SESUG Paper 197-2019

Making Life Easier on the SAS® End: Best Practices for Collecting Survey Data

Julie Plano, Keli Sorrentino, Yale University

ABSTRACT

An analysis-ready SAS dataset is the deliverable, the one and only thing the analysts are waiting for. It is the goal of any research project. Much effort goes into the collecting and cleaning of data before it is ready to be analyzed. Our research team has spent decades learning to design good data collection instruments, resulting in purposeful survey questions with practical response values. Whether the questionnaire is administered in person or collected via online software, it is important to begin with a clear and concise plan. Here we share our expertise and best practices to help research teams effectively design a plan from data collection to analysis ready datasets. Examples include using Qualtrics Research Suite to design surveys, gather data, and export tables. SAS 9.4 is used to create a functional dataset for further manipulation, investigation, and analysis.

INTRODUCTION

Research begins with a question. An investigation is launched to gather information and find the answer. This process is data collection. Building a thoughtful survey is the first step to increasing our knowledge on the subject in question. When collecting information, the questions asked must be concise and relevant. The variables collected should have purpose to ensure the outcome can be evaluated with accuracy. Considerations should include details about the population being measured, the collection mechanism, and survey layout. Communication is key! This is an optimal time to close the gap between field researchers, programmers, and investigators. Errors should be anticipated; the data will not be perfect but there are several ways to set your project up for success. Utilize tools such as skip patterns, variable constraints, and SAS edit check programming. The data management techniques presented in this paper are intended to take the user from the beginning of a research study and implement best practices to ensure datasets imported from the field data entry system are accurate, clean, and require little or no manipulation before analysis.

DATA COLLECTION

All data collected will need to be linked to a specific person, case or location. Unique identifiers should be assigned at the earliest possible time to enable accurate tracking of collected data for each case. When creating an instrument consider the population being surveyed. Questions should be written at a sixth-grade reading level so that all possible audiences are given the opportunity to comprehend each word or phrase. Don't use abbreviations or slang words as these can cause the meaning of the question to change based on the study subject's perception. Focus on one topic at a time, include only one question, and use a follow-up question if needed for further clarification.

Finally, keep it short; questions should be concise and to the point. Extra verbiage could confuse or even sway the study subject to answer differently. Planning how response values will be assigned and recorded will help streamline the process from field collection to analysis.

SUBJECT IDENTIFIER

The first step in tracking all pieces of your study data is assigning a unique identifier to each case. A study ID of some type is typically used in research to de-identify the data and make it confidential. But this number can be incredibly useful with some thoughtful planning. A numerical study ID should be assigned to every potential study case as soon as a person or location is added to the project. This study ID can be used to track each piece of data associated with the case including personal information, questionnaire data, environmental samples, and results of sample analysis. Projects may require data to be stored in multiple, separate datasets and formats. The study ID makes it easy to link every segment of data for each case together in SAS to compile all the information. Having several options for collecting and processing raw data at your disposal enables the research team to adjust to the needs of all aspects of different research projects. Access databases, Excel spreadsheets, and online survey tools such as Qualtrics are some data collection options. Researchers often choose to put all the data in one Excel spreadsheet, adding each set of data to one row as it is collected or developed. While this method can work, input of data is challenging, and errors are common. Field questionnaire data may be collected using paper and pen and then later input into an Access database or collected directly as electronic data using Qualtrics. Figure A1 (see Appendix) illustrates how to use a unique study ID across multiple data collection and storage options. Figure A2 incorporates the unique identifier for a family and includes the individual and samples.

QUESTION DEVELOPMENT AND LAYOUT

Good data collection instruments are critical to producing quality research data. Begin with having a clear list of the data you need to collect. Determine any covariates that will be important to the final analysis questions. Once you know what you want to ask, steps should be taken to develop how the questions will be asked and options for collecting responses. This may sound simple, but it can be very tricky to ensure the data you collect will correctly answer your research questions. Bring the entire research team together; each group—principal investigators, data management, and the field team— will have different input which will help make the instrument's design solid. Having a quality design will help with your data analysis and ensure consistent data collection. Questions and their answer choices should be clear, so they are not left open to interpretation.

