

National Human Exposure Assessment Survey (NHEXAS)

Maryland Study

Quality Systems and Implementation Plan for Human Exposure Assessment

Emory University
Atlanta, GA 30322

Cooperative Agreement CR 822038

Standard Operating Procedure

NHX/SOP-D02

Title: Questionnaire Data Entry and Preparation

Source: Harvard University/Johns Hopkins University

U.S. Environmental Protection Agency
Office of Research and Development
Human Exposure & Atmospheric Sciences Division
Human Exposure Research Branch

Notice: The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here. This protocol is part of the Quality Systems Implementation Plan (QSIP) that was reviewed by the EPA and approved for use in this demonstration/scoping study. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

1. Title of Standard Operating Procedure

Harvard University/Emory University/Johns Hopkins University Standard Operating Procedures:
D02 Questionnaire Data Entry and Preparation, Rev 1.0

2. Overview and Purpose

The purpose of this SOP is to describe how data is transcribed from the questionnaire forms and entered into the official database also known as the Complete Dataset (CDS). This SOP also describes the structure and purpose of the CDS and the Analysis-ready Dataset (ADS) and how data is entered into the ADS.

3. Discussion

Data will be entered into the CDS and ADC incrementally. To help ensure that all data is accounted for during the data entry process each participant's questionnaires will be assigned to a *unit* of data. A unit of data will be handled together and processed as a unit. All questionnaires collected from all individuals who are participating during a specified interval of time will be defined as a unit of questionnaire data.

The size of a unit of questionnaire data will be controlled by the Project Data Coordinator. Number of subcycles will be used as the measure defining the number of participants included in a unit of data. This will give some consistency to the scheduling of data management. For example, a unit of questionnaire data may be all participants in the first two-week subcycle of the study. For the first two or three Cycles the unit of questionnaire data will be defined as a two week subcycle worth of questionnaires. After the first two or three Cycles the size of a unit of data may be increased to 3 subcycles (a Cycle) because the data collection methods will have been checked for irregularities and lab results will be returning, adding to the data management burden.

Structure of the CDS.

The CDS is the official database. The CDS will be used to create any database used for analysis, e.g. used to construct the ADS, and will be the database sent when data is requested from other investigators. There are three time scales on which questionnaire data is collected: questionnaires may only be asked once, questionnaires may be asked each cycle, and information may be asked for each day of the cycle. Ideally, each questionnaire is stored in a separate table. There are some exceptions to this. For example, the follow-up questionnaire is too large to fit into one table and the information about individuals living in the household is more efficiently stored in a separate table from the rest of the descriptive questionnaire.

The CDS will reflect exactly what was written on the questionnaires except in one respect. Many questions are asked only if a certain response to a previous question was given. These questions are skipped during the interview, because the question is irrelevant or the answer is obvious given the answer from the previous questions. For those questions which are skipped where an answer may be inferred from previous responses, the inferred answer is included in the CDS. A list of inferred questions and the inferred answer are given in Appendix B. The inferred answers will be automatically included in the initial data files returned from data entry.

Structure of the ADS

The data frame of interest is the combination participant ID-cycle Number. The ADS is a predesigned set of adjustments to the CDS so that all data is on the same framework and is in a form which is ready for analysis. There are three types of adjustments made: summarizing, deletion, and rearrangement. Many variables will need to be summarized before they can be used in analysis. The most obvious examples are activity and food data which are collected each day of the cycle. This information needs to be averaged over the cycle to create variables on the framework of interest. Some variables may not be needed or wanted in the ADS, the confidential questions are an example. A list of these transformations is given in Appendix B.

Since most analyses will focus on certain aspects of the data collected, *e.g.* food intake or activity patterns, the questions can be rearranged to group all topically related questions into the same table. The questions have been reorganized into five major groups: demographics, food, activity, physical housing characteristics, and living arrangements. These adjustments are not meant to be conclusive and all encompassing, but should cover most of the general research questions of interest at the beginning of the study.

These transformations of the CDS are stored as Paradox queries. This makes creation and updating of the ADS a simple procedure which can be repeated quickly and easily as new data arrives and is entered into the CDS. Typically, links between the tables in the ADS will be on the participant-cycle level. The ADS is derived from the official database CDS, but should not be construed as an official database.

