PharmaSUG 2017 - Paper HA08

Put the “K” in Your KAB: Know How to Efficiently Program with Knowledge, Attitudes, and Behavior Survey Data

Cara Lacson, United BioSource Corporation
Jasmeen Hirachan, United BioSource Corporation

ABSTRACT

Data from Knowledge, Attitudes, and Behavior surveys can present challenges, but these can be alleviated by a solid, efficient programming strategy. Producing KAB tables, listings, and figures can be very tedious to program, from the number of SAS® data set variables that the survey data requires, to skip logic within the survey, and the volume and detail of question and response text that must be reported. With lots of typing involved, this is certainly something you want to do only once within your project, and in one location, to reduce the chance for error. This paper presents the basic concepts of a KAB survey from a programming perspective: what to expect in your data, recommendations for your analysis data set structure, and how to handle permitted skip questions. This paper also explores a solution to storing long question text, responses, and correct answers in one central location for use throughout all of your SAS programs in a KAB study.

INTRODUCTION

A Knowledge, Attitudes, and Behavior (KAB) survey is used to obtain information about a group of individuals with common attributes. The KAB survey or questionnaire is designed to accurately measure trends in knowledge, attitudes, and behaviors associated with a select product. As a result, it is a powerful tool in evaluating risks associated with a particular product and is commonly used as a part of a Risk Evaluation and Mitigation Strategy (REMS).

Each type of survey targets a different stakeholder, for example, patients, pharmacists, or healthcare providers. Some research programs target multiple stakeholders, with a unique survey for each participant type, and some may only target one type of participant, depending on the requirements in the REMS. The respondents within a particular stakeholder have a common attribute, such as patients who take a certain prescription medication to alleviate a health condition, pharmacists who distribute the prescription medication to individuals, or healthcare providers who prescribe the medication to their patients.

The KAB survey helps industry leaders gain valuable information from the self-reported level of knowledge, attitudes, and behaviors of the targeted group of individuals with questions that help determine the knowledge with which the respondent is using, distributing, or prescribing a medication. It is critical that all individuals involved are well informed and understand the risks associated with the product, in order to maintain good practices to enhance safety and efficacy.

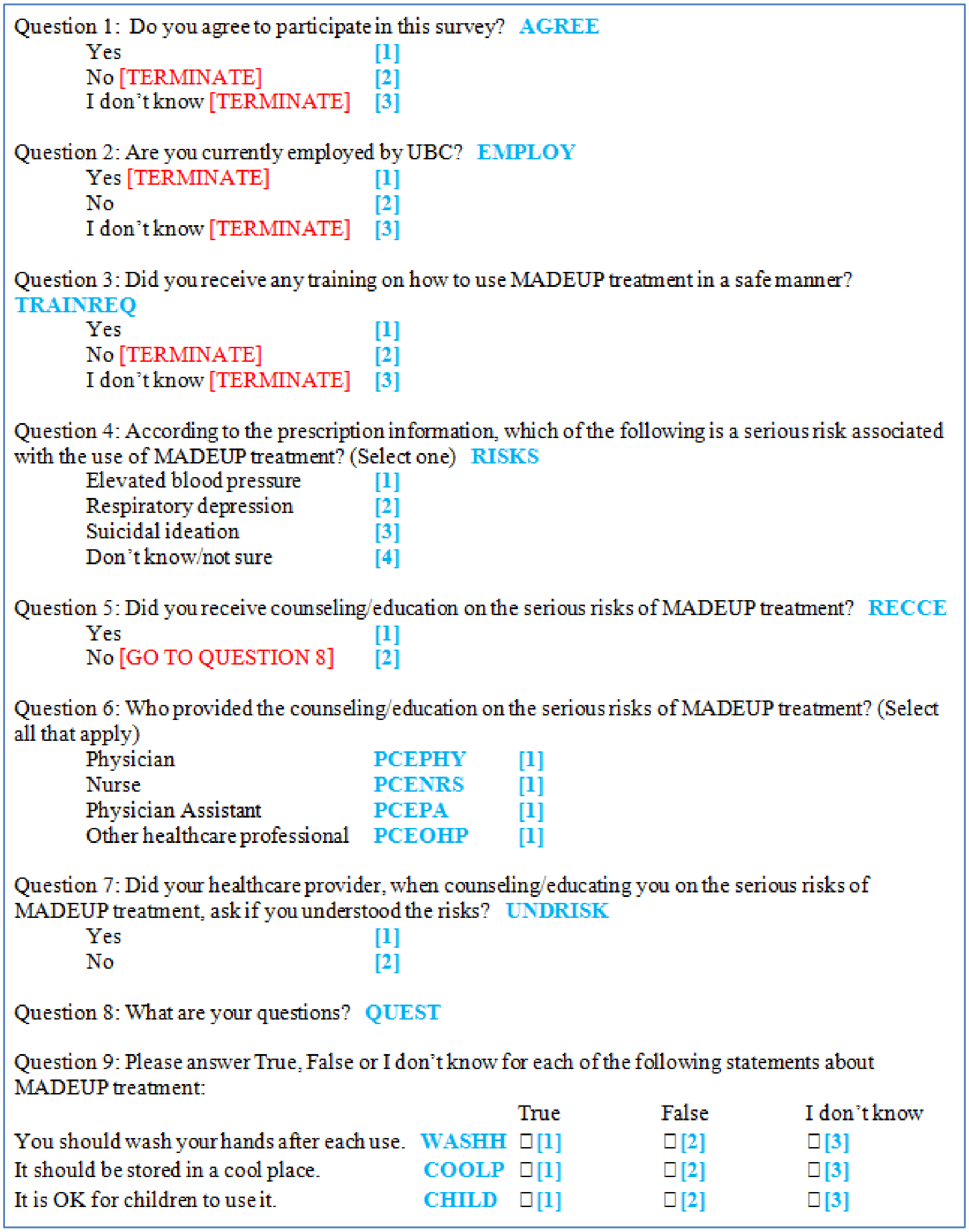
Because surveys can be quite lengthy with detailed text, resulting in SAS data sets with lots of variables, and summary tables displaying exact question and response text, it is important to have a plan to manage this type of survey data effectively and efficiently.