When creating a research survey, it is important to consider if each question being asked will produce the data you need. Often the original question is too broad in its scope to provide useful information. The question may need to be broken down into a set of questions to get at the answer you really want.

The instrument should be thoroughly tested. Reading through a questionnaire to yourself is very different from administering it to another person. This is an important step and testing should be done several times with different members of the design team. The instrument should also be tested under the same circumstances it will be used in the field, whether that is in person, over the phone, or an online survey completed by the subject. Testing across a wide group of people will help identify any issues with deployment or instrument design and allow time to rectify problems.

At the development phase, communication between the principal investigator (PI), data team (coordinator, manager, programmers, analysts), and field researchers is invaluable. Each have their own goals and perspective. The PI has a question to answer. The data team has the task of drilling down to the core of this issue to guarantee the right questions are asked and the responses provide valid and valuable information. The field researchers have the task of data collection and will consider how the subjects might respond and if the interview can be followed with ease. Working together and communicating early on creates a successful platform to ensure a scientifically sound study.

RESPONSE VALUES

Why create more work? Thoroughly examining variables and their values at the development phase will make your life easier on the SAS end. Always use numbers when coding responses. Online survey software will capture the response information but if responses are not assigned with a numeric value then the software will default to auto numbering. If all responses are not recorded consistently then this may present a problem. Consider if the list of responses is inclusive of all potential options. Needing the addition of a choice partway through the study could either present inconsistencies with numbering at data entry or the online entry mechanism (i.e. Qualtrics) could re-code the values entirely. When discussing instrument design with your team, request one variable for each question and code appropriately with numeric responses. This will reduce error and save time re-coding variables during the SAS programming phase.

For accuracy, it is better not to ask the study subject to do calculations. Instead, get their best estimate, offer ranges where possible and let SAS do any computations. This will help to avoid the "I don't know" responses.

Despite good question design there is a chance the only valid response is "I don't know" or it is left missing. Missing values can happen because the question was overlooked in error, the response values did not encompass an accurate choice, or the question was refused. In research it is important to differentiate between these kinds of non-response. If the study subject responds "I don't know" as their answer this is important data to capture. Options for collecting missing data could include using a numeric value in the choice list as well as having a code to distinguish the type of missing response. For example, 8 = Don't know, 9 = Not Applicable. At the SAS level we could look through our response values using edit check programming (see section on Edit Checks) to help the coders determine if the answer was truly missing or an error occurred that can be resolved.

Where possible you should always avoid open-ended questions that lead to text answers. If you can't use PROC FREQ to list response values easily then your cleaning and/or analysis will require more work. However, there are times when a question may require an "other" to accommodate unforeseen responses. Targeted comment variables can be used to collect information that cannot be put into a category. Valid and consistent responses are the goal for any research project but, when necessary, a 'comment' or 'specify' variable can be attached to ensure all the information is being recorded. For example, when asking about race (Figure 1), we can numerically categorize what the common responses are but not all individuals feel they fit into one of our choices. The "Other" option can contain a line that will be input at data entry. In SAS we can easily look at "variable_x" responses and accurately re-code them if applicable or even create a new group. Having the text entered as data makes it easy to look through the responses in a SAS dataset.

