Generating MUD (Messy Ugly Data): A Program for Courses in Research Methods and Introductory Statistics

Proposed as Creating MUD (Messy Ugly Data) with QuesGen: A Program and Projects for Introductory Statistics

The *MUD: Messy Ugly Data Generator* computer program is available at:
http://oak.cats.ohiou.edu/~brooksg/mud.htm

Abstract

The purpose of this paper and presentation is to demonstrate a new computer program (*MUD: Messy Ugly Data Generator*) that instructors can use to help create realistic projects for students in their introductory statistics courses. In particular, the free and downloadable Windows-based computer program will be provided on CD to participants along with several projects based on the software. Because reasonable existing databases are not always available for students to use, the MUD program was developed to generate data, including either good clean data or messy ugly data, to simulate realistic research scenarios for real-world-like projects. Examples of program input and data screening, cleaning, and analysis projects will be shared with participants.

Real Data in Statistics Education

Applied statistics courses and introductory research methods often contain components that require students to collect data, to perform statistical analyses using a computer software package, and/or to interpret the results of these analyses. Additionally, students in these applied courses must often learn data management skills such as entering, screening, and transforming the data using a statistics program such as SPSS, Minitab, or SAS.

Many faculty try to have students use real data for such purposes, data which (a) may already exist locally or in a national database, (b) may be found in textbook examples or on the internet, or (c) may be collected during the course. Indeed, several scholars have recommended real data for statistics courses. For example, Cobb (1993) summarized National Science Foundation projects in statistics education, several of which included real data. Cobb noted that "no student who spends time with [real] data will come away thinking of statistics as just a collection of empty numerical rituals" (p. 8). Others, including Sullivan (1993), Crow (1996), and Mittag and Taylor (1996), have advocated that students analyze real data, and even data that they have collected themselves. Although most of these studies have focused on the K-12 curriculum, several scholars have made similar recommendations for undergraduate students in mathematical and psychological statistics courses (e.g., American Statistical Association, 2005; Beins, 1993; Conners, McCown, and Roskos-Ewoldsen, 1998; Thompson, 1994). The most important issues related to using real data seem to be (a) providing context for the data analysis and answering real research questions, (b) giving students an opportunity to participate in the data collection process, and (c) providing more interesting and relevant analyses.

Unfortunately, there are several potential problems with using data obtained in one of these ways. For example, data cannot always be collected during a course due to IRB, time, or cost constraints. Even if data can be collected, it is often necessarily limited to small samples and to problems thrown together quickly and not of real interest to the students. Unfortunately, several of the examples provided in the literature include data based on information collected from the class (e.g., height and weight) or activities based on things like M & M candy and temperature—real data, but not terribly relevant to most students' courses of study. Similarly, existing data (e.g., available on the Internet or previously collected by a faculty member) may not be sufficiently relevant or interesting to students. Further, such existing data are usually already in clean condition, thereby minimizing students' data management activities. Finally, reasonable data for a real problem—with appropriate variables, numbers of variables, and numbers of cases—may simply not exist.

Realistic Data in Statistics Education

The MUD program and its associated projects were inspired by learning activities such as that described by Holcomb and Spalsbury (2005). Holcomb and Spalsbury's project is innovative, using real data with real problems (i.e., erroneous data) in an easy-to-manage project for both instructors and students. The difficulty with such a project, however, is that the variables are given by the real data and not always appropriate for students in particular courses. For example, the data they provide on a web site is from a study about gender differences in the amount of calcium present in patients over 65 years old—data not entirely appropriate for social science statistics courses.

Alternatively, the MUD program allows instructors or students to generate data for projects with realistic data and variables more appropriately tailored for their particular needs. In order to individualize projects, instructors can easily create separate data for each student or group of students. Instructors can also create a large database from which students are assigned unique subsets of cases. Or, potentially, instructors can show students how to create data to fit their own unique research problems (e.g., data to fit a questionnaire created as part of a project in a research methods course).

Random data can be generated to fit the characteristics reported in a research article, thereby making an instructional pseudo-replication of interesting studies relatively easy. Indeed, scholars such as Barcikowski, Johanson, Rich, & Robey (1989), Derry, Levin, and Schauble (1995), and Singer and Willett (1990) have advocated using realistic scenarios, like those provided by research articles, for classroom presentations and projects. Data can also be generated to fit or violate a particular set of distributional characteristics defined by the teacher, giving the student an opportunity to "find" problems in the data. Likewise, data can be generated to model problems, such as violations of assumptions of statistical tests.

Students then analyze and interpret these data using a statistical computer package introduced in their course (e.g., SPSS or Minitab). The process enables the students to appreciate the complexity and appeal of practical, "real world" data analyses. Also, students who generate their own data think about the data differently than those who use existing secondary data sources provided to them by textbooks or instructors, or available to them on the Internet. For example, data generation reinforces the understanding of the research problem at hand. Working with data that fit a relevant and interesting research context helps them make better sense of their variables and the connections among them. It requires students to think sensibly about the characteristics of the data (e.g., means, variances, minimum and maximum reasonable values), consider issues like effect sizes, and account for correlations among variables.

Although most statistics programs and a host of standalone computer programs will create data for a variety of distributions, such generation methods are limited in a variety of ways. In particular, most ignore or do not allow correlations among variables and generate only "clean" univariate data (clean data has no inherent problems, like extreme, missing, or impossible values). Although there are algorithms available for non-normal distributions, even these methods create only clean non-normal data.

Program Description

The MUD Generator program provides many options, including specification of the correlations among variables and the ability to create items that can be combined into a summated scale (see Sample Data Generations Steps). Although MUD is designed primarily for instructors to use to create data for students, students may also use it with some guidance. MUD is designed primarily to create data that mimics survey data collected via questionnaires. Such an approach provides flexible data that can be used for a variety of purposes and analyses in an introductory statistics course. For example, dichotomous or categorical variables can be used as independent variables for mean comparisons; multiple variables can be used for correlation and regression analyses. The software and the associated materials will be made available for free download from the Internet so that faculty and students may use them in support of statistics instruction and learning.

The MUD program uses descriptive information such as that often reported in research articles as input parameters for the generation of data. The program will generate data for up to 25 variables, which may include a number of items for a summated scale. Users interactively enter the (a) name for the database, (b) the percentage of "messy data" to be generated, (c) variable names, means, standard deviations, minimums, maximums, and numbers of decimals for each variable, (d) item names, positive/negative direction, and item intercorrelations, (e) correlations among the variables, including the correlations between the variables and the total scale score, and (f) the number of cases to generate. In working with students, perhaps to generate data for a survey created as part of a classroom project, the instructor can discuss the data generation process with the students. For example, talking about reasonable standard deviation, correlations and effect sizes, sample size, and summated scales. Several data files are created by the program, most notably a comma-delimited text file containing the data, an information file about the data that were generated, a text file that contains generated surveys/questionnaires, and a specially formatted file with input information for potential use again later.

Perhaps the most important aspect to MUD is the ability (but not requirement) to generate messy ugly data easily. That is, by default, 3% of the data values are automatically created by the program that are purposefully extreme, missing, or bad (a clean file with 0% messy data can also be specified by the user). For example, Figure 4 shows a Q01 value of -20 for case 10, a Q04 value of 33.58 for case 6, and a Q05 value of 777.0 for case 2 – all impossible values given the data generation parameters (see Figure 1). Such ugly data values are created thoughtfully to mimic realistic data and data entry problems, including transposition of numeric values (e.g., 19 instead of 91), double entry of values (e.g., 33 instead of 3), and extreme numeric values, both inside and outside of acceptable values (e.g., 4 standard deviations above the mean). The projects require students to screen and examine data, looking for problem values using descriptive statistics. Sometimes the values in the data set are reasonable, but wrong. For example, case 3 in Figure 4 shows a Q10 value of 4, a completely reasonable value on a scale of 1 to 5; however, Figure 5 shows that the Q10 value for case 3 was actually 3. Problems such as this cannot be identified through descriptive statistics, helping emphasize to students that there are no shortcuts to research.

Several projects have been created for use with MUD, including (a) data entry projects, (b) data screening projects, and (c) data analysis projects (see Sample Data Analysis Project). The data entry projects require students to enter data they are given in the survey/questionnaire into a statistics program (e.g., SPSS). Specifically, the project requires students to (a) enter data, (b) create useful variable names, variable labels, (c) set up missing values, (d) determine number of decimals to display and types of variables, and (e) transform existing data (e.g., recode variables, compute total scores). The data screening projects require students to screen a data file for extreme, missing, and bad values—and then clean or fix the data based on original questionnaire values. The data analysis projects typically require students to analyze a given data set using either or both descriptive and inferential statistics. Other projects can be created as well. For example, in a research methods or survey design course, students can create a survey and then generate data for the survey (similar to the project described by Thompson, 1994); in such cases, messy ugly data may not be necessary and can be set to 0%.

Program Development

The MUD program was developed in Borland Delphi 8 Professional, a Pascal compiler for the Microsoft Windows XP operating systems. The L'Ecuyer (1988) uniform pseudorandom number generator was used primarily because such combined generators have been recommended for use with the Box-Muller method for generating random normal deviates, as is the case for the MUD program (Golder & Settle, 1976). Specifically, the FORTRAN code for the L'Ecuyer generator of Press, Teukolsky, Vetterling, and Flannery (1992) was translated into Pascal by the author for the uniform deviate; the Box-Muller algorithm used in the MUD program was adapted by the author from the standard Pascal code provided by Press, Flannery, Teukolsky, and Vetterling (1989). Jointly distributed multivariate normal data are generated following a Cholesky decomposition procedure adapted from Nash (1990) and Bratley, Fox, and Schrage (1987).

Educational or Scientific Importance of the Paper

It is very important for students in applied statistics courses to learn to manage data, including messy ugly data. Unfortunately, finding such data is not always easy. The MUD Generator program will allow educational statistics and research methods instructors to create projects for real-world-like research scenarios with data that reasonably simulate real-world-like data issues.

References

American Statistical Association (2005, February). *Guidelines for assessment and instruction in statistics education (GAISE) college report*. Retrieved March 15, 2006, from http://www.amstat.org/education/gaise/

Barcikowski, R. S., Johanson, G., Rich, C. E., & Robey, R. R. (1989, March). *Creating "real world" multivariate data: A means of providing students with practical data analysis experience*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Beins, B. C. (1993). Writing assignments in statistics classes encourage students to learn interpretation. *Teaching of Psychology, 20*, 161-164.

Bratley, P., Fox, B. L., & Schrage, L. E. (1987). *A guide to simulation* (2nd ed.). NewYork: Springer-Verlag.

Conners, F. A., Mccown, S. M., & Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teaching of Psychology, 25*, 40-42.

Crow, T. (Ed). (1996). *Real data resources for teachers*. Columbus, OH: Eisenhower National Clearinghouse for Mathematics and Science Education.

Derry, S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking trhough situated simulations. *Teaching of Psychology, 22*, 51-57.

Golder, E. R., & Settle, J. G. (1976). The Box-Muller method for generating pseudo-random normal deviates. *Applied Statistics, 1*, 12-20.

Holcomb, J., & Spalsbury, A. (2005). Teaching students to use summary statistics and graphics to clean and analyze data. *Journal of Statistics Education, 13*(3). Retrieved March 15, 2006, from http://www.amstat.org/publications/jse/v13n3/datasets.holcomb.html

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM, 31*, 742-749, 774.

Mittag, K. C., & Taylor, S. E. (1996). *Using graphing calculator technology in educational statistics courses*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Nash, J. C. (1990). *Compact numerical methods for computers: Linear algebra and function minimisation* (2nd ed.). New York: Adam Hilger.

Press,W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal: The art of scientific computing*. New York: Cambridge University Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

Singer, J. D., & Willett, J. B. (1990). Improving the teaching of applied statistics: Putting the data back into data analysis. *American Statistician, 44*, 223-230.

Sullivan, M. M. (1993). *Students learn statistics when they assume a statistician's role*. Paper presented at the annual conference of the American Mathematical Association of Two-Year Colleges, Boston, MA.

Thompson, W. B. (1994). Making data analysis realistic: Incorporating research into statistics courses. *Teaching of Psychology, 21*, 41-43.

A short survey was given to graduate students in a College of Education course following a library training workshop to assess their attitude toward the usefulness of the training. Here's the survey:

Please answer the following demographic questions.