File naming conventions

The responses for the different questionnaires are placed in separate data files. Templates of these data files have already been created. The names and contents of the templates are given in Appendix A of this SOP. As data enters the system, updated versions of the data files will be created. Each time a data file is updated a new file will be created. The new file will be a copy of the last version of the data file plus the new data entered. The new file will have the same name as the template plus the number of the update. For example the activity questionnaire has the template name *activ.db* and when the first data is entered into the data file it will be named *activ1.db*. When *activ1.db* is updated a new file will be created and named *activ2.db*.

If a correction needs to be made to a data file, *e.g.* *activ1.db*, a new file will be created which contains the corrected version of the data file. The file name of the corrected version of the data file will end with a letter. For example, if corrections need to be made to *activ1.db* a copy of *activ1.db* will be made and named *activ1a.db*. If more corrections need to be made before *activ2.db* has been created then *activ1a.db* will be copied and named *activ1b.db*.

4. Personnel Responsibilities

4.1 Project Data Coordinator is responsible for:

- assuring accuracy and consistency of database
- modifying data entry protocol if necessary, in writing
- tracking questionnaires from collection to entry into database to storage.
- notifying the Data Entry Supervisor of any modifications that affect working databases.

4.2 Data Entry Supervisor is responsible for

- training and supervising Data Entry Assistants
- tracking questionnaires and other forms through data entry and review
- resolving problems and ambiguities that Assistants are unable to handle.

4.3 Data Input Assistants are responsible for

- logging in arriving data, checking for ID, storing questionnaires
- data entry
- reviewing data entry.

5. Required Equipment and Reagents

set of questionnaires and pen for marking the questionnaires.

computer with Borland Paradox V5.0 for Windows and database directories

data in magnetic format returned from the Data Entry Company

data in hardcopy format

CDS backup disks.

data entry instruction sheet appropriate to the type of data being entered

data entry log (forms with initials and dates for data entry; kept in Supervisor's logbook)

6. Procedures

Marking the Questionnaires

The questionnaires will be marked with a black or blue ink pen by the Field or Telephone Interviewer, depending on the questionnaire. The response will be circled by the Interviewer. If a change needs to be made in the marked response, the incorrect mark will be crossed out and initialed. The correct response will then be marked. The Field Interviewer will give these instructions to the participants for questionnaires they mark.

Processing Questionnaires at the FCC

The confidential questions will be separated from the questionnaires by the FCC Clerk on return from the field. They will be stored until a unit of data has been collected. See the discussion section for the definition of a unit of questionnaires and SOP D01 for details on storage procedures. When a unit of questionnaires has been collected the questionnaires minus confidential questions will be photocopied twice and the confidential questions photocopied once by the FCC Clerk. The FCC Clerk will then send the originals of the questionnaires and confidential questions to Emory, one photocopy of the questionnaire to the Data Entry Company. The other photocopies will be returned to storage location described in SOP D01. The FCC Clerk will notify Emory via e-mail which questionnaires have been shipped.

Data Preparation at Emory

On receipt of original questionnaires and confidential questions the Data Entry Assistant will store them as described in SOP D01.

On receipt of each data file from the Data Entry Company the Data Entry Assistant will take the following steps to create the CDS and store new data:

1. The file will be copied to the archive directory. The file will be renamed if necessary to reflect the NHEXAS file naming standards.

2. If the file is renamed the file will be copied onto the same disk as the original with the new name for cross-reference.
3. The disk with the original data will be stored in the designated location for archived disks. See SOP D01 for more details.
4. The data will be imported to the relevant table of the CDS and the resulting file will be renamed to identify it as the latest version of that table. Details are given below.
5. The new CDS will be copied to the CDS back-up disk. See SOP D01 for more details.
6. The old version of the CDS will be removed from the CDS directory so only the current version of the CDS is available.
7. A record of completion will be entered into the data entry logbook.

Step 4: Entering new data units to the CDS

New records are added to copies of the latest version of the tables by importing the new records into a template table and then using the Add/Append functions to put the results temporarily stored in the template table into the new version table. An introduction to these methods can be found on p. 180 and p. 187 of the *Paradox* User's Guide.

The procedure will be the same for all tables. The steps in preparing the CDS tables for questionnaire are:

1. Gather all new units of initial data files to be entered. Then for each new unit of data:
2. Rename a copy of the last version of the CDS table to identify it as the latest version of the CDS table.
3. Import the new data to the relevant template table by using the Tools|Utilities|Import commands.
4. Take the newly created table from Step 3 and use the Tools|Utilities|Add command to add the new table to the latest version of the relevant CDS table created in Step 2.
5. Clear the template table by using the Table|Empty command.