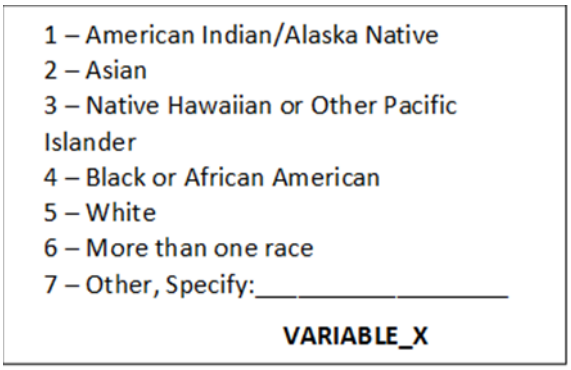


Figure 1. Example Response Choices for Race

DATA ENTRY MECHANISMS

While there are many options for collecting data directly into an electronic format, for most of the research projects done at our Center we have found nothing beats paper and pen to enable the data collector to adequately record all the information provided by the study subject. Once the data are collected, reviewed, and verified for correct coding, it is then entered in an electronic format. Typically, this involves double-entering the data into a form in Access. We have had great success with this method for many large research projects. However, we also find alternatives to collecting data are required in certain circumstances. Qualtrics online survey software is ideal for creating an instrument that can be completed by the study subject or administered by a member of the research team in the field.

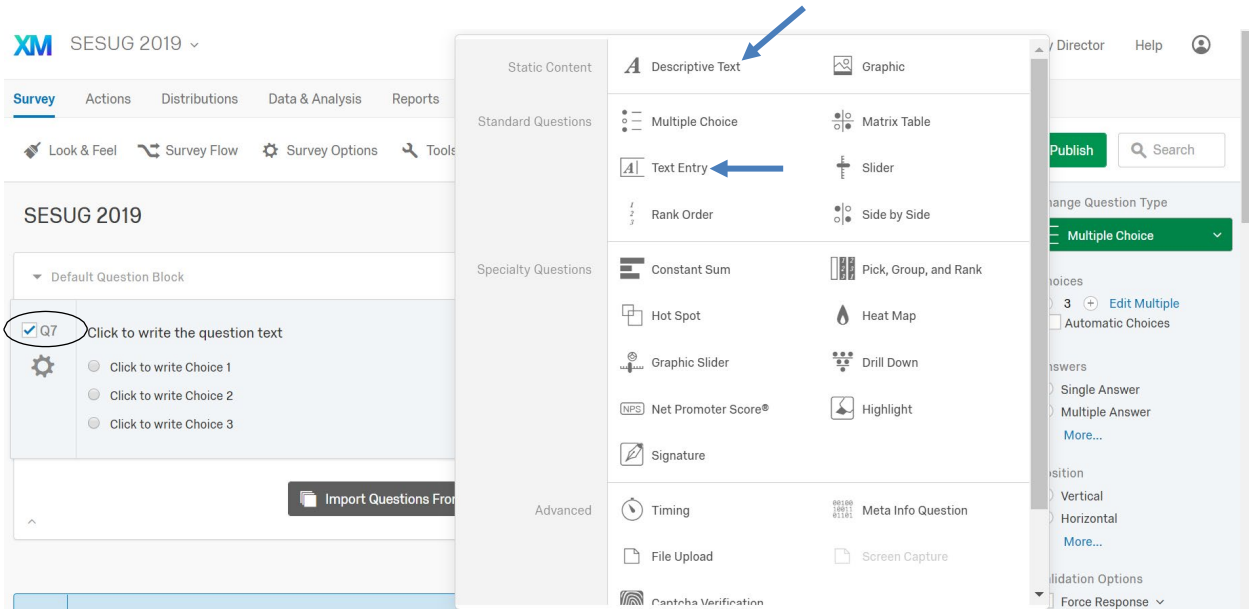
Being ready for the unforeseen is critical. We provide our field staff with paper versions of all electronic data collection tools in case technical difficulties arise. Data can be entered into Qualtrics when the staff member returns to the study office.

BUILDING A SURVEY IN QUALTRICS

Qualtrics is simple to use and it is easy to customize a survey to your needs whether small or large. With Qualtrics you can format questions and collect responses in multiple ways. Then, as a first step to reducing errors and making your data analysis ready, you can easily set up skip patterns and place constraints on the responses allowed for a question.

