROC APPEND combines the output data sets produced by the FREQ procedures into one manageable file, which is then thoroughly described by PROC SUMMARY. The summary statistics produced allow us to easily pinpoint and examine from different angles both questionnaire items and sections with high rates of non-response. This informs us as to whether troublingly high rates necessitate redesign of our instruments or, perhaps more importantly, are reasonable when considering item subject contents and contexts.

ACKNOWLEDGEMENTS

This paper uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors via email at:

Mariah Mantsun Cheng, Ph.D.
UNC Carolina Population Center
CB# 8120
Chapel Hill, NC 27516 USA

Email: Mariah_cheng@unc.edu

Timothy Monbureau, MS
UNC Carolina Population Center
CB# 8120
Chapel Hill, NC 27516 USA
Email: tim_monbureau@unc.edu