Correcting the CDS

See Quality Assurance Procedures for methods used to detect errors in the CDS. If an error is observed by the Data Entry Assistant, they will notify the Data Entry Supervisor who will notify the Project Data Coordinator. The Project Data Coordinator will decide whether the correction will be made and will notify the Principle Investigator about the decision. If the CDS needs to be corrected a renamed copy of the CDS will be made as described in the File Naming subsection of the discussion above. An example: if the CDS name is techw4.db then the renamed copy will be techw4a.db. The Data Entry Assistant will make the correction directly to the renamed CDS table and the nature of the correction and the reason for the error if known will be noted in the file qxcom.db. The updated file and the questionnaire comment file will be copied and stored in the designated official database location. The old version of the CDS will then be archived and removed from the CDS directory. See SOP D01 for storage details.

Creating the ADS

For each ADS table there is a short series of queries which need to be run by Data Entry Assistant at the end of the first subcycle of every Cycle, except Cycle 1. The number of queries is dependent on the ADS table being constructed. The queries are already constructed and have names demox.qbe, foodx.qbe, housex.qbe, and livingx.qbe, where *x* is the number in the series of queries for the respective table being created.

Starting from the CDS the steps in preparing the question tables for ADS are the same for each ADS table:

1. Using the latest version of all CDS files start with query 1 for the ADS file being constructed. Repeat for each query in the series running the queries in increasing numeric order.
2. After all queries have been run rename the final answer table to reflect the version of the ADS created. See the section on file naming conventions in the discussion.
3. Remove the intermediary files which are created during the process.

7. Quality Assurance Procedures

All personnel have appropriate training, see SOP D04.

Quality assurance procedures used by the Data Entry Company are the following:

[RB: looking for some fill in from Data Entry Company if we can get it]

1. Double entry techniques to ensure the correct entry of the questionnaires.
2. Use of Range limits on field inputs to reduce opportunity for incorrect entry.

The ranges used by the Data Entry Company will be specified by the Project Data Coordinator and will be used in the CDS tables. If an imported value is outside the stated ranges the Data Entry Assistant will automatically be notified by the *Paradox* software and the Data Entry Assistant will investigate the nature of the error and appropriate action will be taken. See SOP G06 on Problem Management.

To guarantee no complete records have been missed, the Data Entry Assistant will check the number of records in the initial data file, which contains the new data, and the number of records in the old version of the official database to make sure the sum equals the number of records in the new version of the database. Discrepancies will be checked by looking directly at the initial data file sent by the Data Entry Company to see which records may have been lost or duplicated. If no error is observed the Data Entry Assistant will check to make sure that all old records were correctly incorporated into the new file.

Spot checks of the questionnaires will also be carried out. See SOP D05 for details.

At the conclusion of the study, a comparison of previous versions of the official database with the final version of the official database will be undertaken to check for consistency in database creation. Any discrepancies will be checked.

8. References

Borland Paradox Relational Database V5.0 User's Guide and Online Help.

Harvard University/Emory University/Johns Hopkins University Standard Operating Procedures:

D01 Data Flow Procedures, Rev 1.0

D03 Lab Results Data Entry and Preparation, Rev 1.0

D05 Exploratory Data Analysis and Summary Statistics, Rev 1.0

Appendix A - List of Questionnaire Data files and description of contents

The CDS files

activ.db -	Activity Questionnaire
addgr.db -	List of grocery stores from additional questions.
addqx.db -	Additional information about living arrangements in household.
basea.db -	Baseline questionnaire, except confidential questions, part 1.
baseb.db -	Baseline questionnaire, except confidential questions, part 2.
conf.db -	Confidential Questions T5, D10b, D10d.
confb.db -	Confidential Questions from baseline questionnaire.
confn.db -	Confidential Question D6a, names of individuals living in the household.
descr.db -	Descriptive Questionnaire, except D5 and confidential questions.
desci.db -	Descriptive question D6, information on individuals in household.
follow.db -	Follow-up questionnaire.
folllb.db -	Follow-up questionnaire Part B (can't get in one table).
foodf.db -	Food Diary Followup Questionnaire.
foode.db -	Food Diary Checklist.
parti.db -	Participant list contains participant ID and cycle.
techw.db -	Technician walk through questionnaire, except T6, T12.
techf.db -	Technician walk through question T12, information on carpets.
visit.db -	Names of technicians and sampling notes.