1. Age. . . . . _____

2. Gender. . . . . Male (0) Female (1)

3. Department. . . . . Teacher Education (1) Educational Studies (2) Higher Education (3) Other (4)

4. Current GPA. . . . . _____

5. What was your Library Knowledge (exam) score? _____

6. What was your Library Skills (project) score? _____

Circle the rating that most closely describes your feelings about these statements concerning our two class meetings in the library on the scale from Strongly Disagree (SD) to Strongly Agree (SA).

	points:	1	2	3	4	5
7.	I learned several things I did not know.	SD	D	N	A	SA
8.	All students should be encouraged to participate in this training.	SD	D	N	A	SA
9.	The library training was <u>not</u> useful for me.	SD	D	N	A	SA
10.	The library training covered content that was important for me to know. .	SD	D	N	A	SA

Variable Information

- For both *LIBRARY SKILLS* and *LIBRARY KNOWLEDGE*, as measured, the minimum possible score was 0 and the maximum score was 100. *CURRENT GPA* ranged from a minimum of 0.00 to a maximum of 4.00. *AGE* was for graduate students (a minimum of 20 and no maximum).
- For *GENDER* and *DEPARTMENT*, the value given to each response category is in parentheses (e.g., Teacher Education students were coded as 1)
- Items 7 through 10 constitute an *LIBRARY TRAINING EVALUATION* scale; responses that reflect a more positive attitude will be scored higher, such that for items 7, 8, and 10 SD is given 1 point, D=2, N=3, A=4, and SA=5 (you will not analyze Q7, Q8, Q9, or Q10 individually – i.e., they are items, not variables)
- Item 9 is negatively worded; that is, <u>agreement</u> with that item reflects a more **<u>negative</u>** attitude toward the library training (therefore, it will need to be recoded during the project)
- You will be assigned data based on this evaluation form to use for the project.

Note: The instructor will provide students with QUESTIONNAIRES generated by MUD prior to this part of the project. The author typically prints the questionnaires out for the students (i.e., usually multiple questionnaires per page to save paper) and/or provides them as unique files on a course web site (i.e., many questionnaire files are uploaded to the web page and students are given a particular numbered file to download for their project). Such Questionnaires look like this:

```
            Survey ID Number:       8
                         Age:     27
                      Gender:      1
                  Department:     -3
                         GPA:     2.71
       Library Knowledge Exam:    67.9
       Library Skills Project:     74
I learned things I did not kno:      5
 Students should be encouraged:      4
Training was not useful for me:      1
 Content covered was important:      4
```

Part I: Data Entry and Database Setup

Setting up your database (often called dataset) and entering the data

1. For every variable in your database, create a *nice* (i.e., brief but sufficiently descriptive) variable **NAME** and variable **LABEL**. Note that later you will **COMPUTE** new variables for which you will also be expected to create *nice* names and labels.

2. For all categorical variables, create *nice* **VALUES** labels.

3. For all variables, set up appropriate **MISSING** value codes (note that it may also be useful to set a **VALUES** label for your **MISSING** value codes). It is sometimes useful to set up 3 discrete missing value codes: one **MISSING** value for data that is missing completely, a different **MISSING** value for data that is erroneous that you cannot fix, and yet a different **MISSING** value for other reasons (e.g., when someone has chosen both options). Using **MISSING** value codes is usually better than just using the SPSS *system missing* value (a period or dot) because you know that you didn't just forget to enter a value.

4. For all variables, set up appropriate **TYPE**, number of **DECIMALS**, and level of **MEASURE**. At your discretion, as needed, change the variable **WIDTH**, **COLUMNS**, and **ALIGN**.

5. Enter the questionnaire/evaluation data assigned to you into SPSS. Enter it exactly as you receive it, even if some values seem strange for the given variables. If someone is missing a particular value, enter an appropriate **MISSING** value code (i.e., the code value for data that is missing completely).

6. Using SPSS procedure(s), **DISPLAY DATA FILE INFORMATION** for your working database. Print this report (do your best to fit this report on a single page of printed output).

7. Using SPSS procedure(s), **REPORT CASE SUMMARIES** for your working database. Print this report (do your best to fit this report on a single page of printed output).

8. Turn in the SPSS output printed at #6 and #7.

Note: The instructor will provide students with text DATA FILES generated by MUD prior to this part of the project. The author typically the text databases as unique files on a course web site (i.e., many database files are uploaded to the web page and students are given a particular numbered file to download for their project). Such Database Files look like this:

```
id,q01,q02,q03,q04,q05,q06,q07,q08,q09,q10
1,26,0,2,3.90,61.0,67,4,3,3,3
2,25,0,4,3.14,85.7,90,4,4,2,5
3,24,1,4,3.66,100.0,100,5,5,2,54
.
.
.
```

Part II: Data Retrieval

Opening a text file with the same data as Part I. Instead of entering data yourself, you will receive a text file that you will pretend someone else has created for you.

1. Use the SPSS **READ TEXT DATA** wizard (under the **FILE MENU**) to open the file.
   - At Step 1 of 6, click ***Next >***.
   - At Step 2 of 6, verify that **DELIMITED** is marked for the VARIABLE ARRANGEMENT and CHANGE the option **ARE VARIABLE NAMES INCLUDED AT THE TOP OF YOUR FILE?** to **YES**. Click ***Next >***.
   - At Step 3 of 6, verify that the FIRST CASE OF DATA BEGINS ON **LINE NUMBER 2**, that **EACH LINE REPRESENTS A CASE**, and that you will import **ALL OF THE CASES**. Click ***Next >***.
   - At Step 4 of 6, verify that the DELIMITER THAT APPEARS BETWEEN VARIABLES is **COMMA** and that THERE IS **NO** TEXT QUALIFIER. Click ***Next >***.
   - At Step 5 of 6, click ***Next >***.
   - At Step 6 of 6, click ***Finish***.
2. Change every variable **NAME** to the same **NAME** used for the variable in Part I.
3. Use the SPSS procedure **COPY DATA PROPERTIES** wizard (under the **DATA MENU**) to copy your setup information from the file you created in Part I.
   - At Step 1 of 5, enter or BROWSE for your Part I file so that it is listed in the box.
   - At Step 2 of 5, verify that you will APPLY PROPERTIES FROM SELECTED COURSE FILE VARIABLES TO MATCHING WORKING FILE VARIABLES. Click ***Next >***.
   - At Step 3 of 5, verify that you will **REPLACE** all VARIABLE PROPERTIES. Click ***Next >***.
   - At Step 4 of 5, click ***Next >***.
   - At Step 5 of 5, verify that you will EXECUTE THE COMMAND and click ***Finish***.
4. Save the file with a different name than the file you saved in Part I.
5. Using SPSS procedure(s), **REPORT CASE SUMMARIES** for your working database. Print this report (do your best to fit this report on a single page of printed output).
6. Using SPSS procedure(s), report appropriate **descriptive statistics** for each variable (i.e., note that different descriptive statistics are appropriate for categorical variables than for scale variables). Print this report (do your best to fit this report on a single page of printed output).
7. Turn in the SPSS output printed at #5 and #6.

Part III: Data Screening

Examining your dataset for data entry errors and otherwise bad data. Find and use variables with interesting data wherever possible for each item below (where no such interesting data exists, explain what it would look like in the output).

1. Using SPSS **DESCRIPTIVES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **minimum** and **maximum values** to screen data for a scale variable (e.g., *AGE*, *CURRENT GPA*, *LIBRARY KNOWLEDGE*, *LIBRARY SKILLS PROJECT*).

2. Using SPSS **DESCRIPTIVES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **standardized (*z*) scores** to screen data for a scale variable.

3. Using SPSS **EXPLORE**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use a **box plot** to screen data for a scale variable.

4. Using SPSS **EXPLORE**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use an **extreme values** report to screen data for a scale variable.

5. Using SPSS **EXPLORE**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use a **stem-and-leaf display** to screen data for a scale variable.

6. Using SPSS **FREQUENCIES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **a histogram** to screen data for a scale variable (suppress tables).

7. Using SPSS **FREQUENCIES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **a frequency table** to screen data for a categorical or ordinal variable (e.g., *GENDER*, *DEPARTMENT*, Items 7-10).

8. Using SPSS **FREQUENCIES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **a bar chart** to screen data for a categorical or ordinal variable.

9. Using SPSS **FREQUENCIES**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **a pie chart** to screen data for a categorical or ordinal variable.

10. Using SPSS **SCATTER/DOT**, along with any necessary additional analysis, describe and demonstrate (with SPSS output) how you can use **a scatter plot** to screen data for a combination of scale variables (e.g., *CURRENT GPA* and *LIBRARY KNOWLEDGE*).

11. Turn in the output printed and annotated (i.e., with your demonstrations) for #1 through #10.

Note: The instructor will provide students with COMPARISON files generated by MUD prior to #8 of this part of the project. The author typically prints the Comparison files out for the students and/or provides them as unique files on a course web site (i.e., many comparison files are uploaded to the web page and students are given a particular numbered file to download for their project). Such Comparison Files look like this:

Survey	Variable Name	Original Value	Survey Value	Dataset Value	S<>D
3	Content covered was important	5	5	54	**
8	Department	1	-3	-3	
11	Content covered was important	5	5	55	**
13	GPA	2.96	2.96	0.30	**

.
.
.

Part IV: Data Cleaning

Fixing your dataset so that data is as accurate as possible.

1. Based on your data screening done in Part III, you will fix the errors in your Part II data where possible. That is, you will find the "correct" data for problem cases from the questionnaires given to you in Part I. Note that sometimes even your supposedly-correct questionnaire data has missing or strange values; in such cases, change the strange data to appropriate missing value codes.

2. Explain how you identified every problem value that you fixed or replaced (e.g., which data screening method helped you identify the problem, or did you only find the problem after looking at the original questionnaires.

3. Save the fixed file with a different name than the file you saved in Part II.

4. Using SPSS procedure(s), **REPORT CASE SUMMARIES** for your working database. Print this report (do your best to fit this report on a single page of printed output).

5. Using SPSS procedure(s), report appropriate **descriptive statistics** for each variable (i.e., note that different descriptive statistics are appropriate for categorical variables than for scale variables). Print this report (do your best to fit this report on a single page of printed output).

6. Compare all variables before and after fixing data errors. That is, compare the descriptive statistics reported in Part II with those reported here in Part IV. Describe the important differences.

7. Turn in the SPSS output printed at #4 and #5 and your responses to #6.

---

8. You will receive a report of all problems in your data. Make ALL changes so that your database has all the values listed in the ***Original Value*** column of the report instead of the problem data.

9. Save this final fixed file with a different name than the file you saved at #3.

10. Using SPSS procedure(s), **REPORT CASE SUMMARIES** for your working database. Print this report (do your best to fit this report on a single page of printed output).

11. Turn in the SPSS output printed at #10. You will use this database to use for the remaining parts of the project.

Part V: Data Manipulation

Transforming and computing variables so that your data will answer your research questions.

1. Using SPSS procedure(s), **RECODE** the *LIBRARY TRAINING EVALUATION* scale Item 9 into the same variable so that it has the same scale direction as the other 3 scale items. That is, Item 9 needs to be recoded so that 5 represents the most positive attitude/evaluation.

2. Using SPSS procedure(s), **COMPUTE** a single variable that represents the **total** of Items 7-10 (a new single variable for the *LIBRARY TRAINING EVALUATION* scale). Remember that because of the way individual items were scored, higher scores on the *LIBRARY TRAINING EVALUATION* scale reflect a more positive evaluation of (or attitude toward) the library training.

3. Using SPSS procedure(s), create a new variable (perhaps called *LTE_RANK*) for the **RANK** of the students' scores on the *LIBRARY TRAINING EVALUATION*.

4. Using SPSS procedure(s), **COMPUTE** a new variable that represents the **average** of *LIBRARY KNOWLEDGE* and *LIBRARY SKILLS*, which will be considered the *LIBRARY UNIT GRADE*.

5. Using SPSS procedure(s), **RECODE** *LIBRARY UNIT GRADE* into a different variable (perhaps called *LETTER GRADE*) to represent categorical letter grades, using the following grading scale: A≥90, B≥80, C≥70, D≥60, and F<60.

Part VI: Data Screening (again)

Examining your dataset for additional data problems. Here you will examine variables for some additional specific problems.

1. Using appropriate SPSS procedure(s) from Part III to examine the new variables you created in Part V (e.g., *LIBRARY TRAINING EVALUATION*, *LTE_RANK*, *LIBRARY UNIT GRADE*, *LETTER GRADE*). Report how you investigated (screened) the data, any additional problems you identified (if any), and which techniques you used to find them.

2. Using SPSS procedure(s), examine the **normality** (both graphically and statistically) of the scale variables in your database (e.g., *AGE*, *CURRENT GPA*, *LIBRARY KNOWLEDGE*, *LIBRARY SKILLS PROJECT*, *LIBRARY TRAINING EVALUATION*, and *LIBRARY UNIT GRADE*). Hint, you can get such information in several ways using **EXPLORE**.

3. Using SPSS procedure(s), examine the **normality** (both graphically and statistically) of one of the scale variables in your database – separately for males and females.

