

THE COMPUTER ASSISTED QUESTIONNAIRE AND DATASET DEVELOPER

**Shobana Raghupathy, Ph.D.,
James L. Peterson, Ph.D.,**

**Sociometrics Corporation
170 State St., Suite 260
Los Altos, CA 94022**

**Ph: 650 949 3282 *209
Shobana@socio.com**

THE COMPUTER ASSISTED QUESTIONNAIRE AND DATASET DEVELOPER

ABSTRACT

The Computer Assisted Questionnaire and Dataset Developer is a research tool developed by Sociometrics Corp., that is aimed at facilitating the development and documentation of data collection instruments, as well as data entry and error checking. The software stores information entered by the researcher in a database, and then draws on it to (1) produce a fully formatted questionnaire in print format; (2) produce a printed codebook, flow chart, and data file map; (3) provide for data entry from completed questionnaires, with simultaneous error checking; (4) produce a raw data file in ASCII format, and (5) build the machine-readable program statement files needed to transform the raw data file into SPSS and/or SAS system files. The database is fully editable, permitting the user to modify, update, correct, or restructure the study (including by reordering questions or sections) at any time. The software automates tasks best done by computer, improves instrumentation and documentation by providing a complete, high-quality structure and format, and reduces the post data-collection effort of documenting a public-use data set. This greatly eases the dataset development process—from the questionnaire creation through the data entry process—by automating and standardizing the entire process. This minimizes errors in the data entry and processing, thereby saving valuable time and effort for the researcher.

The software also includes an item bank of some 8000 variables from premier studies in the family research field. Researchers may search the item bank for questions suitable to their research purposes, and drop them into their own questionnaires. This facilitates the use of proven questions, enabling users to draw on the best of the extant research rather than inventing new questions.

SIGNIFICANCE

In a typical social science research project, researchers need to devote a considerable amount of time and energy towards the *front-end of the research process* (such as survey development and documentation, data collection, as well as data processing). This involves drafting and continually revising questionnaires or interview forms; processing data once it has been collected; entering the data into a database; and checking the data to remove errors. In addition, researchers need to prepare the data for analysis by generating codebooks and creating analysis files for statistical software programs such as SAS and SPSS. Much of this work is tedious and repetitive. Text in the questionnaire is repeated in the codebook; byte positions given in the codebook are repeated in the program set-up files for SAS or SPSS; variable and value labels in the set-up programs repeat text in the questionnaire and codebook. Moreover, if a change is made in a questionnaire, corresponding changes must be made in all the related research documents and files. In addition, considerable effort must be spent in formatting the questionnaires, codebooks, and set-up files.

Additionally, social science investigators who collect original data are increasingly being asked by their sponsors and other researchers to share their data sets with other researchers. Data sharing requires a standard of documentation that is not often adhered to by researchers who develop data sets for their own use. Yet, many investigators are either unable or unwilling to put

in the extra effort and resources required to bring a data set up to minimum standards, especially if they turn to this task only at the end of their research projects, when funds are exhausted.

The Computer Assisted Questionnaire and Dataset Developer is a software program that promotes such high standards by easing the entire process of questionnaire creation, documentation, data entry, error checking, and creation of analysis program set-up files. It does so by permitting the researcher to enter common information only once, and by automating such tasks such as formatting and generating all research materials (including the questionnaires, codebooks, and set-up files), checking data files for errors, and preparing files for analysis. This would free up the researchers time for scientific rather than clerical tasks.

FUNCTIONS OF THE SOFTWARE

The CAQDD performs the following functions:

- Survey Generation—generates fully formatted survey questionnaires or instruments in a printed format.
- Codebook Generation—generates the data set documentation in a printed codebook, flow chart, and data file map.
- Data Entry—provides for data entry from completed questionnaires, with simultaneous error checking.
- Program File Generation—produces a raw data file in ASCII format, and builds the program set-up files needed to transform the raw data file into SPSS and/or SAS system files.

The software stores information entered by the researcher in a database. The researcher enters data into this database just once. The program then draws on this database to generate a questionnaire; create the corresponding variable and value labels, codebooks, as well as SPSS and SAS set-up files to read in the data once it is entered. In addition, the program creates a file map and checks data files for errors during data entry. The database is fully editable, permitting the user to modify, update, correct, or restructure the questionnaire (including reordering questions or sections) at any time.