WHAT TO EXPECT IN YOUR KAB SAS DATA

Since this data is generally not submitted to regulatory authorities, CDISC standards rarely (if ever) apply. For this presentation, the original SAS data set that stores the survey responses is in a format with one record per respondent, with as many variables as needed to capture each of the survey responses.

GENERAL MAKEUP OF A KAB SURVEY

A KAB survey typically contains a section with eligibility questions, followed by questions that further explore the level of knowledge, attitudes, and behavior that the respondent has about a particular product. Display 1 below introduces the sample Patient KAB survey that is used in this paper, with annotation of variable names and values in **blue**, and instructions for termination and skips in **red**.



Question 1: Do you agree to participate in this survey? AGREE

Yes [1]
No [TERMINATE] [2]
I don't know [TERMINATE] [3]

Question 2: Are you currently employed by UBC? EMPLOY

Yes [TERMINATE] [1]
No [2]
I don't know [TERMINATE] [3]

Question 3: Did you receive any training on how to use MADEUP treatment in a safe manner? TRAINREQ

Yes [1]
No [TERMINATE] [2]
I don't know [TERMINATE] [3]

Question 4: According to the prescription information, which of the following is a serious risk associated with the use of MADEUP treatment? (Select one) RISKS

Elevated blood pressure [1]
Respiratory depression [2]
Suicidal ideation [3]
Don't know/not sure [4]

Question 5: Did you receive counseling/education on the serious risks of MADEUP treatment? RECCE

Yes [1]
No [GO TO QUESTION 8] [2]

Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)

Physician	PCEPHY	[1]
Nurse	PCENRS	[1]
Physician Assistant	PCEPA	[1]
Other healthcare professional	PCEOHP	[1]

Question 7: Did your healthcare provider, when counseling/educating you on the serious risks of MADEUP treatment, ask if you understood the risks? UNDRISK

Yes [1]
No [2]

Question 8: What are your questions? QUEST

Question 9: Please answer True, False or I don't know for each of the following statements about MADEUP treatment:

		True	False	I don't know
You should wash your hands after each use.	WASHH	□ [1]	□ [2]	□ [3]
It should be stored in a cool place.	COOLP	□ [1]	□ [2]	□ [3]
It is OK for children to use it.	CHILD	□ [1]	□ [2]	□ [3]

Display 1: Sample KAB Survey with Annotation

Questions 1, 2, and 3 represent “Eligibility” or “Inclusion/Exclusion” questions, which determine if the respondent is eligible to participate in the remainder of the survey based on their answers. Some responses in the Eligibility section are noted as “TERMINATE” in red. If a TERMINATE response is selected, the respondent is not allowed to continue participation in the survey, and the survey is discontinued. Questions 4 through 9 explore the respondent’s understanding of the serious risks associated with the medication, and also the level of information that was provided to the respondent by a healthcare professional.