Part VII: Describing Individual Variables

1. Using SPSS **FREQUENCIES**, report descriptive statistics (including **mode**) for *LIBRARY KNOWLEDGE* — be sure to **suppress** tables since this is a scale variable
2. Using SPSS **DESCRIPTIVES**, report **descriptive statistics** for *LIBRARY SKILLS*
3. Using SPSS **EXPLORE**, report **descriptive statistics** for *CURRENT GPA*
4. Using SPSS **CASE SUMMARIES**, report **descriptive statistics** for *LIBRARY TRAINING EVALUATION*
5. Using SPSS **FREQUENCIES**, report appropriate **descriptive statistics** for *GENDER*
6. Using SPSS **FREQUENCIES**, report a **frequency table** for the *letter grades* of *LIBRARY KNOWLEDGE*
7. Using SPSS **FREQUENCIES**, create a **histogram** for *LIBRARY TRAINING EVALUATION* with the **normal curve** displayed
8. Using SPSS **FREQUENCIES**, create a **bar chart** for *DEPARTMENT*
9. Using SPSS **FREQUENCIES**, create a **pie chart** for *GENDER*
10. Using SPSS **EXPLORE**, create a **histogram** for *CURRENT GPA*
11. Using SPSS **DESCRIPTIVES**, calculate **standardized scores** for *LIBRARY TRAINING EVALUATION*
12. Using SPSS **EXPLORE**, report which cases have the 5 highest **standardized scores** for *LIBRARY TRAINING EVALUATION*
13. Using SPSS **FREQUENCIES**, report **how many** respondents were 22 years old and how many were over 25
14. Using SPSS **FREQUENCIES**, report which value for *LIBRARY SKILLS* is at the **75th percentile**
15. Using SPSS **EXPLORE**, report which value for *LIBRARY TRAINING EVALUATION* is at the **75th percentile**
16. Using SPSS **GRAPH ERROR BAR**, create an **error bar plot** for the variable *LIBRARY TRAINING EVALUATION*

For the remaining items in this Part, determine for yourself which SPSS procedure to use (or work it out by hand if SPSS will not easily give you the answer) — where you use SPSS, provide outputs

17. Report **range** for *LIBRARY KNOWLEDGE*
18. Report the **percentage** of respondents who were in the Department of Educational Studies
19. Report the **standard error of the mean** for *LIBRARY SKILLS*
20. Report a **95% confidence interval** for the mean of *LIBRARY KNOWLEDGE*

Part VIII: Describing Combinations of Variables

1. Using SPSS **EXPLORE**, report the **mean** *AGE* for respondents grouped by *DEPARTMENT*
2. Using SPSS **SPLIT FILE** and **FREQUENCIES**, report the **mean** *AGE* for respondents grouped by *DEPARTMENT* (be sure to UN-split the file before running any other analyses)
3. Using SPSS **FREQUENCIES**, create **bar chart(s)** for the variable *GENDER* for respondents grouped by *DEPARTMENT* (hint, split the file)
4. Using SPSS **EXPLORE**, create **histograms** for *AGE* for respondents grouped by *DEPARTMENT*
5. Using SPSS **CROSSTABS**, create a table of *GENDER* by *DEPARTMENT*
6. Using SPSS **SELECT CASES** and **FREQUENCIES**, create a **pie chart** for *GENDER* for only those respondents in the Department of Teacher Education (be sure to UN-select the data before running any other analyses)
7. Using SPSS **EXPLORE**, create one box plot for the variable *LIBRARY TRAINING EVALUATION* with respondents grouped by *GENDER*
8. Using SPSS **EXPLORE**, create one box plot for the variables *LIBRARY SKILLS* and *LIBRARY KNOWLEDGE* displayed together
9. Using SPSS **CORRELATE BIVARIATE**, calculate both the **Pearson correlation** and the **Spearman Correlation** between *AGE* and *LIBRARY TRAINING EVALUATION*
10. Using results from SPSS **CORRELATE BIVARIATE**, report the magnitude of **shared variance** between *LIBRARY SKILLS* and *AGE*
11. Using results from SPSS **CORRELATE BIVARIATE**, report which variable has the highest Pearson correlation with *LIBRARY TRAINING EVALUATION*: *LIBRARY KNOWLEDGE*, *LIBRARY SKILLS*, or *CURRENT GPA*.
12. Using SPSS **REGRESSION LINEAR**, report the **mean** for *LIBRARY TRAINING EVALUATION* and report the **bivariate correlation** between *CURRENT GPA* and *LIBRARY TRAINING EVALUATION*
13. Using results from SPSS **REGRESSION LINEAR**, report the **raw score regression line (equation)** used to predict *LIBRARY TRAINING EVALUATION* from *LIBRARY KNOWLEDGE*
14. Using results from SPSS **REGRESSION LINEAR**, report the **standardized regression line** used to predict *LIBRARY TRAINING EVALUATION* from *CURRENT GPA*
15. Using results from SPSS **REGRESSION LINEAR**, report the **actual value**, **predicted value**, and **error of prediction** for the case with ID=5, when predicting *LIBRARY TRAINING EVALUATION* from *CURRENT GPA*
16. Using results from SPSS **REGRESSION LINEAR**, report the standard deviation of the errors of prediction (i.e., **standard error of the estimate**) that result from predicting *LIBRARY SKILLS* from *LIBRARY KNOWLEDGE*
17. Using results from SPSS **REGRESSION LINEAR**, report **shared variance** between *CURRENT GPA* and *AGE*
18. Using SPSS **SCATTER/DOT**, create a **scatter plot** for the independent variable (x-axis) *AGE* and dependent variable *LIBRARY TRAINING EVALUATION* — add the regression line to the scatter plot
19. Using SPSS **GRAPH INTERACTIVE SCATTERPLOT**, create a **scatter plot** for the independent variable (x-axis) *AGE* and dependent variable *LIBRARY TRAINING EVALUATION*
20. Using SPSS **SCATTER/DOT**, create a **matrix of scatter plots** for the following variables: *LIBRARY TRAINING EVALUATION*: *LIBRARY KNOWLEDGE*, *LIBRARY SKILLS*, or *CURRENT GPA*

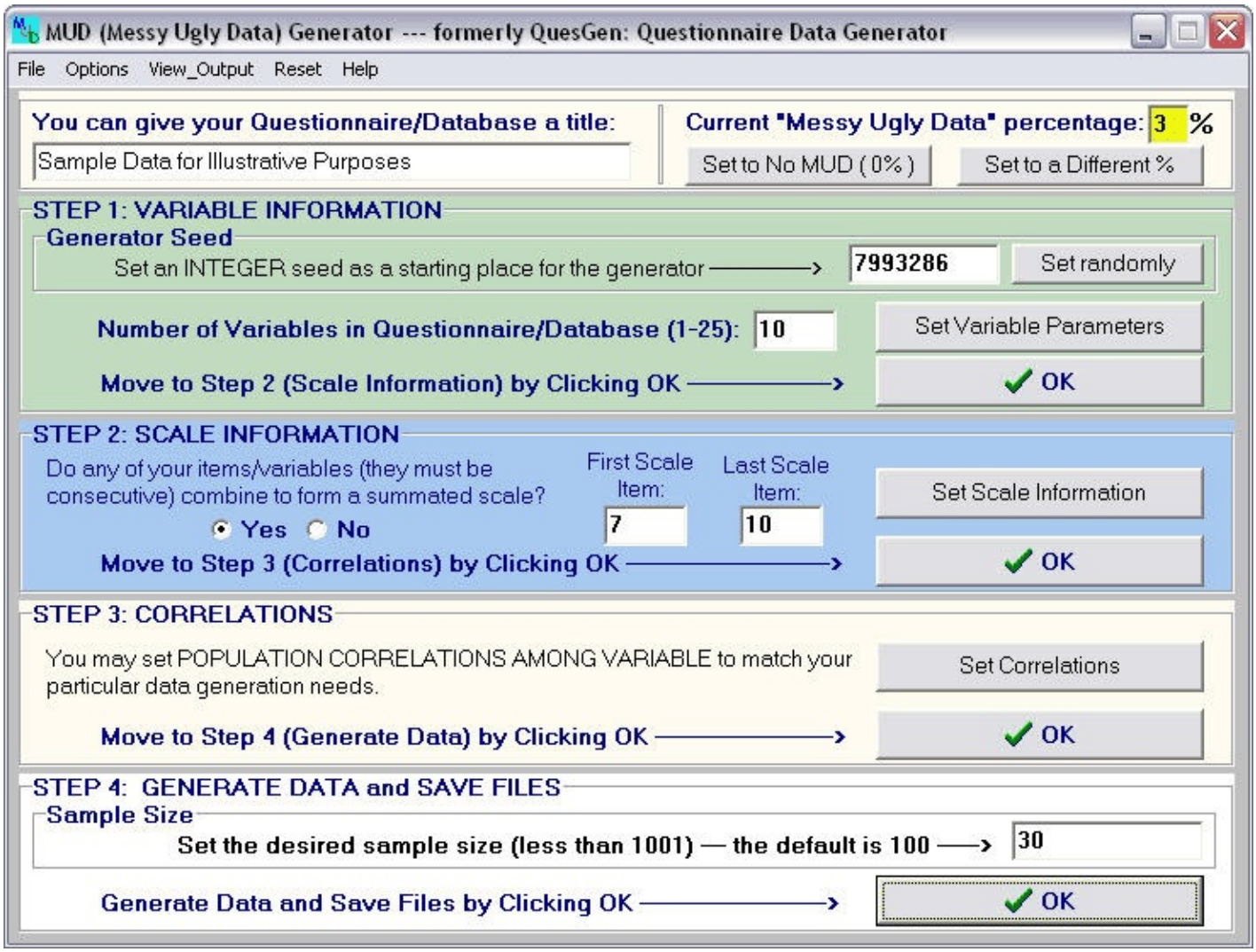
Part IX: Inferential Statistics and Hypothesis Testing

FOR EACH ITEM BELOW, you will answer research questions by performing hypothesis tests. For *each* Research Question (#1–8 below):

- provide both the Statistical Null Hypothesis and Alternative Hypothesis, both in symbols and in words
- state all appropriate statistical assumptions for the statistical method used, test the ones you can, and discuss the ones you cannot test
- use a level of significance of .05 (i.e., $\alpha = .05$)
- calculate the sample size that SHOULD have been used for the analysis based on a medium effect size (specify) and a statistical power of .80
- report and discuss your results in paragraph form **USING APA STYLE**, including:
  - (a) **descriptive** statistics, graphs, and tables appropriate for the analyses being performed to answer the research question
  - (b) the **inferential** test statistic(s), degrees of freedom, and the significance of the test statistic(s)
  - (c) the decision about any Statistical Null Hypotheses performed
  - (d) any appropriate effect sizes and/or confidence intervals
  - (e) at least one table and at least one graph to present interesting, supportive results
  - (f) an interpretation of the results in terms of the original Research Question (i.e., #1–8)
- include all necessary statistics, tables, and graphs within your written text paragraph responses; that is, although you will provide all of your SPSS outputs, you should not refer to them in your writing—rather you must report all necessary information in text and/or refer to tables and figures you have created and included in your written text paragraphs

1. Can you confidently conclude that there is a relationship between *CURRENT GPA* and *LIBRARY UNIT GRADE* in the population?. . . . . Bivariate Correlation
2. Can you confidently conclude that, in the population, *LIBRARY UNIT GRADE* can be used to predict *LIBRARY TRAINING EVALUATION*?. . . . . Regression Linear
3. Can you confidently conclude that the average population *CURRENT GPA* is different than 3.50?. . . . . One-Sample *t* Test
4. Can you confidently conclude that, in the population, students have different *LIBRARY SKILLS* and *LIBRARY KNOWLEDGE*?. . . . . Paired-Samples *t* Test
5. Can you confidently conclude that, in the population, male students differ from female students in their *LIBRARY TRAINING EVALUATION*?. . . . . Independent-Samples *t* Test
6. Can you confidently conclude that there are different proportions of males and females in the population?. . . . . Chi-Square Test
7. Can you confidently conclude that, in the population, there is a relationship between *GENDER* and *DEPARTMENT*?. . . . . Crosstabs Analysis
8. Can you confidently conclude that, in the population, students differ on *LIBRARY UNIT GRADE* depending on which *DEPARTMENT* they are in?. . . . . One-Way ANOVA

The MUD: Messy Ugly Data Generator program after all fields have been filled.



The program generates three basic types of files for students to use: (1) questionnaire files that mimic individual surveys collected by researchers, (2) databases that mimic data already entered into a computer program, and (3) a comparison of survey values, database values, and originally generated values. It generates several other files of particular use for the instructor. An example questionnaire (ID number 8) from the questionnaire file:

```
                Survey ID Number:       8
                             Age:     27
                          Gender:      1
                      Department:     -3
                             GPA:     2.71
          Library Knowledge Exam:    67.9
          Library Skills Project:     74
I learned things I did not kno:      5
 Students should be encouraged:      4
Training was not useful for me:      1
 Content covered was important:      4
```

This survey has Case 8's responses to the following hypothetical evaluation form:

Please answer the following demographic questions.

