National Human Exposure Assessment Survey (NHEXAS)

Arizona Study

Quality Systems and Implementation Plan for Human Exposure Assessment

The University of Arizona
Tucson, Arizona 85721

Cooperative Agreement CR 821560

Standard Operating Procedure   **SOP-UA-D-5.0**

Title: Global Coding Used by NHEXAS Arizona (Hand Entry)

Source: The University of Arizona

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

Global Coding Used by NHEXAS Arizona

1.0 Purpose and Applicability

The purpose of this procedure is to define the global coding scheme to be used in all NHEXAS working and master databases.

2.0 Definitions

2.1 DATA CLEANING = The process of locating and correcting data processing errors. They can be individual level errors in the electronic and physical data, or they can be system level errors in the data collection, packaging, coding, entry, and cleaning procedures themselves. This process is also referred to as "data validation."

2.2 DATA FIELD = An area on a data entry form (i.e., screen) where datum from a physical form is entered.

2.3 DATABASE, MASTER = Accumulative database generated from validated data processing batches. Newly cleaned batches are appended to the master database. Copies of this database are used in analyses. All corrections made to copies of the master are made to the master database itself. Thus, it is the most complete and accurate database of its kind.

2.4 DATABASE, WORKING = A database earmarked for or in the process of cleaning that contains one or more data processing batches. When cleaned, this will be appended to the master database.

2.5 FORM, PHYSICAL [DATA] = The paper or "hard copy" version of a data form. This is also referred to as a "physical data form."

2.6 GLOBAL CODING = A set of standard codes used in data within the Respiratory Sciences Center designating the status of a data field in three cases: (1) datum refused, (2) datum non-applicable, and (3) datum missing. It can also be a standard coding approach that pertains to questions with the same response structure.

2.7 NHEXAS Arizona: Acronym for National Human EXposure Assessment Survey, a research project conducted in Arizona by the University of Arizona/Battelle/Illinois Institute of Technology consortium.

3.0 References

N/A

4.0 Discussion

First, global coding schemes can be divided into two types: field-dependant, and fixed. Field-dependant global coding schemes are used in

three situations: (1) subject refusals, (2) non-applicable data, and (3) missing data. Codes in these cases are "field-dependent" because the exact value of the global code depends on the length and type of the data field receiving the code. In particular, if the datum is **refused** by the subject, then the data field receives a code containing one or more **fives** (or a combination of fives and zeros if it is a date or time field). If the datum is **non-applicable**, then the data field receives a code containing one or more **eights** (or a combination of eights and zeros if it is a date or time field). Finally, if the datum is **missing**, then the data field receives a code containing one or more **nines** (or a combination of nines and zeros if it is a date or time field).

Examples:

(1)   A subject refused to reveal the annual income of his or her residence. The data field for the variable *INCOME*, which is defined as a three-character integer, would receive a code of *555* for refused.

(2)   A timer malfunction occurred while sampling in the field and the field technician could not obtain the actual stop time. The data field for the variable *ENDTIME*, which is defined as an eight-character time field, would receive a code of *09:09:09* for missing.

Fixed global coding schemes, on the other hand, are independent of the length and type of the data field receiving the code. They are simply standard manners of approaching the coding of questions that have the same response structure (see examples below in section 7.3).

Second, the global coding scheme, whether field-dependent or fixed, follows the principle of mutual exclusivity. This means that fields receive codes that are never within or near the valid range for real data. For example, lab data are coded with negative global codes (eg., -55, -88, or -99). Although data entry is slower with negative codes, their use greatly decreases the chance that global codes are inadvertantly included in calculations; further, if they were included, then detection of this mistake is very likely. Questionnaire data, in contrast, receive positive global codes in order to speed up data entry. The codes are one to two characters greater than the highest valid response value for the question, however (eg., for a question that has a one- or two-character response value, the global codes would be 555, 888, or 999). It is thus very unlikely that the global codes would be confused with real responses.

Finally, certain constraints inherent in our primary database software (*R:BASE for DOS version 2.11*) as well as our customized data entry and cleaning program (*Data.exe*) have shaped our global coding scheme. With respect to *R:BASE for DOS*, it only accepts valid dates and times. One cannot, for instance, enter 99/99/99 for a date or 99:99:99 for a time. With respect to *Data.exe*, it has a built-in feature for automatic global coding of dates and times during data entry. If one fills a date or time field with either fives, eights, or nines, then it will convert them to the global code. For example, if one enters 99999999 into a time field, then it will be converted to 09:09:09. Given this, any ex-

ceptions to the global coding scheme will be handled on a case by case basis and will be documented in the appropriate coding instructions when *Data.exe* is used for data entry.