GETTING STARTED

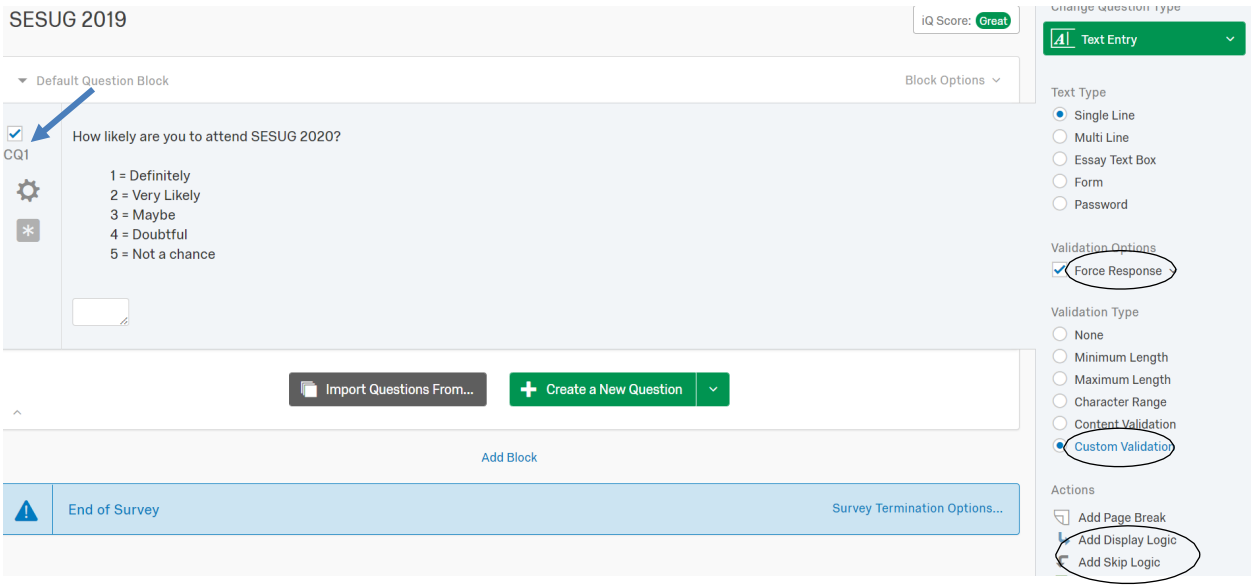
Start by opening a Blank Survey Project. It is easy to begin adding text and questions to the survey. The option "Descriptive Text" can be used to add introductions, directions or transitional statements. Be aware that Qualtrics will automatically name your variables in the order they are created. For example, in Display 1, several questions were deleted prior to this version and now the first question has a variable name of Q7. By clicking on the variable name, you can change it to suit your needs. Display 2 shows the updated variable name for the question.