1. Age.	_____				
2. Gender.	Male (0)	Female (1)			
3. Department.	Teacher Education (1)	Educational Studies (2)	Higher Education (3)	Other (4)	
4. Current GPA.	_____				
5. What was your Library Knowledge (exam) score?	_____				
6. What was your Library Skills (project) score?	_____				

Circle the rating that most closely describes your feelings about these statements concerning our two class meetings in the library on the scale from Strongly Disagree (SD) to Strongly Agree (SA).

	points:	1	2	3	4	5
7.	I learned several things I did not know.	SD	D	N	A	SA
8.	All students should be encouraged to participate in this training.	SD	D	N	A	SA
9.	The library training was <u>not</u> useful for me.	SD	D	N	A	SA
10.	The library training covered content that was important for me to know.	SD	D	N	A	SA

The database created is a comma-delimited text file that includes variable names (designated as q01, q02, etc.). Here is an example of 10 cases (note Case 8 from above – you'll see one example of messy ugly data, a -3, for the *DEPARTMENT* variable):

```
id,q01,q02,q03,q04,q05,q06,q07,q08,q09,q10
1,26,0,2,3.90,61.0,67,4,3,3,3
2,25,0,4,3.14,85.7,90,4,4,2,5
3,24,1,4,3.66,100.0,100,5,5,2,54
4,25,1,2,3.64,73.2,78,4,3,2,4
5,26,0,4,3.49,80.1,84,5,4,2,5
6,24,1,4,2.68,54.3,62,4,4,2,5
7,23,0,4,1.96,60.4,68,5,4,2,4
8,27,1,-3,2.71,67.9,74,5,4,1,4
9,23,0,4,2.66,84.7,89,4,4,1,5
10,23,1,2,2.95,85.5,90,4,3,1,4
```

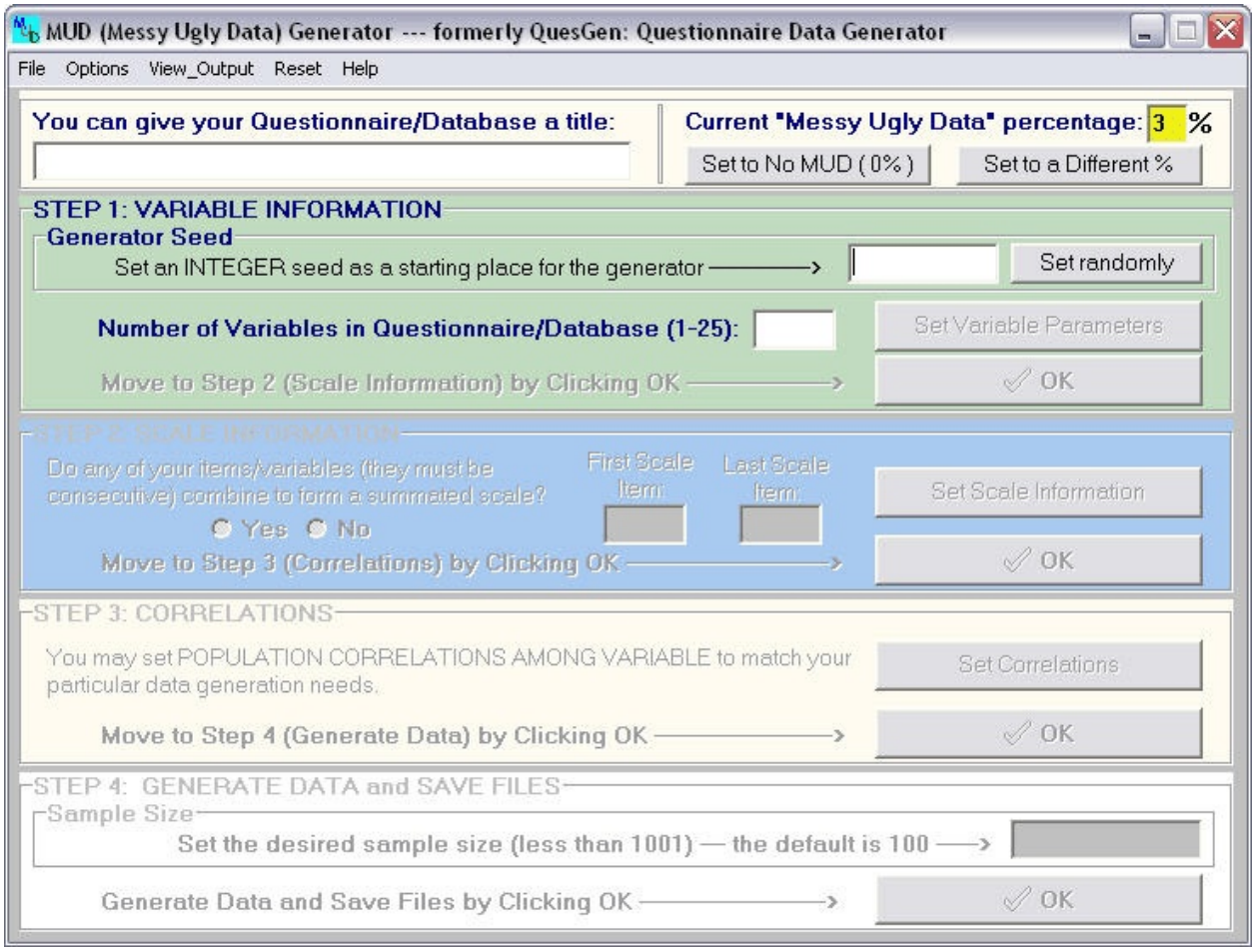
The third primary file of interest can be used to fix the file back to its original form. The comparison file shows the original generated value, the survey value, and the database value (which is not always the same as the survey value).

Survey	Variable Name	Original Value	Survey Value	Dataset Value	S<>D
3	Content covered was important	5	5	54	**
8	Department	1	-3	-3	
11	Content covered was important	5	5	55	**
13	GPA	2.96	2.96	0.30	**
16	Department	4	4	43	**
19	Library Skills Project	83	.	.	
25	Gender	1	4	4	
25	I learned things I did not kno	3	3	33	**
28	Age	25	125	125	

```
================================================================================
A total of 9 data points were changed based on the rules set by user.
A total of 8 cases were changed based on the rules set by user.
A total of 5 cases have surveys different from values in the dataset (S<>D).
================================================================================
```

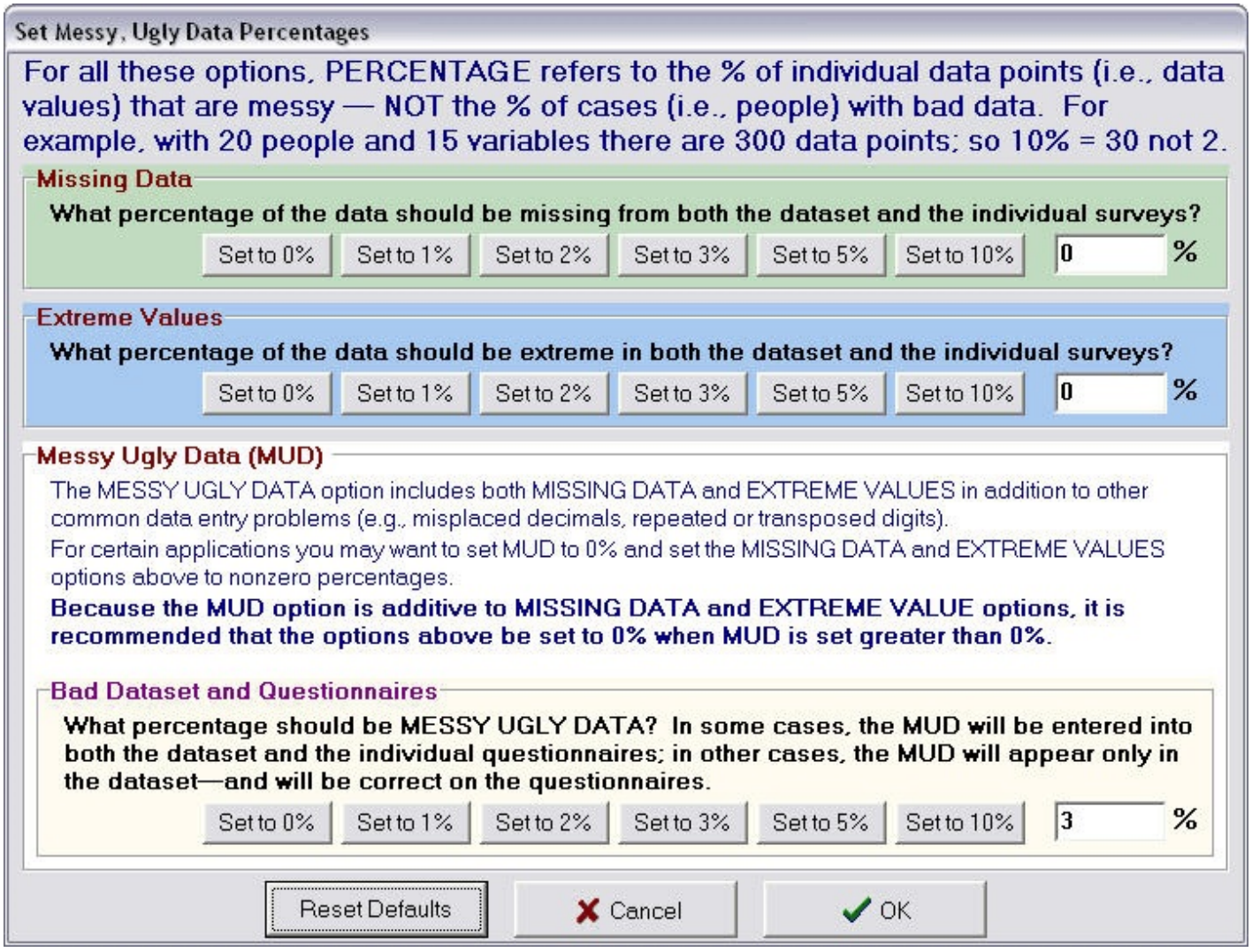
OK, let's start generating some data...

Before the fields are filled in sequentially, many options are unavailable and sometimes gray. For example, until after Step 1 has been completed, the Step 2, Step 3, and Step 4 boxes are not able to be accessed.



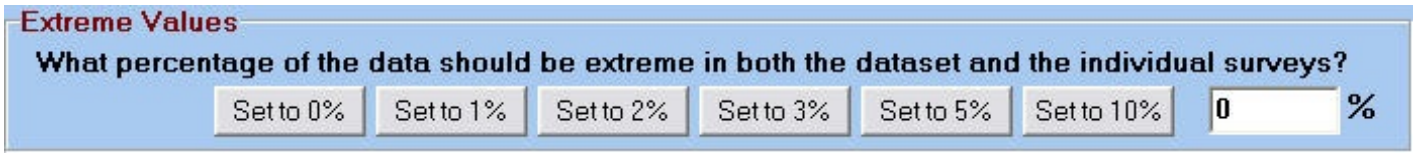
The first step in the MUD (Messy Ugly Data) data generation process is to provide a name or **Title** for your database and to determine how much MUD you want. By default, 3% MUD is generated. This usually works out pretty well, but for real practice in finding bad data you can ask for more. You can also ask for no MUD (i.e., 0%) if there is no need to practice screening data.

By clicking the **Set to a Different %**, you can open a dialog box to allow you greater control over MUD generation.

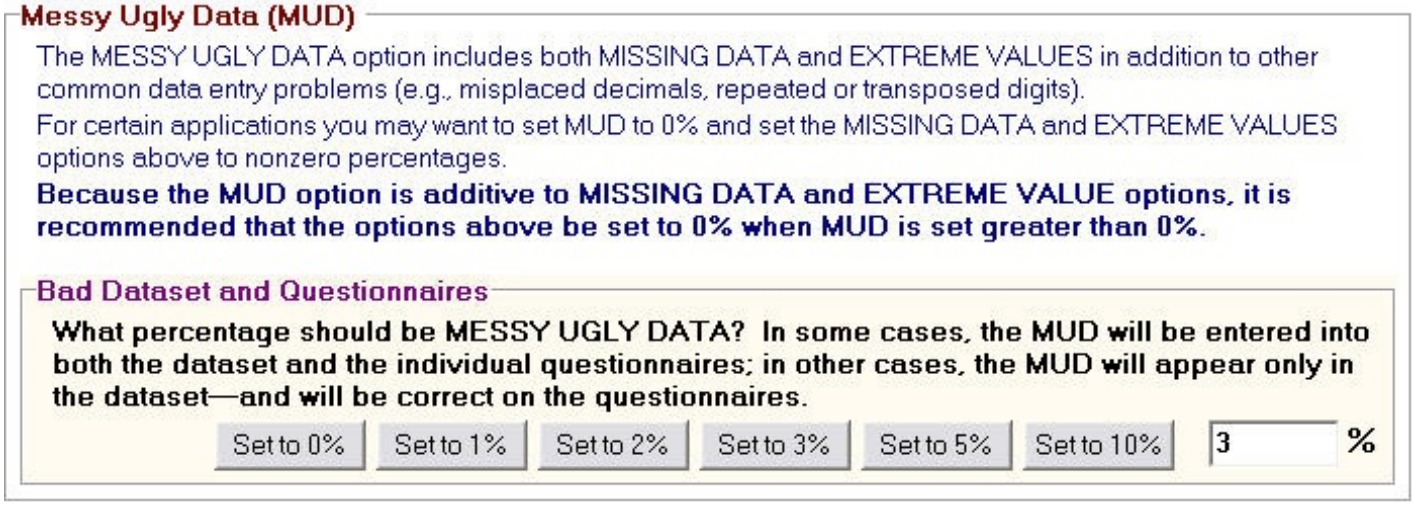


Note the comments at the top of the screen. The generator works such that a given percentage of total data is messy ugly, not the number of cases with messy ugly data. Individual cases can have multiple messy ugly data values. It is possible to designate different percentages for different kinds of messy ugly data: Missing Data, Extreme Values, and Messy Ugly Data.