Complex elements such as question headers and footers, question-level skip patterns, value-level skip patterns, the assignment of global missing value codes, and notes and comments are also allowed. The program has the ability to store information about and format a variety of types of questions: fixed response, fill-in, open end, multiple response, event histories, and multi-item questions. The codebook automatically assigns data record positions and formats and generates a data file map based on information about variable widths and formats supplied by the user. Automatic checking of response data upon entry prevents the user from entering invalid codes. The user can generate any number of studies with the software, and may draw on information in one study to create items for another study, without having to reenter the data.

The user enters info into this database through a series of windows that appear on screen. In the following sections, we will describe some of the key windows, as well as the information that can be entered into them for the simultaneous creation of a questionnaire and a corresponding data set.

A) Creating a Questionnaire with the CAQDD Software

The very first task in primary data collection is the creation of a questionnaire or survey. To create a questionnaire, the user enters information about each question into the CAQDD database through what is known as the **Question Window** (see Figure 1). This information is used by the software for both formatting the question in the questionnaire, documenting the question in the codebook, as well as creating a corresponding variable in the eventual data set.

Figure 1. Question Window

The screenshot shows the 'New Question' dialog box. At the top, the 'Question Type' is set to 'Fixed Response'. Below this are four tabs: 'General', 'Header/Footer', 'Codebook Text', and 'Skips'. The 'General' tab is selected. In the 'General' tab, the 'Position Label' is '4', and 'Use As Identifier Question' is unchecked. The 'Question Text' field is a large cyan text area. Below it is the 'Fill In Text' field. The 'Data Type' is a dropdown menu, 'Input Format' is a dropdown menu, 'Field Width' is a cyan text field, 'Min Value' is a text field, 'Max Value' is a text field, 'Total Items' is a text field, and 'Field Decimal Places' is a text field. At the bottom of the 'General' tab, there are three checkboxes: 'Limit to Value Set' (unchecked), 'Suppress Question in Instrument' (unchecked), and 'Display Other/Specified' (unchecked). Below these is the 'Class Topic' dropdown menu. On the right side of the dialog, there are 'OK' and 'Cancel' buttons.

To create a question, the user needs to fill in information on each question regarding *Question Type* (such as whether the question is a fixed response question), its position or order in the questionnaire (*Position Label*), the actual text (*Question Text*), as well as any header and footer information that goes with the question. The software allows for a several different types of questions that can be entered into the questionnaire:

- Fixed-response questions: These are questions in which the respondent is presented with a set of pre-determined responses that are categorical in nature.
- Fill-in questions: These are questions in which the respondent fills in a short answer (often a number or a word) without being presented with choices.
- Open-ended questions: *Open-ended* questions are similar to fill-in questions except that they allow for longer verbatim responses.

- Multi-item questions: These are sets of questions that have a common response set and may have a common introduction, and are linked together by their content. A typical example is a set of attitude items that, together, make a scale measuring an underlying construct.
- Multi-response questions: These are questions to which the respondent may give more than one response.
- Rosters or event histories: These are sets of questions (of any mix of the above types) that are repeated for each of several persons, events, or situations.

The following is a printout of a section of a questionnaire as produced by the software that has both fixed response and fill-in questions.

Figure 2: Sample Print out

D. BOYFRIENDS, GIRLFRIENDS AND SEXUALITY

1. Have you ever had sexual intercourse?

Yes	<input type="text" value="1"/>	➔ Skip to Question 3.
No	<input type="text" value="2"/>	

2. How old were you when you had sexual intercourse for the first time?

Years

3. Will you choose to have sexual intercourse in the next year?

Yes	<input type="text" value="1"/>
No	<input type="text" value="2"/>
Don't Know	<input type="text" value="3"/>

As noted above, the software also allows for more complex question types such as Multi-item questions and Rosters. Multi-item questions incorporate a cluster of sub-items that have the same set of responses, while a *Roster* occurs when the same set of sub-questions is repeated several times, each time pertaining to a different event or case. Examples of rosters are birth histories, employment histories etc. where information is asked about each case before moving on to the next one.

Complex elements such as question **headers and footers**, and **skip patterns** are also available with the CAQDD software. Skip patterns occur when a question is asked of just a subset of the respondents. In such instances, all other respondents are required to skip the question (see Question 1 in Figure 2 above). Skip patterns can be of two types: (a) *question* skip, and (b) *value* skip. Question 1 (Fig.2) is an example of a *value* skip as the skip pattern is based on the particular response to a question (all respondents who have answered “No” are required to skip to Question 3). A *question* skip occurs when the skip pattern is associated with a question. In the example

below (Figure 3), respondents not having siblings are asked to skip Question 3 entirely. Both types of skip patterns are possible with the CAQDD software.