Display 1. Qualtrics Research Suite – Basics

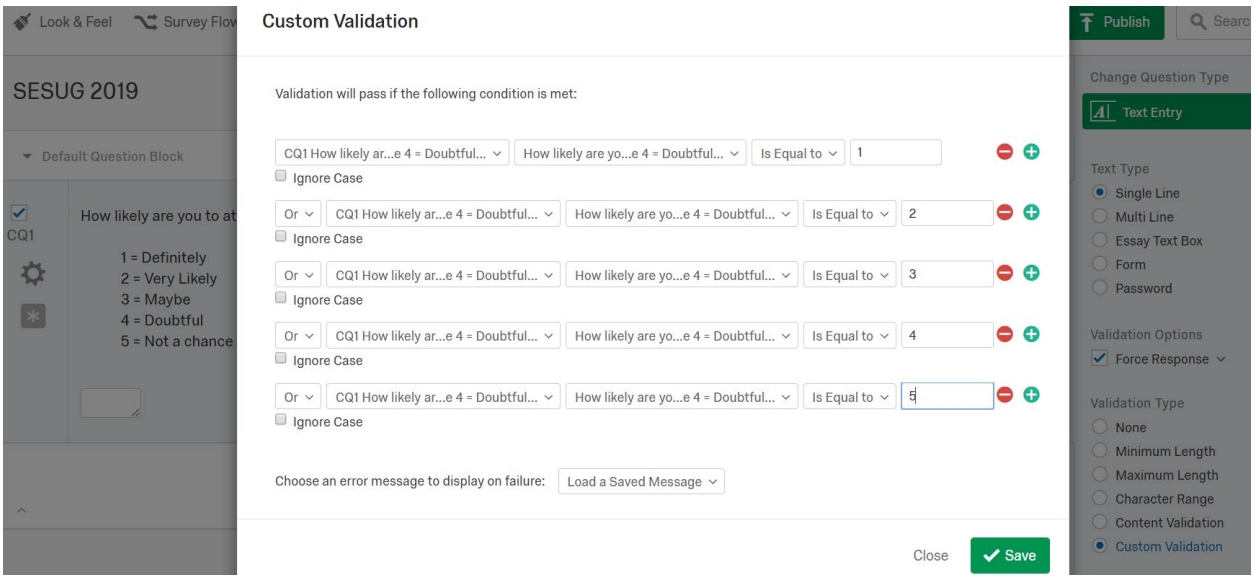
ERROR REDUCTION METHODS

Reduce the number of errors during data collection by making use of the design options in Qualtrics. The first step to creating the survey is choosing the question type (see Display 1). Use the question type "Text Entry" to collect data. This format provides the best variable control options for data entry. If the standard "Multiple Choice" question format is used, then Qualtrics will assign the variable values. If additional options or changes are required later, you will not have control over the new variable values. This can lead to inconsistencies in the coding. Display 2 shows a question written using the "Text Entry" option where data will be input to the box provided. It also highlights how to add several constraints to variable responses. Choose "Force Response" and the question must be answered to move on with the survey.



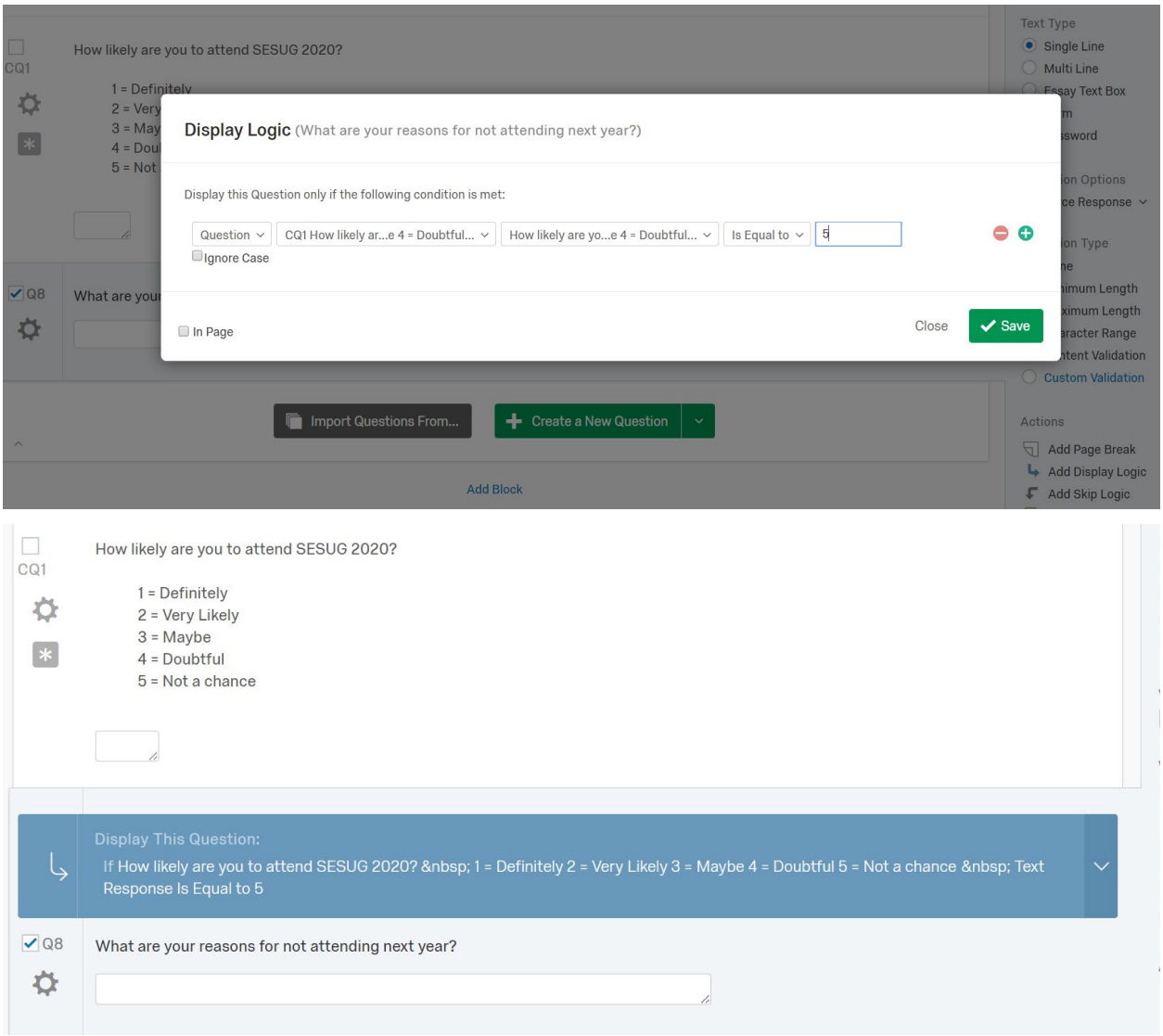
Display 2. Qualtrics Research Suite – Reducing Errors