MUD begins by generating complete data sets that follow the parameters given by the user. During the data generation process, however, data values are randomly set as missing at the rate designated. Both the surveys and the database are changed as a result of Missing Data.



Similarly, MUD will randomly set data values to scores outside a normal range of values. Essentially, data values are created at an appropriately random rate that are more than 3 standard deviations above or below the mean. The algorithm used is: **PopMean ± [PopSD * (3+random value between 0 and 1)]**. Both the surveys and the database are changed as a result of Extreme Values.



A complicated process is used to generate MUD. At the rate given, data values are changed randomly based on several rules.

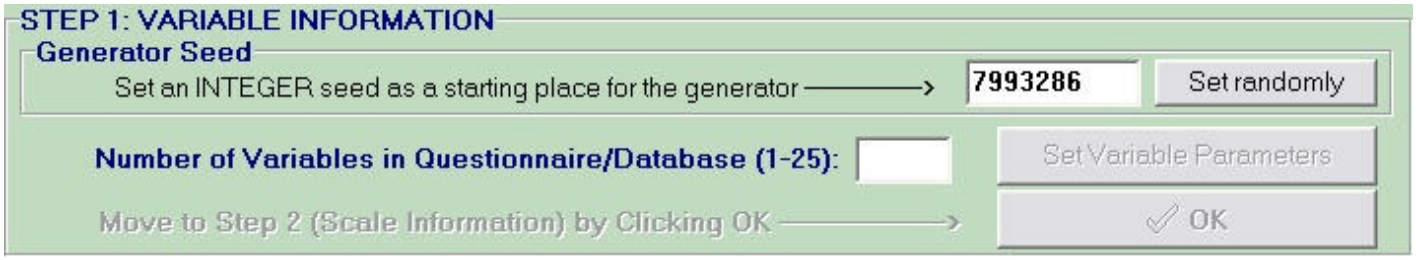
- Type 1 MUD follows the rule given for extreme values (above)
- Type 2 MUD sets data values to commonly used missing value codes (e.g., 9, 99, 999, -1, system missing)
- Type 3 MUD repeats a digit for numbers under 100 (e.g., 2 becomes 22, 3.31 becomes 33.31, 93.4 becomes 993.4), it adds 1000 to larger numbers (e.g., such that 100 becomes 1100)

- Type 4 MUD randomly changes numbers of categorical variables to letters (sometimes the correct letter where 1 becomes A, sometimes the wrong letter where B might become 3); doubles or halves scale values (e.g., 4 randomly becomes either 2 or 8); multiples or divides scale values by 5; or adds 10, 100, or 1000 depending on the size of the original value
- Type 5 MUD adds a digit one lower than an original digit (e.g., 2 becomes 21), makes a value negative (e.g., 23 becomes -23), or moves the decimal place (e.g., 2.63 becomes 26.30, or 88.6 becomes 8.9)
- Type 6 MUD changes a value to the (Maximum + 2), changes a value to the (Minimum – 2), transposes digits (e.g., 73 becomes 37, 65.4 becomes 4.7 due to rounding to 1 decimal place), or adds 10000 to very large numbers (e.g., 12345 becomes 22345)

Randomly,

- sometimes both the surveys and the database are the same, but are changed from the original data,
- sometimes the surveys remain as originally generated while only the database is changed, and
- sometimes the survey is changed from the original data and the database is changed from the survey data.

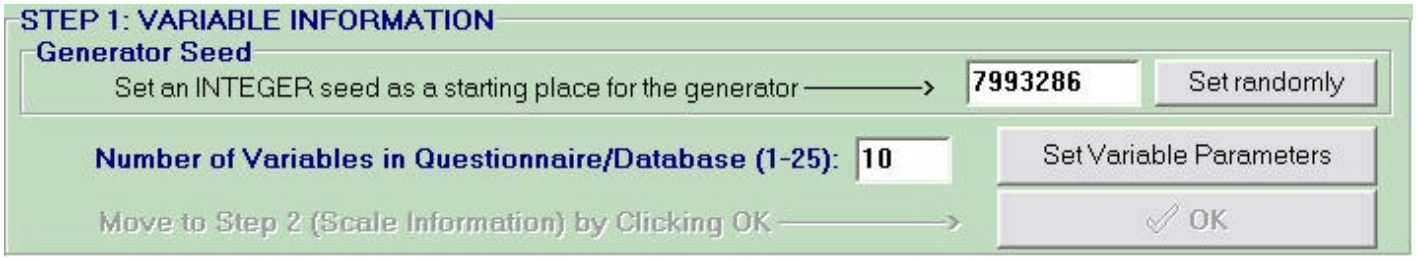
After clicking OK to set changes, you are returned to the main MUD program screen. It is now time to move on to **STEP 1: VARIABLE INFORMATION**, where you begin by setting a **Generator Seed**.



An important property of computerized pseudorandom number generators (with both good and bad implications), is that the numbers generated are not truly random. While this can result in obviously non-random data when the computerized generator is poorly implemented, it is useful for methodological researchers because a stream of pseudorandom numbers can be exactly duplicated when the same seed is used repeatedly.

Because the computerized generator uses a mathematical function that derives new pseudorandom numbers from previous ones, the process needs a starting place. This is the purpose of the **Generator Seed** — to provide a starting place for the generation function.

Next, we need to set the **Number of Variables** (or items) in the questionnaire/database. For our continuing example in this paper, we'll use 10 variables.



After setting the **Number of Variables**, we can **Set Variable Parameters**. The variable parameters define the population parameters for our variables. The pseudorandom number generation process will create data that would reasonably be sampled from the given population. The screen below has already had the variable information filled in for our example.

Step 1: Variable Information

Cancel OK

Defaults are standard normal data. To create data to mimic test percentages, maybe set M=75, SD=10, Min=0, Max=100, and # Decimals=0 or 1. Setting SD is the hardest part. When in doubt, SD=M/5 usually seems to work well. Or you can use the fact that roughly 95% of your normally distributed data will be between (M-2SD) and (M+2SD), but note Min/Max values. To create roughly equal-size groups, set Decimals=0, set MIN to lowest group # and MAX to highest. For 2 groups (0,1), use M=0.5 and SD=1. For more than 2 groups, use M=(Min+Max)/2 and SD=M*0.6. For example, for 3 groups (1-3), M=(1+3)/2=2 and SD=M*0.6=1.2; for 5 groups (1-5), M=(1+5)/2=3 and SD=M*0.6=1.8. Note that the group data will actually be ORDINAL.

To edit a value, you'll need to double click the appropriate cell or press F2 or Enter in the appropriate cell.

Question #	Variable Name (30 characters)	Pop Mean	Std Dev	# Decimals	Min Score	Max Score
Q_01	Age	24	2	0	20	
Q_02	Gender	0.7	1	0	0	1
Q_03	Department	2.5	1.5	0	1	4
Q_04	GPA	3.2	0.6	2	0	4
Q_05	Library Knowledge Exam	75	10	1	0	100
Q_06	Library Skills Project	80	8	0	0	100
Q_07	I learned things I did not know	4	0.5	0	1	5
Q_08	Students should be encouraged	3.5	0.5	0	1	5
Q_09	Training was not useful for me	2	0.5	0	1	5
Q_10	Content covered was important	4	0.5	0	1	5

Number of Items / Variables: 10

Set column values to the default values (in parentheses) for:

Set ALL Columns to Defaults | Variable Name (Q_##) | # Decimals (2) | Pop Mean (0.0) | Pop SD (1.0) | Min Score (none) | Max Score (none)

Set all column values equal to the value in ROW 1 for:

Set ALL Columns to Values in ROW 1 | Var Name | Decimals | Pop Mean | Pop SD | Minimum | Maximum

OK saves changes and/or sets any empty cells to their default values and closes this window.
CANCEL resets cells to blanks and closes this window.

Cancel OK

The key to the MUD program is that data are generated from a survey research perspective. That is, all data are from a hypothetical administration of a survey. Here the data are based on the evaluation form illustrated earlier. The actual sample statistics will not equal the population parameters set here. Rather, the sample statistics will reflect reasonable sample values from a population with these parameter values.

Question #	Variable Name (30 characters)	Pop Mean	Std Dev	# Decimals	Min Score	Max Score
Q_01	Age	24	2	0	20	
Q_02	Gender	0.7	1	0	0	1
Q_03	Department	2.5	1.5	0	1	4
Q_04	GPA	3.2	0.6	2	0	4
Q_05	Library Knowledge Exam	75	10	1	0	100
Q_06	Library Skills Project	80	8	0	0	100
Q_07	I learned things I did not know	4	0.5	0	1	5
Q_08	Students should be encouraged	3.5	0.5	0	1	5
Q_09	Training was not useful for me	2	0.5	0	1	5
Q_10	Content covered was important	4	0.5	0	1	5

Variables should be defined as realistically as possible. That is, if a variable has a Minimum or Maximum value, you should set it. You should also try to define a reasonable Mean and Standard Deviation for the variable. An appropriate number of decimal places should be set (e.g., if data are always integer, set Decimals to 0, if using GPA maybe set Decimals to 2). A reasonable Mean is usually relatively easy to determine, but designating a Standard Deviation is not often easy.

Through trial and error, two ways have emerged as the most reasonable methods for setting the SD. Very frequently, the Mean divided by 5 works quite nicely. However, it is usually preferable to use the properties of the normal distribution to set a Standard Deviation (data are generated following a normal distribution). Therefore, using the (Mean ± 2 SD) as the range for 95% of your data can usually be accomplished fairly easily. For example, if a mean test score is 75, then you might expect 95% of the data to be between 55 and 95 — set SD=10.

Note the buttons for default options. You can set all data to their default values, set just one column (e.g., Mean, SD, Decimals, etc.) to default values, set all values equal to whatever is set for Q_01 (if all variables are measured with similar scales, such as items of an scale or instrument), or set all values in one column to the value set for that variable in Q_01. The buttons for defaults indicate what those defaults are.

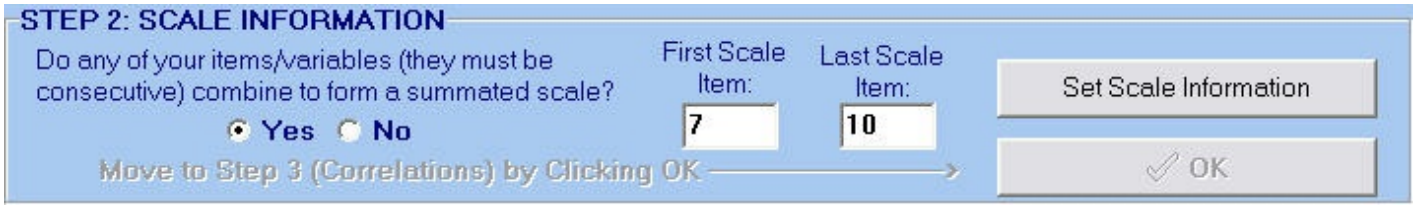
Several recommendations are listed at the top of the Step 1: Variable Information screen.

A few other things to note:

- no Minimum or Maximum values must be set, only one or the other, or both
- integer data will have 0 decimals
- the variable name will only allow 30 characters
- several variables (items) may be combined to form a summated scale—but must be listed consecutively here

After finishing with the variable definitions, the **STEP 2: SCALE INFORMATION** box becomes accessible. The first decision is whether any of your variables (or items) will be combined to form a total score from a summated scale. You should click either **Yes** or **No**.

If there is no summated scale imbedded within your questionnaire, when you click **No**, the **OK** button becomes active so that you may move directly to Step 3 (Correlations).



If there is a summated scale within your survey, when you click **Yes**, the two boxes for **First Scale Item** and **Last Scale Item** become active. In these boxes you will put the item number for the first item and last, respectively. Remember, these items must be listed consecutively in **STEP 2: VARIABLE INFORMATION**.