Figure 3

If you do not have any siblings

➔ Skip to C. THINKING ABOUT THE FUTURE

3. How many siblings i.e. brothers and sisters (including step brothers and sisters) do you have?

Siblings

Question or Item Bank

While researchers can create their own questions for the questionnaire, they also have the option of using questions from an already existing **question (or item) bank** created by Sociometrics. The item bank consists of several thousand commonly used questionnaire items, scales, and other interviewing tools drawn from a variety of premier data sets in the field of family research¹. The software allows a researcher to search the items in the bank, select those of use for the research purpose at hand, and drop them directly into the questionnaire being developed. Although primarily beneficial to family researchers, the item bank consists of a broad range of general questions, and can be used by social scientists focusing on other areas besides family research. Questions in the item bank cover a diverse range of topics such as *crime and delinquency, sexuality, substance abuse, mental and physical health, education, employment history, fertility, marital relationships, aging, etc.*

Questions in the item bank are classified by the study that they are drawn from as well as by the topic they relate to (such as demographics, health, employment etc.). Such a classification allows users to streamline their search by study and topic. Users can also search the item bank for questions that they wish to include in their questionnaire by using appropriate search keywords. At present there are about 8000 questions in the item bank including multi-item and roster questions.

Users can also create their **own question bank** if they wish to reuse questions that they have created for their questionnaire. The questions in the user created item bank can then be copied into a different questionnaire.

Form Integrity

Form integrity is an error verification process at the questionnaire building stage. It ensures that the user does not enter inconsistent field specifications for each question in the questionnaire

¹ The items have been drawn from premier data sets such as the National Survey of Families and Households, 1992, 1988; The National Longitudinal Study of Adolescent Health, Waves 1 and 2 (1994-1996); National Survey of Children, 1976-1987; National Family Violence Survey, 1985; The High School and Beyond Longitudinal Survey, 1980-1986, among others.

(such as specifying a range of values that is outside the maximum or minimum value indicated for a particular question). The user should check form integrity prior to generating reports or beginning the data entry process.

B) Creating Variables and Values with the CAQDD Software

Defining the Variables

For each question in the questionnaire, there is a corresponding variable (or a set of variables, if it is a multi-item or multi-response question). The responses to questions define the variable values or value codes². For every question that is entered into the **Question Window** (see Figure 1), the user also fills out information that allows the software to define the type of variable/s produced by the questionnaire—i.e. whether the variable is a string or numeric variable (*Data Type*), the width of the variable (*Field Width*); the number of decimal places that need to be assigned to the values (*Field Decimal Places*), as well as the maximum and minimum values the variable can take (*Max Value* and *Min Value*). In addition, the software has special formats for inputting variables that are dates and time. Dates can be formatted to be a 10-character field (mm/dd/yyyy), while Time can be entered as an 8-character field separated by colons (hour: minute: second).

The CAQDD automatically assigns unique default names for variables during question entry process. Users have the option of defining their own variable names and are not restricted to the ones automatically assigned by the software.

Value Sets

An important feature of the software is the ability to create and use **Value Sets**. A set of values that tend to be repeated can be stored as a value set. For example, a lot of fixed response questions typically have yes/no responses. Rather than having to repeatedly retype these values, users can create a value set with values of “1=yes” and “2=no”. This value set can then be appended repeatedly to any number of questions with such yes/no responses, with a mere click of a button. Users can create any number of value sets with different values that tend to be repeated. The software allows the user to search for value sets using value labels and value set titles as keywords in the search.

C) Output produced by the Software

Once information is entered into the CAQDD database about the questions and the corresponding variables and values, the software, at this stage, is ready to automatically produce 3 types of output for the user: (1) the actual **Questionnaire** (or Survey Report), (2) **Codebook Report** (for a complete description of the variables), **Data File Map, Flow Charts**, (3) **SAS and SPSS** program statements for reading in the raw data when it is eventually collected.

The *Questionnaire or Survey Report* is the actual questionnaire produced by the software. The report is generated and sent to the screen from which it can be viewed for completeness and

² For example, the question “Are you male or female” can generate a variable *Gender* with 2 value codes: 1 and 2. The value labels for these codes are: 1 “female” and 2 “male” respectively.

accuracy. From the screen report it can be printed by a click of a button to produce the paper questionnaire. The survey report cannot be directly edited; changes to the questionnaire text (such as questionnaire title, headers and footers etc.) must be done in the appropriate windows. This ensures that changes will be reflected in all program outputs, such as the codebook and set-up files as well as the questionnaire.

The Codebook Report and the *Data File Map* both provide summary statements of what will be the eventual dataset. The codebook report, for example, shows each question in the questionnaire, the associated variable name and label, variable type (whether string or numeric), values and value labels, as well as the column placement of the variable in the data set. The following example shows a section from a sample codebook report.