TYPES OF QUESTIONS AND VARIABLES

With regard to data variable structure, there are 3 types of questions:

1. A “single response” question is an objective multiple choice question where the respondent can choose only one response. It is usually represented by a numeric variable with values of 1 through x, where 1 represents the first response text, and x is the maximum number of responses for the question.
2. A “multiple response” question is an objective question where the respondent is directed to “select all that apply” and can choose more than one response. The question is represented by multiple variables; it has a unique numeric variable for each possible response, usually with values of missing or 1, where 1 represents the response having been selected.
3. A “verbatim text” question is a subjective question in which the respondent can answer with a single free text response. It is represented by a unique character variable that captures the respondent’s answer exactly as it was provided.

The previous example (Display 1) shows single response Questions 1, 2, 3, 4, 5, 7, and 9, multiple response Question 6, and verbatim text Question 8. The single response (multiple choice) questions are represented by a single numeric variable per question: AGREE for Q1, EMPLOY for Q2, TRAINREQ for Q3, RISKS for Q4, RECCE for Q5, UNDRISK for Q7, and WASHH, COOLP, and CHILD for each of the individual sub-questions within Q9. The multiple response (select all that apply) question contains multiple numeric variables, PCEPHY, PCENRS, PCEPA, and PCEOHP, each of which will be set to a value of 1 if selected. The verbatim text question is represented by a single character variable QUEST.

SKIPPED QUESTIONS

Some questions may be allowed to be “skipped” based on responses to previous questions, in which case there may be some missing data in the original SAS data set. In the previous example (Display 1), if the respondent answers “No” to the single response Question 5, then the subsequent multiple response Question 6 and single response Question 7 are not asked. The result is that variables PCEPHY, PCENRS, PCEPA, and PCEOHP from Question 6, and UNDRISK from Question 7 will all be missing (or filled with a value that represents missing) in the original SAS data set.

SURVEY TEXT

Survey question and response text usually needs to be reported in the analysis tables exactly as it is presented in the survey, including all punctuation, spacing, special characters, font styles (bold, italic, underline), etc. It can be quite lengthy and tedious to type, not to mention prone to errors, especially if typed more than once and/or in multiple files. This text is not normally included in the original SAS data set. A solution for efficiently storing and accessing question and response text is explained in subsequent sections of this paper.