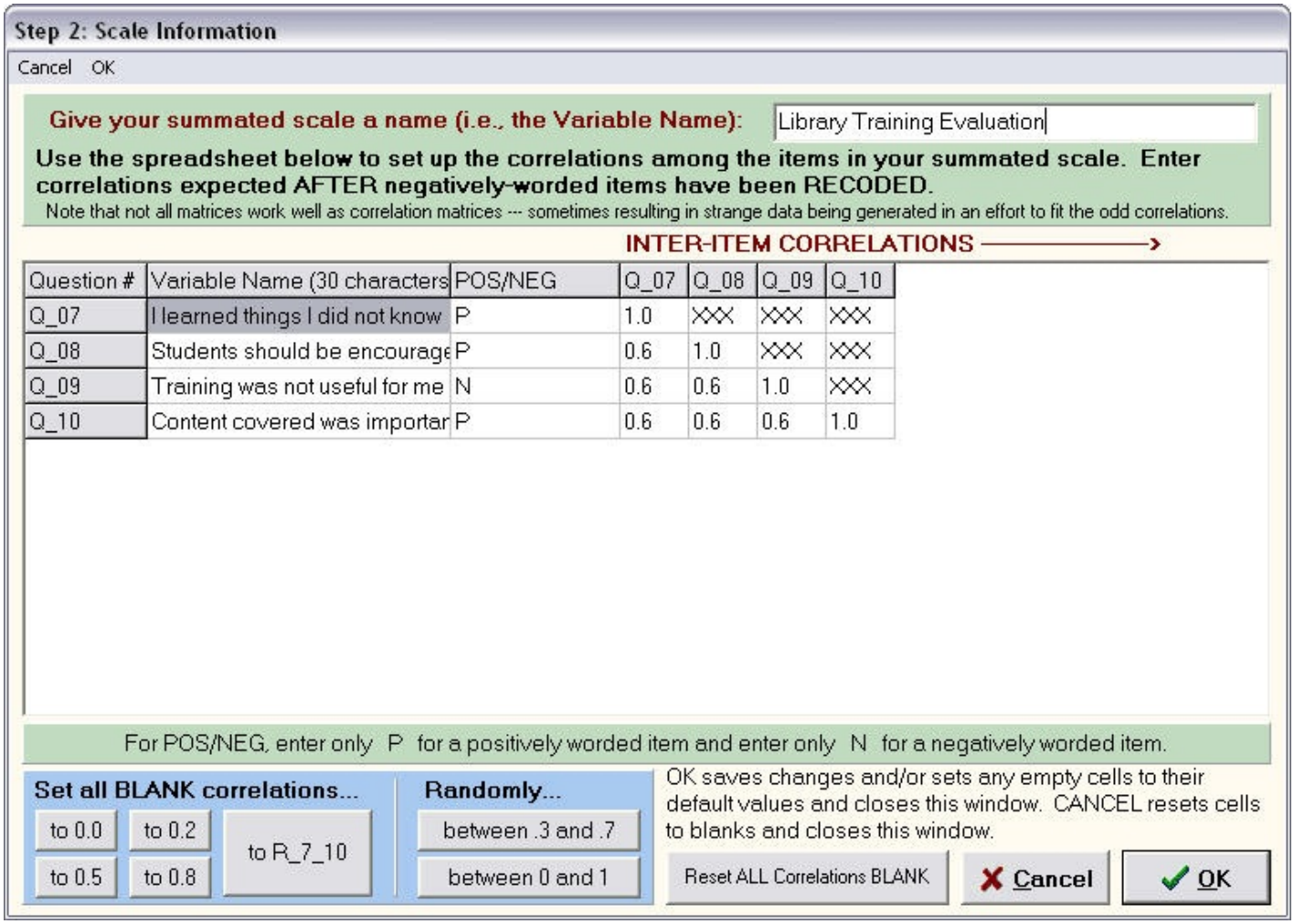
In the screen above, these values have been set to 7 and 10, respectively, for our example survey.

After designating the item numbers for the scale, you must **Set Scale Information**.

The first task is to name the scale (not required). This should be the name of the variable that will be measured by the total score for the items designated in the scale.

After naming your scale, you can list the items (up to 30 characters) and whether they are **Positively** or **Negatively** worded. For example, in our scale, 3 of the 4 items are worded such that if a respondent agrees with the items they are indicating a positive evaluation of (or attitude toward) the library training. However, Item 9 ('training was not useful for me") is negatively worded, which means that disagreement suggests a more positive evaluation of the training.

MUD generates data such that all items are scaled in the same, positive direction. After the scale data are generated, MUD then recodes negatively worded items so that they are appropriate for the actual survey questions.

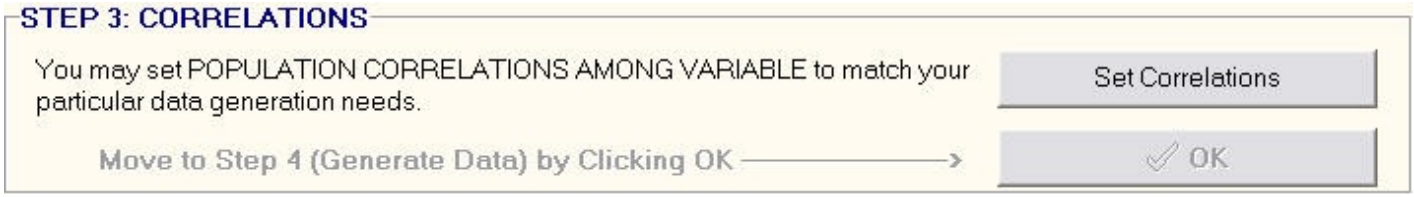


The last requirement is to provide a correlation matrix for the items. The correlations are based on all items having the same directionality (i.e., positive wording).

It is important to note, that just like with the Means and Standard Deviations, the actual sample Correlation values generated will not be exactly the same as the population values set here. Rather they will be sample correlations determined by the random sample of data created by the MUD program.

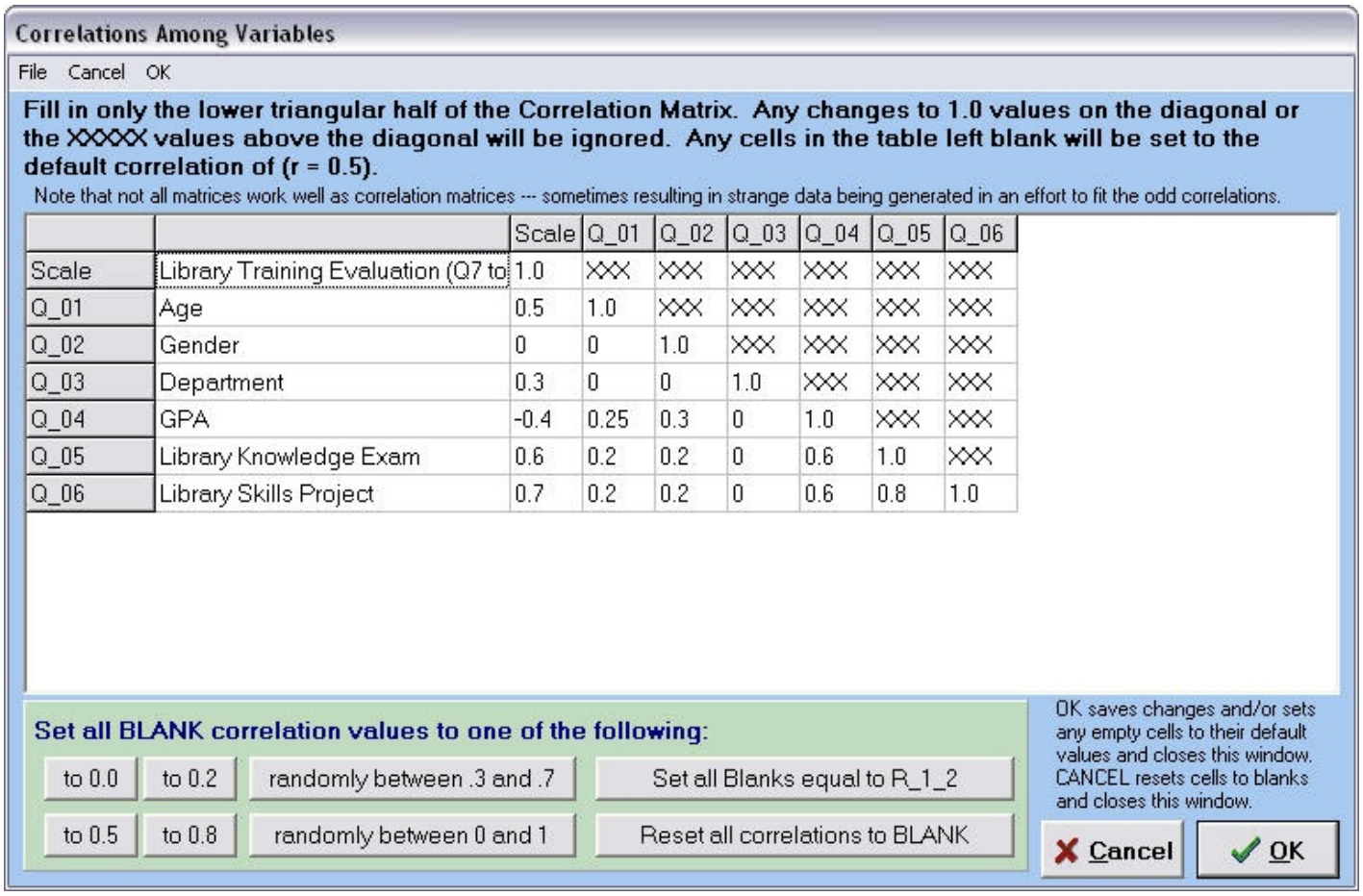
Here, with most scale items expected to measure roughly the same thing (an attitude toward or evaluation of the library training), all correlations are set to be the same moderately strong correlation (0.6).

Note that there are several options available for setting the correlations automatically. But note that these will only change those correlations that are currently BLANK (i.e., the cells in the table are empty). Also, do not change the XXXXX values in the upper triangular half of the matrix — only fill in correlations in the lower triangular half.



After you set the correlations for the scale items, **STEP 3: CORRELATIONS** becomes active. You must click on the **Set Correlations** button in order to set the correlations among the variables in your survey.

It is important to note again, that just like with the Means and Standard Deviations and the Scale Item Correlations, the actual sample Variable Correlation values generated will not be exactly the same as the population values set here. Rather they will be sample correlations for the random sample created by MUD.



		Scale	Q_01	Q_02	Q_03	Q_04	Q_05	Q_06
Scale	Library Training Evaluation (Q7 to	1.0	XXX	XXX	XXX	XXX	XXX	XXX
Q_01	Age	0.5	1.0	XXX	XXX	XXX	XXX	XXX
Q_02	Gender	0	0	1.0	XXX	XXX	XXX	XXX
Q_03	Department	0.3	0	0	1.0	XXX	XXX	XXX
Q_04	GPA	-0.4	0.25	0.3	0	1.0	XXX	XXX
Q_05	Library Knowledge Exam	0.6	0.2	0.2	0	0.6	1.0	XXX
Q_06	Library Skills Project	0.7	0.2	0.2	0	0.6	0.8	1.0

You need to set a correlation matrix among the variables in your survey. The first item will be your scale (if you defined one). Other variables will be identified by their Item Number (e.g., Q_01 or Q_02), with their names labeled on the rows (columns will have only the Item Numbers). You should determine a reasonable correlation between your variables (note that these are guesses, but try to be reasonable).

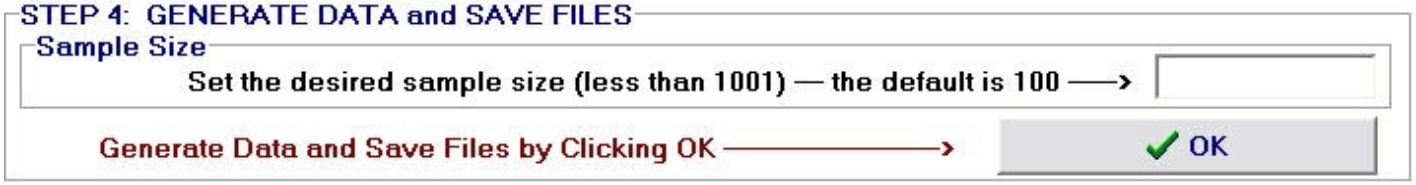
For example, expecting no correlation between *GENDER* and *AGE*, *GENDER* and *DEPARTMENT*, *DEPARTMENT* and *AGE*, *DEPARTMENT* and *GPA*, *DEPARTMENT* and *LIBRARY KNOWLEDGE*, and *DEPARTMENT* and *LIBRARY SKILLS*, all these correlations have been set to 0. Similarly, no correlation is expected between *GENDER* and the *LIBRARY TRAINING EVALUATION* scale.