Example 3: Sample Codebook Report

A. ID SECTION					
This information is confidential and will not be revealed to anyone.					
<u>Identifiers</u>					
1. Respondent's ID number					
<u>Name</u>	<u>Variable Label</u>			<u>Record</u>	<u>Columns</u>
<u>Format</u>					
Aa001__	Respondent ID Number			1	1 - 3
		A3			
CODE					
2. Time of interview					
<u>Name</u>	<u>Variable Label</u>			<u>Record</u>	<u>Columns</u>
<u>Format</u>					
Aa002__	Interview Time			1	4 - 11
F8					
CODE					
3. Date of Interview					
<u>Name</u>	<u>Variable Label</u>			<u>Record</u>	<u>Columns</u>
<u>Format</u>					
Aa003__	Interview Date			1	12 - 21
F10					
CODE					

The data file map is another report that summarizes the data set created by the user (Example 4).

Example 4: Sample Data File Map

<u>Values and Attitudes Questionnaire</u>					
Variable	Record	Start	End	Format	Variable Label
Aa001__	1	1	3	A3	Respondent ID Number
Aa002__	1	4	11	F8	Interview Time
Aa003__	1	12	21	F10	Interview Date
Ab001__	1	22	22	F1	Gender

Ab002__	1	23	23	F1	Ethnicity
Ab003__	1	24	25	F2	Number of Siblings
Ab004A01	1	26	27	F2	Sibling Age

In addition, the CAQDD software also automatically generates, in machine-readable format, the program statements necessary for reading in the data by two standard statistical software packages: SAS and SPSS. Examples 5 and 6, below, show sample sections of an SPSS and an SAS program set-up file, respectively, as generated by the software.

Example 5: A Section of the Sample SPSS Program File

```
* Sociometrics Corporation (Los Altos, CA)

SET PRINTBACK = NO/WIDTH = 80
FILE HANDLE INPUT/NAME='datafile.txt'
DATA LIST FILE=INPUT RECORDS =1
/1
Aa001__ 1-3 (A)
Aa002__ 4-11
Aa003__ 12-21
Ab001__ 22-22
Ab002__ 23-23
```

Example 6: A Section of the Sample SAS Program File

```
* Sociometrics Corporation (Los Altos, CA);

LIBNAME LIBRARY '/your path';
PROC FORMAT LIBRARY=LIBRARY;
VALUE Ab001__
  1="Female"
  2="Male";
VALUE Ab002__
  1="American Indian"
  2="Asian-American"
  3="Black"
  4="Hispanic"
  5="White"
  9="No Answer";
```

EXPORT OPTIONS

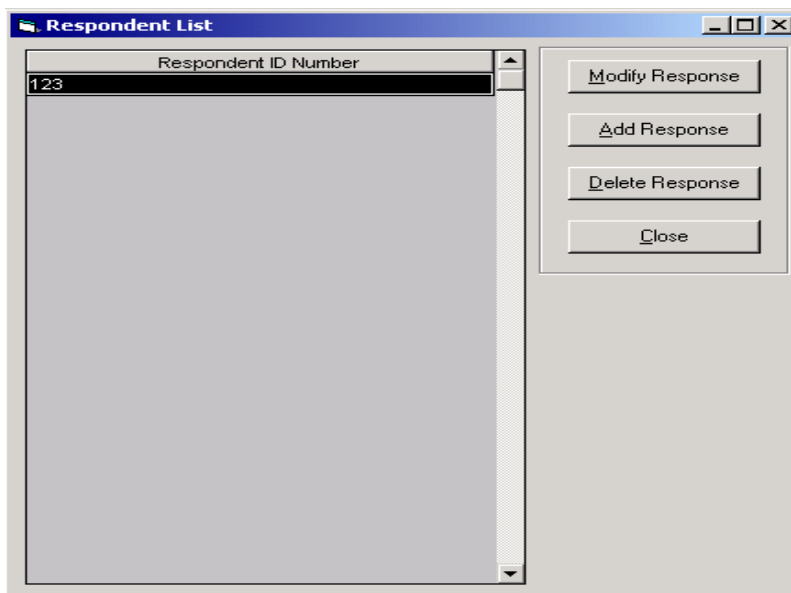
The software can export all the reports mentioned above in several different formats such as MSWord, HTML, Rich Text Format, Excel, ASCII, and MSWord. The user can export the survey he or she has created into a different software program for making more complex formatting changes in design and style (such as making font changes, inserting graphics and ClipArt's etc.). However, the user is cautioned not to make substantive changes to the content of the report in this way, as these changes would not be reflected in the underlying database and would not therefore be carried forward to the other associated reports.