The ADS files

activa.db -	Activity related questions.
demo.db -	Demographics questions.
fooda. -	Food related questions.
house.db -	Physical housing characteristics.
living.db -	Living arrangement characteristics.

Appendix B - Set of Confidential Questions

Each questionnaire has an initial page which contains the name, address and phone number of the participant. This information is not going to added to the data set. The questions are:

- D1: Address
- D5a,b: Names of Individuals in household
- D10: Telephone number
- B21: Health information.
- B44: Income levels.
- F6: Medications: prescription
- F7: Medications: non-prescription.
- F8: Pregnancy
- T5: Indicate nearest major intersection.

Appendix C - Transformations of Questions Before Data Analysis

Many questions in the questionnaire are not asked if certain answers are given to a previous question. These questions will be referred to as conditional questions. In nearly all cases the answers to these questions can be inferred from the question which was answered. To reduce the number of missing values in the data set the inferred responses will be entered into the database. A second major source of alterations in the data are multiple response questions. These are questions that have several choices and the respondent is asked to mark all that apply. These questions are changed by assuming that the question was asked for each choice separately and the response is yes or no. In the data file this expands the number of variables by creating a variable for each choice in the multiple response question. This appendix outlines the changes that need to be made to the conditional questions asked in the different questionnaires. The questions are divided by questionnaire and the answered question is stated first, followed by the answer which leads to the conditional question(s) being skipped and the suggested answer.

MV=missing value, DK =Don't Know

Descriptive Questionnaire

- D5. Gives information about all individuals living in the residence. Suggested summary variables:
 1. Average or median age. Also maybe youngest and oldest.
 2. Dominant ethnic group (in numbers) also variables indicating presence or absence of different ethnic groups.
 3. Dominant schooling level (in numbers) also variables indicating presence or absence of different schooling levels.
 4. Number of smokers
 5. Total person hours at home (represents dust stirring activity)
 6. total number of individuals who use chemicals/pesticides

Baseline Questionnaire

- B6a. Yes answer implies Yes to B6b.
- B6a and B6b could be combined into do you now or ever have smoked regularly? Then B6c can be rephrased to be "When was the last time tobacco used" (answer of zero means currently in use.)
- B7a-d. For non-smokers answer of zero is implied.
- B13 Major conditional response. B14-B16 answered only if B13 is positive. Attempt to use B14-B16 will require use of subset of sample.
- B14b-d. Three verbal responses. Will need to be categorized. Can we do it ahead of time. Particularly B14b,c.
- B14e answered *no* implies B14f is *no* for the five categories. *DK* is missing value.
- B14g answered *no* implies B14h is *no* for the six categories. *DK* is missing value.
- B14i answered *no* implies B14j is *no* for the six categories. *DK* is missing value.
- B14k answered *no* implies B14l is *no* for the eight categories. *DK* is missing value.
- B14m answered *no* implies B14n is *no* for the three categories. *DK* is missing value.
- B15-B16 are identical to B13-B14 in construction.
- B17 is identical to B13, B14a.
- B18 for children under 6. MV for others, unless B18a combined with B14a, B16a, and B17b to generate an answer to "How many hours were you out of the house due to work or school?"
- B27a answer of *no* should be placed in new category of B27b *no attached garage*.

- B27c can be combined into B27b to create compound categories B27b categories with and without doorway.
- B27a answer of *no* implies B27d answer of *no*.
- B29a answer of *no* should be placed in new category of B29b *none*.
- B29c is MV for those with no air conditioning.
- If B34a answer is *no* then B34b is zero and B34c has new category *none*.
- Similarly for B35 - B37.
- B36d, B37d also need to have category *none* added when answer to B36a, B37a is *no*.
- B38a answer of *no* implies B38b, B38c new categories of *none* and B38d, i-iii) are *zero*.
- B38e will be MV for those not using pesticides in last 6 months.
- B38g, h will be MV for those not using pesticides in last 6 months.
- B39 is similar to B38.
- B40 similar to B38.
- B43a answer of *no* implies B43b-B43d are *zero* and B43e is *no*. B43f is MV.

Time Diary and Activity Questionnaire

Two sections to this questionnaire: location and activity.