Low correlations are expected between *GENDER* and *LIBRARY KNOWLEDGE*, *GENDER* and *LIBRARY SKILLS*, *DEPARTMENT* (as an ordinal variable) and the *LIBRARY TRAINING EVALUATION* scale, *AGE* and *LIBRARY KNOWLEDGE*, *AGE* and *LIBRARY SKILLS*, *AGE* and *GPA*, and *GENDER* and *GPA*. These are all set arbitrarily between 0.2 and 0.3.

A moderate correlation is expected between *AGE* and the *LIBRARY TRAINING EVALUATION* scale (0.5). Notice that a moderate NEGATIVE correlation is expected between *GPA* and the *LIBRARY TRAINING EVALUATION*, based on the theory that better students already knew most of what was taught in the library training and therefore found it to be a waste of time.

Relatively large correlations were arbitrarily set between 0.6 and 0.8 among the remaining pairs of variables.

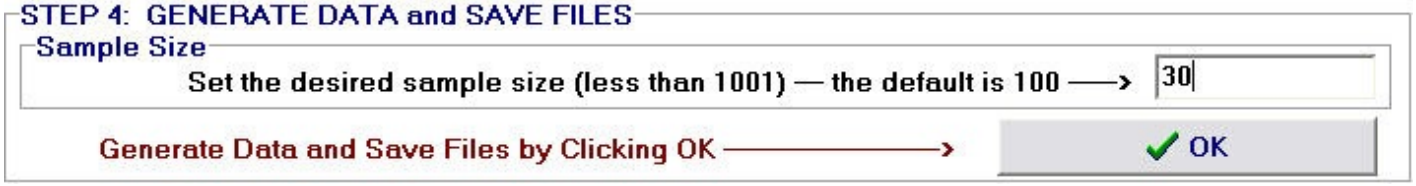
Note that there are several options available for setting the correlations automatically. But note that these will only change those correlations that are currently BLANK (i.e., the cells in the table are empty). Also, do not change the XXXXX values in the upper triangular half of the matrix — only fill in the lower triangular half.



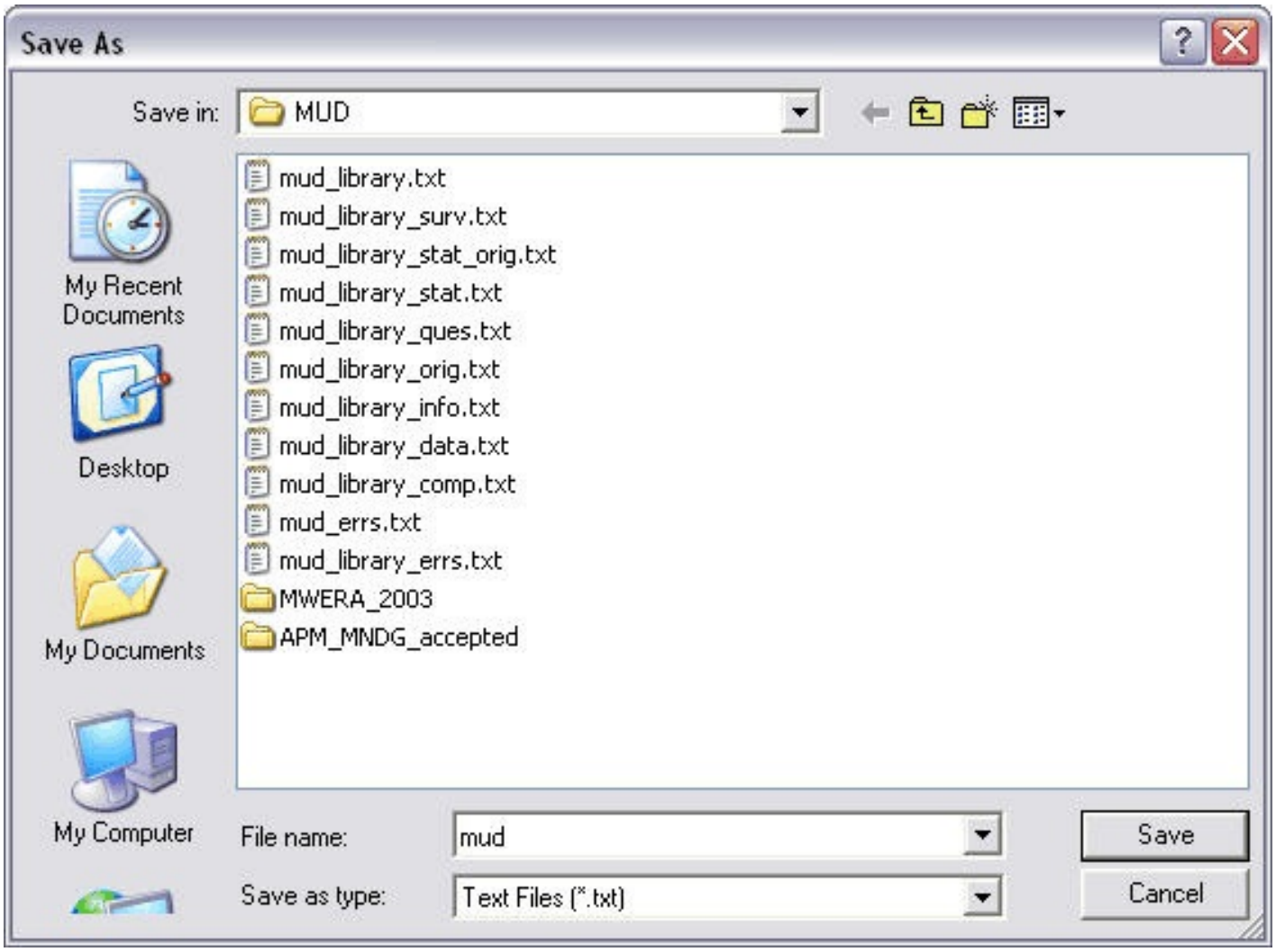
After you set the correlations for the scale items, **STEP 4: GENERATE DATA AND SAVE FILES** becomes active. You must click on **OK** button in order to set the correlations among the variables in your survey.

The first task is to determine your sample size. This can be based on a variety of factors, including perhaps statistical power. An instructor might generate hundreds of cases from which students will sample for their own data analysis projects. Or the instructor (or student) might generate just the number of cases needed for their own sample.

Here we have chosen to generate a sample of 30, which is a reasonable size for an introductory graduate research class.



After setting the sample size, you click OK to generate and save the data. The Windows Save As dialog box appears.



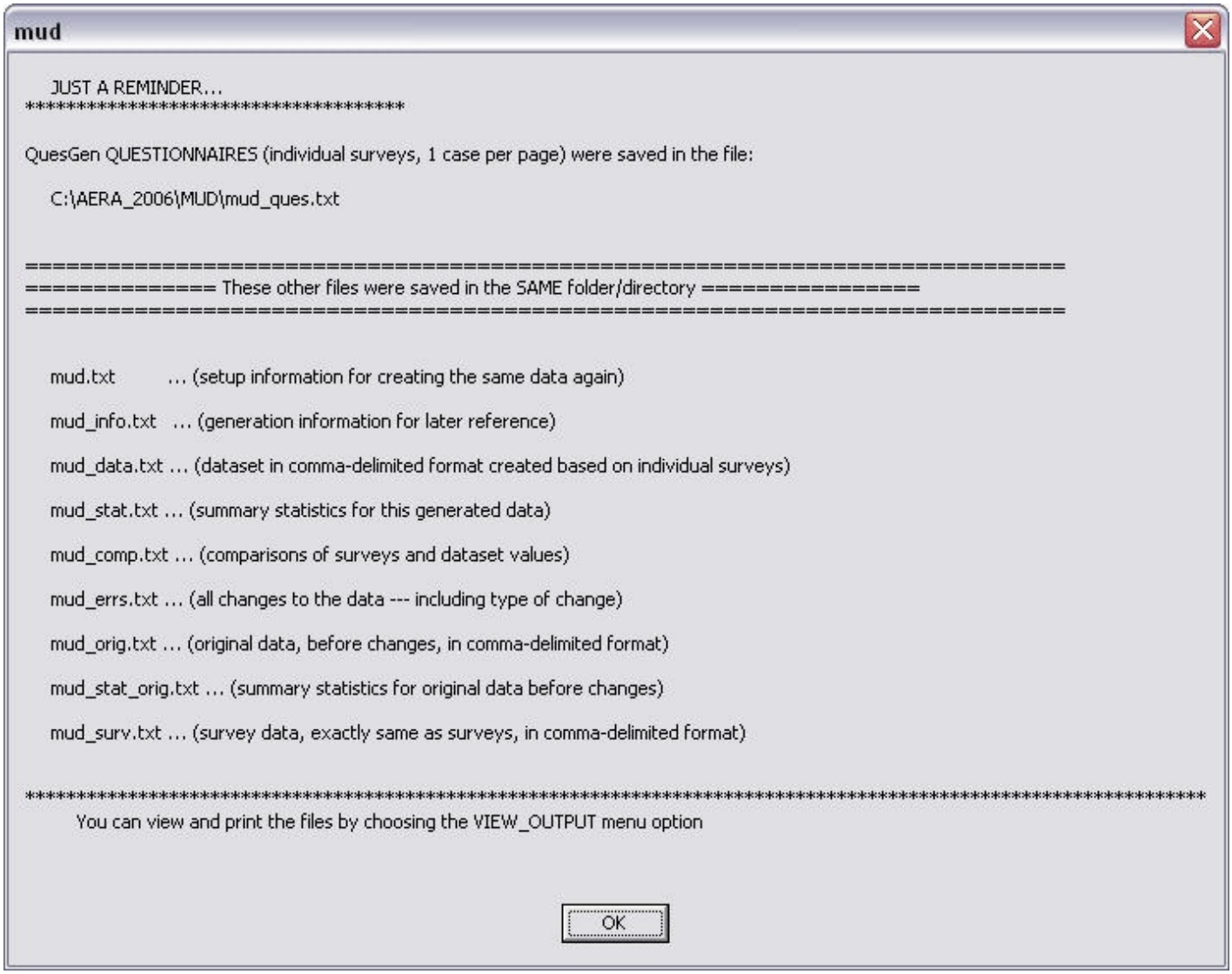
The file will be named "mud" by default, but can be changed or lengthened (e.g., mud_library used for the sample library data used in this paper).

Notice that there are several files saved. They are all basic TEXT files. The most important files for the purposes of this paper are the following:

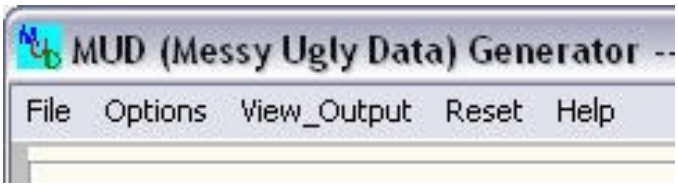
- the survey/questionnaire file (??????????_ques.txt)
- the data file (??????????_data.txt)
- the comparison file (??????????_comp.txt)

These files were illustrated at the beginning of this document.

The reminder screen that appears after the file has been saved lists the files and their contents.



After data are generated, you can view the various files that are created by MUD using the View_Output menu option.



The Information File appears by default when the View Output window opens.