In Display 3, use "Custom Validation" to limit the types of data that can be input or "Content Validation" to format the input of the data into a date, phone number or other option.



Display 3. Qualtrics Research Suite – Custom Validation

Skip patterns and display logic can be used to collect only the data required for a particular situation. Errors will be reduced when inapplicable questions do not display. Display 4 shows the before and after screen views when setting up a display logic requirement for a question. In this example, Q8 will only show on the screen if the response to CQ1 is equal to 5. A skip pattern may skip several questions that are only applicable when a specific response is given to a question. Using display and skip logic also simplifies the instrument completion for the research staff and study subject.



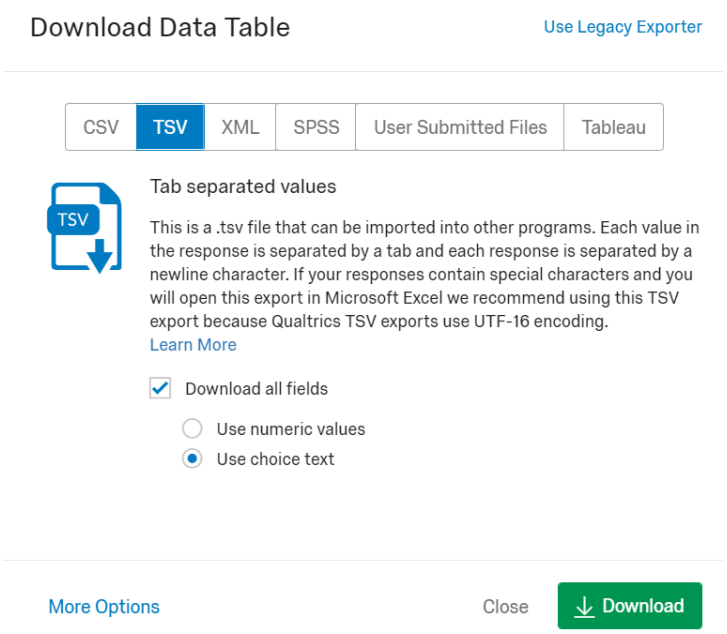
Display 4. Qualtrics Research Suite – Display Logic Steps 1 & 2