D) Data Entry using the CAQDD Software

Having developed a questionnaire, the researcher then uses it to collect data. Data from the completed questionnaires will need to be converted into a machine readable data file for further analysis. This conversion can be done by entering the data into the CAQDD database through the **Response List Window** and the **Data Handler Windows** in order to create a raw data file.

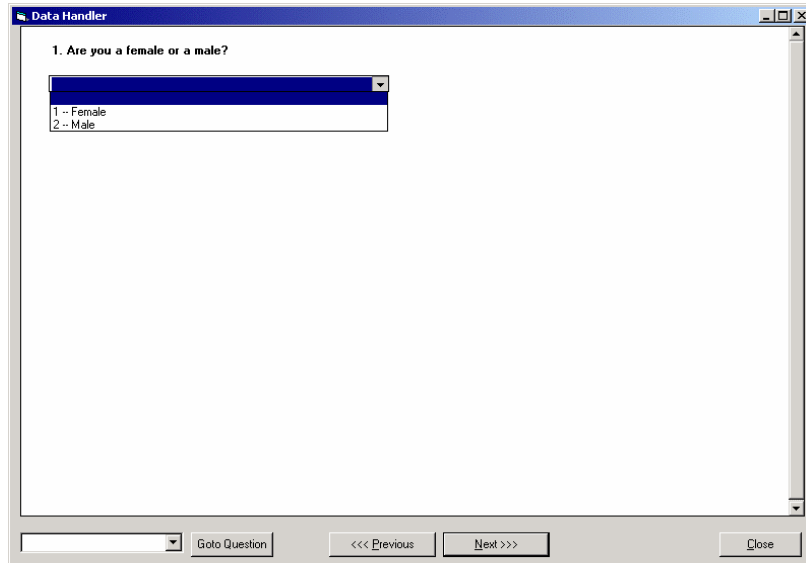
The **Respondent List Window** is the very first window that opens up within the Data Handler (*Fig. 4*). The window provides for entering respondent data. It lists all the respondents for whom data have been entered, and displays identifying information on each of these cases. This identifying information for each respondent is drawn from the ID Section in the questionnaire.

Figure 4 Respondent List Window



The data are then entered by opening the Data Handler Window (*Fig. 5*). The data handler screen displays each question, its response values (including those for missing values) exactly in the order that they appear in the questionnaire. The researcher can then enter data from the questionnaires directly into the software.

Figure 5. Data Handler Window



Minimizing Errors

Data entry through the data handler has several advantages over entering data through spreadsheets or text editing programs. First, because the questions appear exactly as they would in the questionnaire, the process of data entry is made faster and more accurate (*see Fig. 5*). Being able to view questions is particularly important when a large number of data points have to be entered. In the spreadsheet, users have to enter data under a variable name and not the question, which can cause confusion.

But the biggest advantage of the data handler is in minimizing errors during data entry. The data handler prevents the researcher from entering invalid values as it forces the user to choose only the values assigned to the question, including the values assigned to missing cases.

In *Figure 5*, the user will avoid typing in invalid values, such as “3” or “99” for the question on gender as the program will present only the “legitimate” values (i.e. “1” for female and “2” for male). For fixed as well as fill-in questions, the data handler will reject values that fall outside the minimum and maximum range specified in the variable window.

Once the user has entered data, the user can update information on any of the individual cases or respondents by going to the **Response List Window** (*Figure 2*). The user can select the case for which information needs to be modified or updated, and make the necessary changes.

The raw data file is the fourth and final output created by the software. Data can be stored as a machine-readable file in any of the three formats (standard, comma or tab delimited).

SUMMARY

The software is intended to greatly facilitate questionnaire writing, and data set development by automating tasks that can best be done by computer. Its biggest advantage is that it saves time and effort for the researcher by greatly simplifying data processing and questionnaire writing. By putting in information just once in the Question Window, the researcher is able to generate a fully usable questionnaire; has the variables and values defined in a codebook; and has the SAS and SPSS program statements ready for reading in the data once it has been collected. Equally important, the program has a number of error-checking routines built in to prevent common or critical errors both in the questionnaire creation (Form Integrity) as well as during the data entry process. Finally, the availability in the item bank of several thousand variables makes it possible for the user to quickly identify and include questions in his or her own questionnaire that have been used and tested in premier studies in family research. The simplicity of the program as well as its capacity to minimize errors makes the program attractive as a research tool to both *seasoned data researchers as well as novices*.