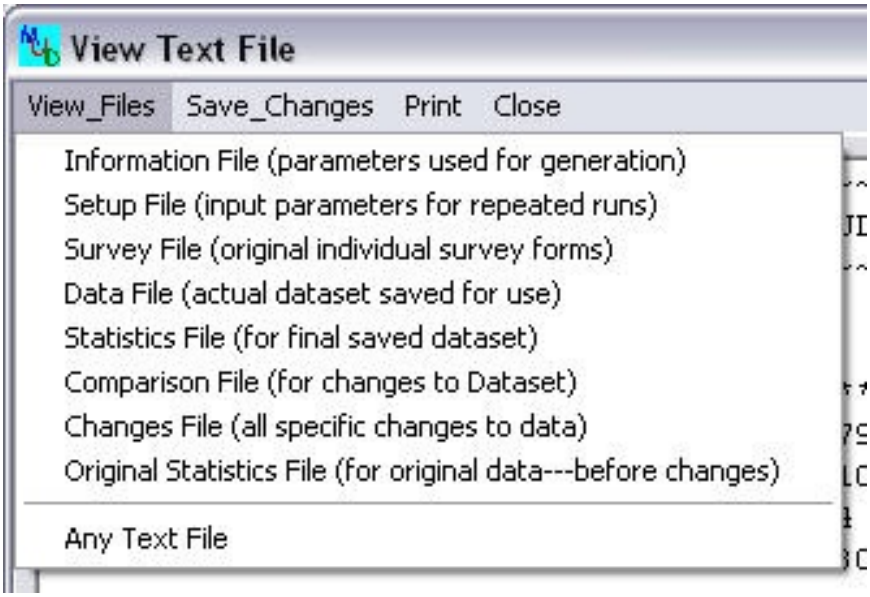
View Output --- NOW SHOWING: Information File

View_Files Save_Changes Print Close

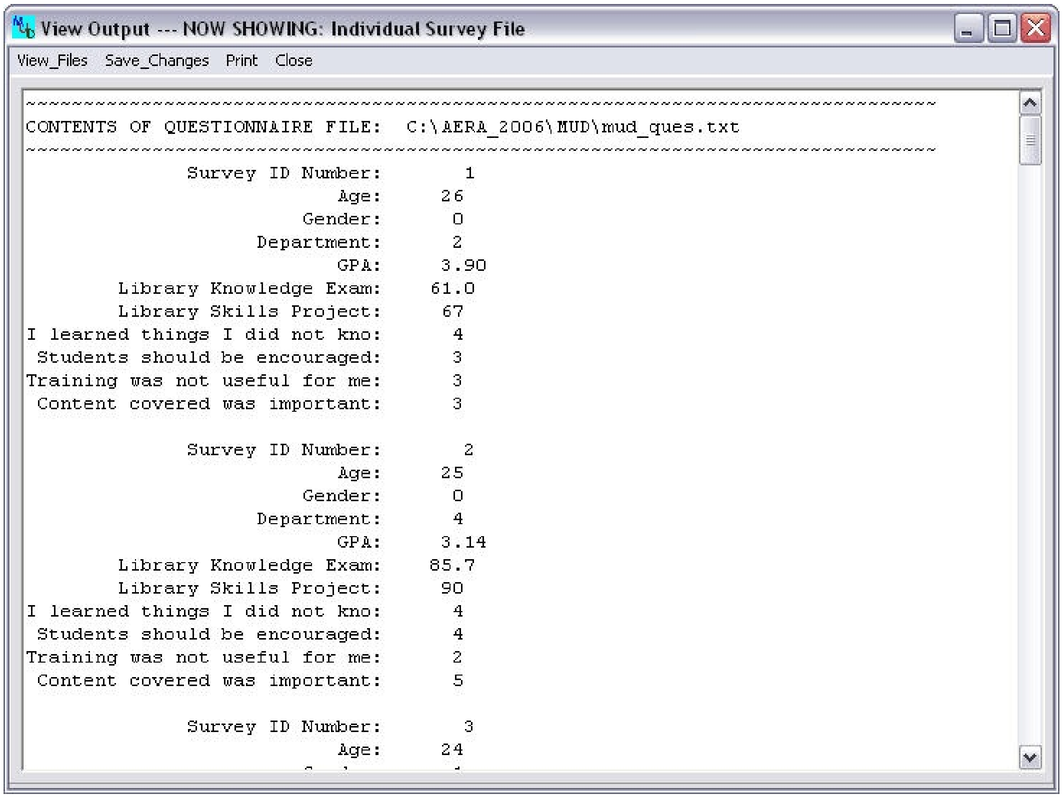
```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
CONTENTS OF INFORMATION FILE NAMED: C:\AERA_2006\MUD\mud_info.txt
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Sample Data for Illustrative Purposes

**********************************************************
                      Generator Seed: 7993286
     Number of Total Items/variables: 10
Number of Items Identified for Scale: 4
               Total Number of Cases: 30

             Missing Data Percentage:   0%
             Extreme Data Percentage:   0%
        Messy Survey Data Percentage:   3%

========================================================
Variable Information (population parameters:
========================================================
```

Q_##	Variable Name	Pop Mean	Std Dev	#Decimals	Min Score	Max Score
Q_01	Age	24.000	2.000	0	20.000	n/a
Q_02	Gender	0.700	1.000	0	0.000	1.000
Q_03	Department	2.500	1.500	0	1.000	4.000
Q_04	GPA	3.200	0.600	2	0.000	4.000
Q_05	Library Knowledge Exam	75.000	10.000	1	0.000	100.000
Q_06	Library Skills Project	80.000	8.000	0	0.000	100.000

Other files can be opened from the View Output menu.



For example, the Questionnaire/Survey file:

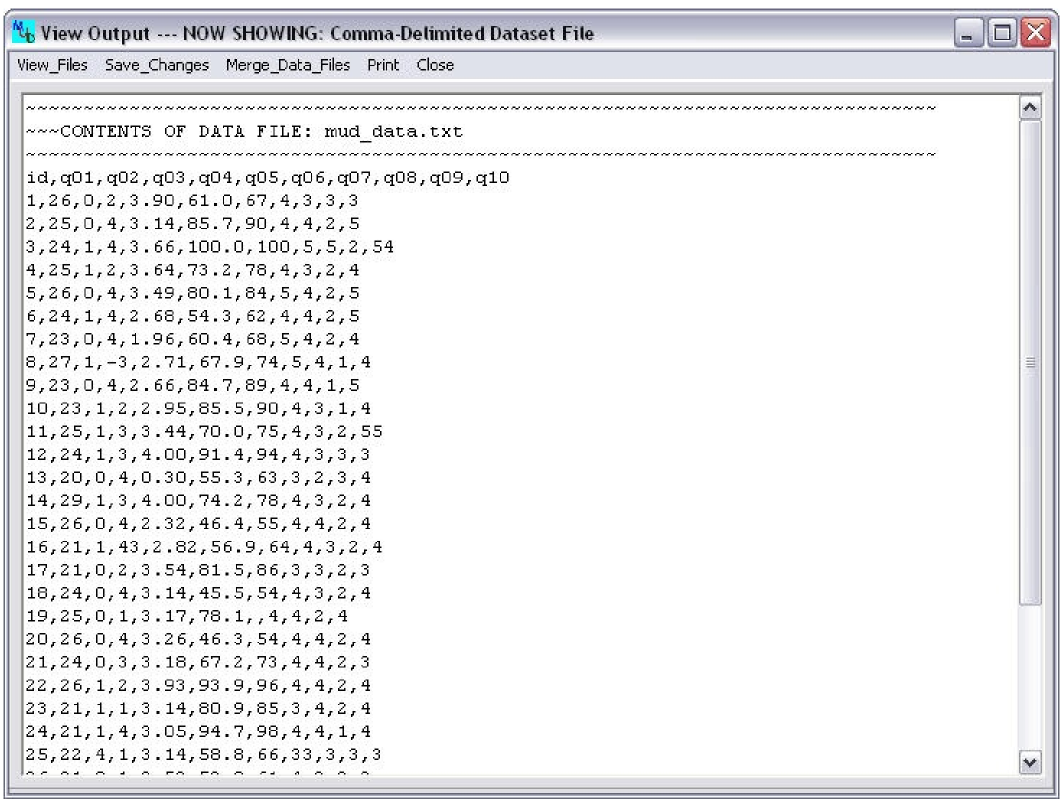


View Output --- NOW SHOWING: Individual Survey File

View_Files Save_Changes Print Close

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
CONTENTS OF QUESTIONNAIRE FILE:  C:\AERA_2006\MUD\mud_ques.txt
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
              Survey ID Number:       1
                           Age:      26
                        Gender:       0
                    Department:       2
                           GPA:     3.90
        Library Knowledge Exam:    61.0
        Library Skills Project:      67
I learned things I did not kno:       4
 Students should be encouraged:       3
Training was not useful for me:       3
 Content covered was important:       3

              Survey ID Number:       2
                           Age:      25
                        Gender:       0
                    Department:       4
                           GPA:     3.14
        Library Knowledge Exam:    85.7
        Library Skills Project:      90
I learned things I did not kno:       4
 Students should be encouraged:       4
Training was not useful for me:       2
 Content covered was important:       5

              Survey ID Number:       3
                           Age:      24
```

or the Data File

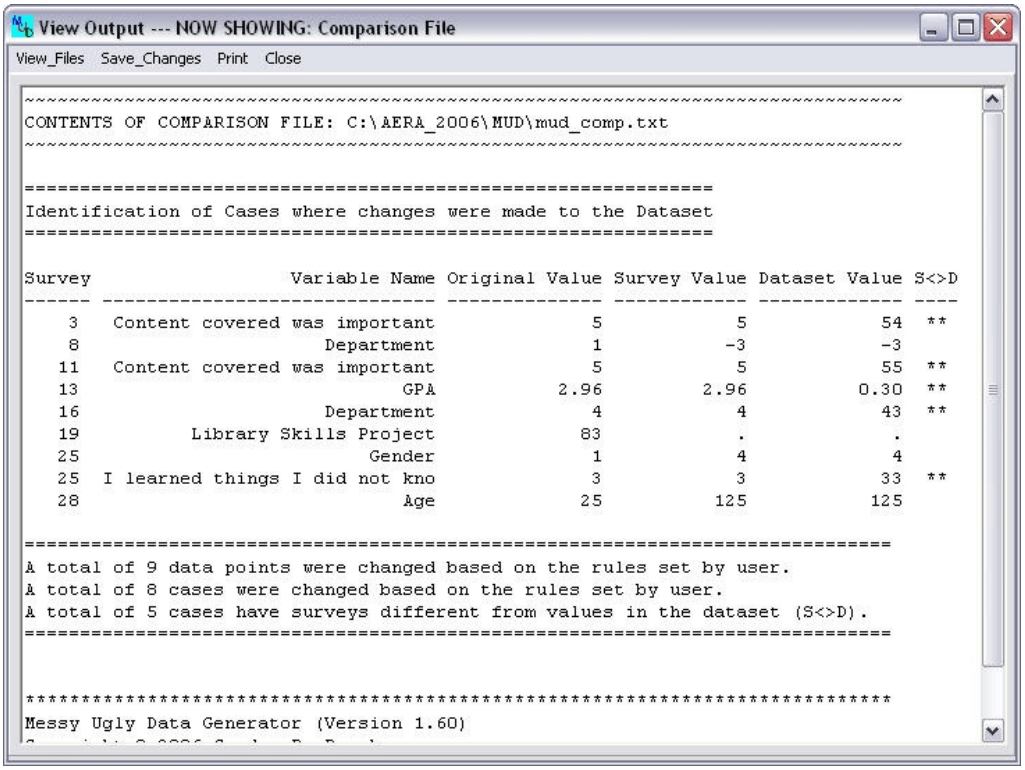


View Output --- NOW SHOWING: Comma-Delimited Dataset File

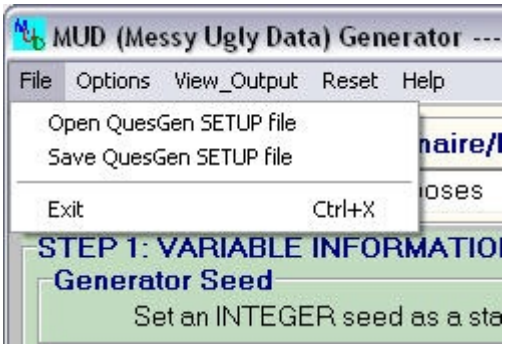
View_Files Save_Changes Merge_Data_Files Print Close

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~CONTENTS OF DATA FILE: mud_data.txt
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
id,q01,q02,q03,q04,q05,q06,q07,q08,q09,q10
1,26,0,2,3.90,61.0,67,4,3,3,3
2,25,0,4,3.14,85.7,90,4,4,2,5
3,24,1,4,3.66,100.0,100,5,5,2,54
4,25,1,2,3.64,73.2,78,4,3,2,4
5,26,0,4,3.49,80.1,84,5,4,2,5
6,24,1,4,2.68,54.3,62,4,4,2,5
7,23,0,4,1.96,60.4,68,5,4,2,4
8,27,1,-3,2.71,67.9,74,5,4,1,4
9,23,0,4,2.66,84.7,89,4,4,1,5
10,23,1,2,2.95,85.5,90,4,3,1,4
11,25,1,3,3.44,70.0,75,4,3,2,55
12,24,1,3,4.00,91.4,94,4,3,3,3
13,20,0,4,0.30,55.3,63,3,2,3,4
14,29,1,3,4.00,74.2,78,4,3,2,4
15,26,0,4,2.32,46.4,55,4,4,2,4
16,21,1,43,2.82,56.9,64,4,3,2,4
17,21,0,2,3.54,81.5,86,3,3,2,3
18,24,0,4,3.14,45.5,54,4,3,2,4
19,25,0,1,3.17,78.1,,4,4,2,4
20,26,0,4,3.26,46.3,54,4,4,2,4
21,24,0,3,3.18,67.2,73,4,4,2,3
22,26,1,2,3.93,93.9,96,4,4,2,4
23,21,1,1,3.14,80.9,85,3,4,2,4
24,21,1,4,3.05,94.7,98,4,4,1,4
25,22,4,1,3.14,58.8,66,33,3,3,3
```

or the Comparison File



The SETUP file is saved automatically when data are generated, but can also be saved again or prior to data being generated by using the option of the File Menu. Similarly, previously created SETUP files can be opened for use at a later date.



The sample data used in this paper can be obtained from the Help Menu.