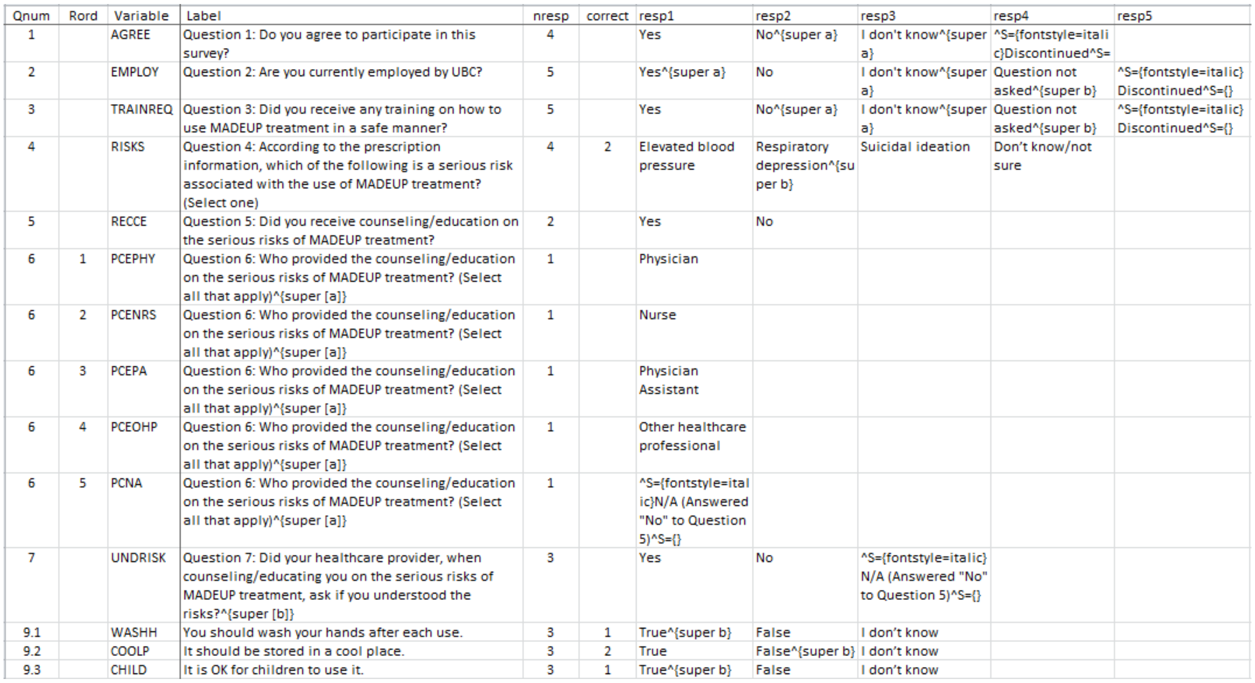
CORRECT RESPONSES

Some questions may have a “correct response” which is the desired response. This is because it is important to know if respondents answered correctly to questions pertaining to key objectives of the survey. Questions that are part of one of these key objectives and/or that relate to the safe use of the product usually have a correct response. The correct responses are not typically flagged in the original SAS data set. Examples of how to identify and flag correct responses for the analysis tables are discussed in subsequent sections of this paper.

“QUESTIONS” SPREADSHEET

The KAB tables need to display survey question and response text, which is generally lengthy. A Microsoft Excel spreadsheet named “Questions” is created to avoid repetitively typing survey text in multiple programs. Along with the survey text, any other relevant values associated with the questions that need to be referred to repetitively in the table programs are also saved in the Questions spreadsheet.

Display 2 below shows a sample Questions spreadsheet corresponding to the survey questions shown in the previous Display 1.



Qnum	Rord	Variable	Label	nresp	correct	resp1	resp2	resp3	resp4	resp5
1		AGREE	Question 1: Do you agree to participate in this survey?	4		Yes	No^{super a}	I don't know^{super a}	^S={fontstyle=italic}Discontinued^S=	
2		EMPLOY	Question 2: Are you currently employed by UBC?	5		Yes^{super a}	No	I don't know^{super a}	Question not asked^{super b}	^S={fontstyle=italic} Discontinued^S={}
3		TRAINREQ	Question 3: Did you receive any training on how to use MADEUP treatment in a safe manner?	5		Yes	No^{super a}	I don't know^{super a}	Question not asked^{super b}	^S={fontstyle=italic} Discontinued^S={}
4		RISKS	Question 4: According to the prescription information, which of the following is a serious risk associated with the use of MADEUP treatment? (Select one)	4	2	Elevated blood pressure	Respiratory depression^{super b}	Suicidal ideation	Don’t know/not sure	
5		RECCE	Question 5: Did you receive counseling/education on the serious risks of MADEUP treatment?	2		Yes	No			
6	1	PCEPHY	Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)^{super [a]}	1		Physician				
6	2	PCENRS	Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)^{super [a]}	1		Nurse				
6	3	PCEPA	Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)^{super [a]}	1		Physician Assistant				
6	4	PCEOHP	Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)^{super [a]}	1		Other healthcare professional				
6	5	PCNA	Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)^{super [a]}	1		^S={fontstyle=italic}N/A (Answered "No" to Question 5)^S={}				
7		UNDRISK	Question 7: Did your healthcare provider, when counseling/educating you on the serious risks of MADEUP treatment, ask if you understood the risks?^{super [b]}	3		Yes	No	^S={fontstyle=italic} N/A (Answered "No" to Question 5)^S={}		
9.1		WASHH	You should wash your hands after each use.	3	1	True^{super b}	False	I don’t know		
9.2		COOLP	It should be stored in a cool place.	3	2	True	False^{super b}	I don’t know		
9.3		CHILD	It is OK for children to use it.	3	1	True^{super b}	False	I don’t know		

Display 2: Questions Spreadsheet

Each row in this spreadsheet includes a question and its responses. The columns in the spreadsheet represent question number, response order, variable name, question text, number of responses, number representing the desired correct response (if applicable), and response text in order of appearance in the survey. Only questions that have a numeric value in the data set will be included in the spreadsheet. Questions that have free text entered by the participant are not included, e.g. Question 8, “What are your questions?” will have text entered by the respondent which will not correspond to a numeric value in the data set.

The QNUM column corresponds to the survey question number. Generally, QNUM should match the survey question number. When a sub-question has its own series of responses, QNUM should also have a similar structure. For example, each sub-question in Question 9 in Display 1 has a set of “True, False and I don’t know” responses. Each sub-question is assigned its own unique Question number, i.e. 9.1, 9.2, and 9.3, as shown in Display 2. The variable RORD is the Response ORDer for questions with multiple responses, such as Question 6 in Display 1. For these questions, each possible response will be treated as a variable, which if selected will have a value of 1 in the data set. Each row in the spreadsheet will then contain one response, i.e. only the RESP1 column is populated with the response as shown in Display 2. RORD controls the corresponding order of appearance in the tables.