MANAGING YOUR DATA FROM QUALTRICS TO SAS

OUTPUT DATASETS

Qualtrics offers multiple file extensions for the downloading of data collected online (Display 5). Moving raw data out of Qualtrics to a tab-separated value file extension creates a simple transfer of data. Another option for downloading survey data is to use SAS to query the API and generate an XML map. In this example, the .tsv file is saved locally and we use SAS to import the survey data:

```
PROC IMPORT OUT= WORK.Example
   DATAFILE= "C:\QDATA\Int_111318.tsv" DBMS=TAB REPLACE;
   GETNAMES=YES;
        DATAROW=2
        ; RUN;
```



Display 5. Qualtrics Research Suite – Download Data Table

SAS INPUT PROGRAM

The SAS import syntax is an opportune place to expand your SAS code to create a full input program where you can format variables, change variable types, account for missing information, assign labels, and create new variables required for analysis. The finished product can be completed at this step.

Assigning variable labels becomes very helpful when sharing data. The recipient will have the option of viewing the dataset by column name or column label, in turn, possibly reducing the number of follow questions if the labels are descriptive. For example, if variables are not labeled the name shown is the variable name, i.e. HQ5. If a label is attached, in one click on the taskbar menu to 'view column labels', the variable HQ5 could have a description, "Do you have pets".

In this import all variables are returned as character. Choose the group of variables to convert, use a SAS array to run through the list and the INPUT function to change the variable type. To change from character to a formatted date create a new variable, use the INPUT function and format, i.e. `Dobdate = input(HQ34, mmddyy10.).`

Keep only the necessary variables, dropping those that will not be used. Rename variables if necessary, for clarity, but also to match other datasets. If you know the key variables that will need to be linked save yourself programming later and include the change here (i.e. change 'study id' to 'studynum' if that is the variable name in other databases and/or files).

Output the cleaned-up file to storage. It is always good practice to have clear and useful file name. Use a key word and the date. When the file is updated keep the key word and change the date each time.

EDIT CHECKS

Creating SAS error check programs are an efficient way to automate the process of scanning data for errors. These can be written as soon as the instruments are field ready. Programs are written for each separate instrument and can be adjusted to accommodate any changes as a study progresses. These types of programs can be as complicated or as simplistic as needed. Example Code A looks at one variable, Race. The overall method can be applied to an entire interview.

Example Code A

The following code returns errors found for the variable Race (Output 1). First the macro "Invalid" is created to customize the output for any errors found. Study ID, the variable name and the variable value are printed. There are seven valid responses for the variable Race1. An IF/THEN statement locates any values not within range. If a question is asked several times with the same ranges—for instance, here we ask the race of up to six people—an array can be included to run through each variable. Below, Output 1 shows an example of the output if an error is found within the data:

```
%MACRO INVALID(VARNAME,VALUE); PUT
    @8 Study_id
    @20 "INVALID " &VARNAME
    @65 &VARNAME " = " &VALUE " _________________" /;
    %MEND INVALID;
 QUIT;


IF NOT (1<= Race1 <=7) THEN DO; %INVALID(' Race1 ', Race1); END;


array R6 (6) Race1 Race2 Race3 Race4 Race5 Race6; IF
COUNT = 6 THEN DO;
do i = 1 to 6;
IF R6[I] NOT in (1,2,3,4,5,6,7) THEN DO; %INVALID('R6[I]',R6[I]); END;
END;
END;
END;
```

```
        8972        INVALID RACE1                                   Q3_C1 = 0
```

Output 1. The output from error macro %INVALID

Example Code B checks the variable ranges as well as if skip patterns are being applied accurately. The question (HQ32) asked if the study subject had an experience. If yes (1), then the following sub-questions are asked, and a response is required for each one. If any of the sub-questions are missing, an error is printed. Looking at a question as a whole makes it easier for the coders to locate exactly where the error is occurring. The macro here is written to print the three variables associated with the question.