- Seven categories (Indoor/Outdoor) by (Home/Work/Other) and In transit which will be summarized by averaging daily location over week observations collected.
 - Also summarize by Total Indoor, Total Outdoor, Total Home, etc.
- There are 29 activity questions. The number of items could be reduced by summarizing related questions. Suggested groups are:
- A1-A3: Auto fuel related.
 - A4, A5: Soil related.
 - A6-A9, A15-A17, A20, A21: Smoke inhalation (not necessarily cigarettes, etc.)
 - A10-A11, A14, A18, A22: water exposure related.
 - A12, A13: pesticide related.
 - A19, A23-A26: are not related to any other questions.
 - A27-A28: Exercise questions. Compute percent of time spent at levels of exercise.



Technician Walk-thru Questionnaire


- T2 = T1 if do not live in multi-family building.
- T4 has verbal comments. How/whether to include in data set?
- T6a multiple response with 6 levels generates 6 yes or no questions.
- T6b cut off at 300 feet, use 300 feet as answer.
- T6c multiple response again indicator variables needed for each possible response.
- T6f same as T6c
- T6g. Response 1 and 3 combined by stating that at wall is same as 0 feet from wall.
- T6h-T6j same argument as T6c.
- T7a *no* implies T7b is *no* for multiple responses 1 and 2 out of 2.
- T8 is same as T7.
- T11 is table of floor and floor cleaning info. Summarize into total floor space, and percent of each type of surface. Date and method of cleaning separated into numerical and multiple choice questions. Last question asks whether anyone uses the floor, not necessarily the participant.




Followup Questionnaire

- F1b1 is MV if F1b is *no*.
- F1g-F1o. Variable in data set will be “Number of days used” and initial answer of *no* to any of these questions will imply answer of zero to “Number of days used”. Side questions will be associated with F1g-F1o. Answer of *no* to side questions for F1g-F1k implies answer of *none*, a new category for second side question. If not used, then answer of *no* to side question is implied.
- F2 a series of compound questions. For each question in series an answer of *no* to initial question implies zero to next question, rest of questions will get MV as phrased.
- F3-F4 similar to F2 except likely to get a larger number of respondents performing these activities.
- If male answer to F8 is implied to be *no*. This is legitimate because two variables (sex, pregnancy) creates three categories (pregnant/not pregnant females and men).
- F10 answer of *no* implies answer of *no* for all parts of F11.
- F12 a series of multiple choice questions. Only asked once?

Food Diary

- Variable number of items with quantities attached. Summarize and categorize items, such as fruit, vegetables, etc. and sum (avg.?) items consumed by categories.
-  4 day checklist, summarize by categories etc. Need to choose level of category detail. The general food groups are reasonable, and would suggest the following sub-categories:
1. Fresh or frozen vs. canned fruits (not including juices made at home)
 2. Fresh or frozen vs. canned vegetables. (not including juices made at home)
 3. leafy vs. non-leafy vegetables.
 4. Fresh or frozen vs. canned meats.
 5. Categories by meat type, poultry, beef, pork, fish and/or shellfish.
 6. breakfast cereals vs. non-breakfast cereals.
 7. grain type: corn, wheat, oat, potatoes, other.
 8. Beverages: including water from home vs. no water from home. (include frozen fruit and vegetable juices here)
 9. Home-baked vs. store bought sweets, etc.
 10. Sweets, etc. vs. Misc.

Food Diary Followup

- FD2, FD5, and FD8 are conditional on the meal being eaten. Since it is probably best to summarize the info in these questions by summing number of times prepared and eaten at each location then An answer of *no* to FD1, FD4, or FD7 implies zero for daily categories in FD2, FD5, and FD8 respectively.
 - Missing food information. Need to set a cut off value to define valid measurement. For example, at least 75% of food must have been collected otherwise it is determined to be a missing value.
 - FD10 can be summarized over all meals.
-  FD11 give each category the values -1, 0, and 1 for less, the same and more food and sum over days. Then matrix multiply FD11 by indicator values of FD12 questions and sum over days. For example, FD11=(-1,0,1,0) for the 4 days and FD12a is (1,0,0,1) then value for FD11 = 0 and FD12a = -1. So average over 4 days they ate about normal amount of food, but under average variability due to travel. The vector for FD12a could be weighted to handle multiple responses on a single day. If FD11 is *same as usual* implies FD12a-h are zero.
- FD13 and FD14 have similar problem as FD11 and FD12. It will be handled in similar fashion, except an answer of *yes* on any day implies an answer of *yes* for whole week.
- 