5.0 Responsibilities

5.1 The Project Data Coordinator is responsible for (1) ensuring consistency of global coding throughout all working databases used in the NHEXAS project; (2) modifying global coding protocol if and when necessary, in written form; (3) notifying the Student Data Assitant of any modifications that affect working databases; and (4) notifying the Project Data Analyst of any modifications that affect master databases.

6.0 Materials and Reagents

N/A

7.0 Procedure

7.1 Criteria for Using Field-Dependent Global Codes

7.1.1 When to Code Data Field as Refused

   (a) Subject has crossed out question or field technician has indicated that subject refused the question.
   (b) Other source(s) indicates that the question, physical form, or questionnaire was refused.

7.1.2 When to Code Data Field as Non-Applicable

   (a) Field technician has written "N/A" on the question, physical form, or questionnaire.
   (b) Sample cannot be taken due to the subject's particular situation. For example, a soil sample cannot be taken in a subject's yard if all of it is paved with cement.

7.1.3 When to Code Data Field as Missing

   (a) The sampler, questionnaire, or datum should have been taken, administered, or gathered according to the standard operating procedure, but was not.
   (b) The sampler or questionnaire was lost prior to data entry.
   (c) The sampling technique or question was determined to be irrevocably flawed.

7.2 Field-Dependent Global Coding Scheme: Follow the coding scheme given for the various data types listed below.

7.2.1 Date Fields

   (a) Refusal: 05/05/05
   (b) Non-applicable: 08/08/08
   (c) Missing: 09/09/09

7.2.2 Time Fields

Note that when real times are recorded to the nearest minute rather than to the nearest second, eight-character global time codes will never be mistaken for real times because they are to the nearest second. If, on the other hand, it is necessary to use a five-character time field for a real time recorded to the nearest minute, then the hour in the global code must be **00**; otherwise, it would be mistaken for a real time.

    (a) Refusal
        1. Eight-character field: 05:05:05
        2. Five-character field: 00:05

    (b) Non-applicable
        1. Eight-character field: 08:08:08
        2. Five-character field: 00:08

    (c) Missing
        1. Eight-character field: 09:09:09
        2. Five-character field: 00:09

7.2.3 Integer and Real Number Fields

    (a) Questionniares and non-calculated fields: Entire field on data entry screen is filled with fives for a refused, eights for non-applicable, or nines for missing.

    (b) Lab data: The first character of the field is a minus sign; the remaining character(s) is filled with fives for refused, eights for non-applicable, or nines for missing.

7.2.4 Alpha-Numeric Fields

In all cases, the entire field on data entry screen is filled with fives for refused, eights for non-applicable, or nines for missing.

7.3 Fixed Global Coding Scheme: Follow the coding scheme given for the various response structures listed below.

7.3.1 Boxes ([ ]) Checked by Subject or by Field Technician

Note that the distinction between box not checked, box non-applicable, and box datum missing can be determined by the information given in the greater context of the question.

    (a) Box checked   ([ ]) = 1
    (b) Box not checked   ([ ]) = 2
    (c) Box item refused   ([ ]) = 5
    (d) Box item N/A   ([ ]) = 8
    (e) Box item missing   ([ ]) = 9

7.3.2 Circled Response Categories

    (a) For numbered response categories, the number preceding the category is coded with one to two leading zeros, and then that number is entered into the database.

    (b) For lettered response categories, the letter circled is entered into the database.

    (c) Any exceptions will be noted in the coding instructions for the physical form in question.

7.3.3 Fill In the Blank

    (a) Numerical responses are entered as they appear.

    (b) Time fields are coded with the European or military time equivalent and then entered as coded.

    (c) Non-numerical responses (including responses to "Other (please specify):_____") are assigned a code from pre-established coding lists.

    (d) Any exceptions will be noted in the coding instructions for the physical form in question.

7.4 Quality Control

7.4.1 Tolerance Limits

N/A

7.4.2 Detection Limits

The Project Data Coordinator ensures global coding consistency throughout all NHEXAS working databases through the quality assurance checks outlined in SOP# UA-D-16.0 and SOP# UA-D-27.0.

7.4.3 Corrective Actions

Any discovered inconsistencies in global coding will be addressed and resolved by the Project Data Coordinator.

8.0 **Records**

Any updates or exceptions to the global coding scheme will be documented and filed in the "Coding Instructions: NHEXAS Phase" notebook, located in room 128 of Health Related Professions, 1435 North Fremont Avenue, Tucson, AZ 85719.