Example Code B

The following code returns errors found for a question with an option for a skip pattern. The initial response must be a '1' for the subsequent response to need values. If the follow-up questions are missing an error is printed. The macro "Conflict3" prints several variables required for investigation:

```
%MACRO CONFLICT3
   (VARNAME_1,VALUE_1,VARNAME_2,VALUE_2,VARNAME_3,VALUE_3); PUT
   @8 STUDY_ID @13 "CONFLICTING DATA FOR --> " &VARNAME_1
   @63 STUDY_ID &VARNAME_1 " = " &VALUE_1 @85 " ____________________" /
   @63 STUDY_ID &VARNAME_2 " = " &VALUE_2 @85 " ____________________" /
   @63 STUDY_ID &VARNAME_3 " = " &VALUE_3 @85 " ____________________" /;
%MEND CONFLICT3;


IF HQ32 NOT IN (0,1) THEN DO; %INVALID ('HQ32', HQ32); END; IF HQ32 = 1
 THEN DO;
 IF HQ32A = '' THEN DO;
      %CONFLICT3('HQ32',HQ32,'HQ32A',HQ32A,'HQ32A_1',HQ32A_1);END;
```

CONCLUSION

Good planning is the first step to having a clean SAS dataset which is ready for analysis in the shortest time frame. Having members of the field team and the data management team participate in the planning will help streamline the data flow from the field. Considering question layout, response options, data collection methods, and data processing are all critical. Implementing SAS edit check programming early on will ensure data is being entered correctly and consistently. The SAS programmer can have vital input when developing data collection instruments. Attention to all these details will expedite getting that final SAS dataset to the investigators.

ACKNOWLEDGMENTS

We would like to acknowledge our colleagues and friends at the Center for Perinatal, Pediatric & Environmental Epidemiology.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Julie Plano
Yale School of Public Health
Julie.colburn@yale.edu

Keli Sorrentino
Yale School of Public Health
Keli.sorrentino@yale.edu

APPENDIX

Access DB #1	Excel SS #1	Access DB #2	Access DB #3	Access DB #4	Access DB #4	Excel SS #2	Excel SS #3
Participant Tracking	GPS Data	Instrument Data	Field Biological Sample Collection	Sample Receipt at Lab	Aliquot Processing & Storage	Lab #1 Results	Lab #2 Results
Stud_Num	Stud_Num	Stud_Num	Sample_ID	Sample_ID	Sample_ID	Sample_ID	Sample_ID
50421	50421	50421	50421-101	50421-101	50421-202 50421-203 50421-305	50421-202 50421-203	50421-305

Figure A1 shows how a unique study number (Study_Num) is used to identify data collected from one study subject. The same number is used across all tables, forms and databases. The unique sample identifier is followed by the study number and can be found in the last 3 digits of the variable (Sample_ID)

Access DB #1	Excel SS #1	Access DB #2	Access DB #3	Access DB #4	Access DB #4	Excel SS #2	Excel SS #3
Participant Tracking	GPS Data	Instrument Data	Field Biological Sample Collection	Sample Receipt at Lab	Aliquot Processing & Storage	Lab #1 Results	Lab #2 Results
Family_ID	Family_ID	Individual_ID	Sample_ID	Sample_ID	Sample_ID	Sample_ID	Sample_ID
50421 Individual_ID 50421-01	50421	50421-01	50421-01-101	50421-01-101	50421-01-202 50421-01-203 50421-01-305	50421-01-202 50421-01-203	50421-01-305
Next Individual							
Access DB #1	Excel SS #1	Access DB #2	Access DB #3	Access DB #4	Access DB #4	Excel SS #2	Excel SS #3
Participant Tracking	GPS Data	Instrument Data	Field Biological Sample Collection	Sample Receipt at Lab	Aliquot Processing & Storage	Lab #1 Results	Lab #2 Results
Family_ID	Family_ID	Individual_ID	Sample_ID	Sample_ID	Sample_ID	Sample_ID	Sample_ID
50421 Individual_ID 50421-02	50421	50421-02	50421-02-101	50421-02-101	50421-02-202 50421-02-203 50421-02-305	50421-02-202 50421-02-203	50421-02-305

Figure A2 includes a unique identifier by family (Family_ID). Unique individuals within a family are coded with two digits following the family identifier (Individual_ID). The sample identifier is followed by the individual id (Sample_ID)