The VARIABLE column contains the variable name for the survey question from the annotated Data Collection Tool (DCT). This helps connect each variable in the analysis data set to associate with each row in the Questions spreadsheet. LABEL is the actual question text as it should appear in the table output including any Rich Text Format (RTF) codes that are required for font styles (e.g. underline, italics) and text (e.g. superscripts) that appear in the Statistical Analysis Plan (SAP). NRESP is the Number of possible RESPonses for each question which includes any derived responses, such as “Discontinued”, “Question not asked”, etc. For questions that have desired correct answers which may require additional statistical calculations, the CORRECT column represents the desired correct response number.

Each RESPx (x being the number 1, 2, 3 and so on) column includes the response text for each question. Any RTF codes for font styles or additional text required for the table outputs are also typed in here. The responses in the RESPx columns follow the order of appearance in the survey. For Question 2, “Yes”, “No”, and “I don’t know” will be in the RESP1, RESP2, and RESP3 columns, respectively, as per the survey questions in Display 1. This corresponds with the values for the variable EMPLOY in the data set, where 1 = “Yes”, 2 = “No”, and 3 = “I don’t know”. The order of the derived values 4 = “Question not asked” and 5 = “Discontinued” are determined by the SAP specifications and are entered in the RESP4 and RESP5 columns, respectively.

IMPORTING THE QUESTIONS SPREADSHEET INTO A SAS DATA SET

This Questions spreadsheet is saved in a separate “Imports” folder along with a SAS program for importing the spreadsheet to a SAS data set. The spreadsheet is imported into a permanent SAS data set named “QUESTIONS” and saved in the same directory as the original survey data so that it is easily accessible; any program can refer to it, and it will also be used to produce the analysis data set ADTQ discussed later in this paper. It should contain the same number of rows and columns as the original spreadsheet.

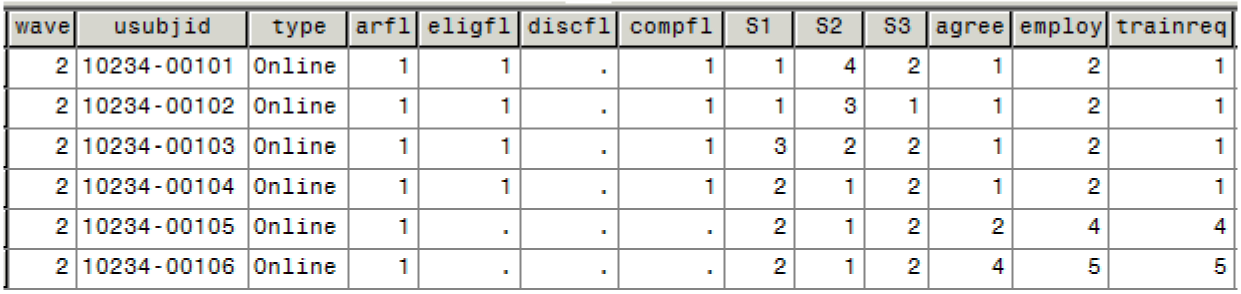
ADPQ (ANALYSIS DATA SET – POPULATION QUESTIONNAIRE)

ADPQ.sas7bdat is a subject-level analysis SAS data set in which analysis variables are derived (such as population flags and subgroups) and, most importantly any question that is skipped by the respondent is assigned an appropriate non-missing value. A variable’s numeric value in the original data set corresponds to the order of appearance of the survey responses which match the order in the QUESTIONS data set. A variable for which a respondent did not pick a response will be blank and a value will be assigned to it in ADPQ corresponding to the QUESTIONS data set.

HANDLING ELIGIBILITY QUESTIONS

If a respondent picks a termination response, the survey is terminated for the respondent before completion. In such cases, the rest of the Eligibility questions will be assigned a value of “Question not asked” in the survey response. If, at any time during the survey, a respondent drops out of the survey, then the remaining Eligibility questions are assigned a value of “Discontinued”.

In the following example in Display 3, the variable AGREE has a value of 1 for “Yes” or 2 for “No” per the survey question. If a respondent logged into the survey but did not answer any question, then all Eligibility questions will get a value corresponding to “Discontinued”. As shown in the Questions spreadsheet in Display 2, AGREE will get a value of 4, EMPLOY will get a value of 5, and TRAINREQ will also get a value of 5 for a “Discontinued” respondent. If a respondent answered “No” to the first question, then the survey will be terminated as the respondent becomes ineligible, in which case the variables EMPLOY and TRAINREQ will get values of 4 corresponding with RESP4 “Question not asked”. Display 3 below shows these variables in the ADPQ data set.



wave	usubjid	type	arfl	eligfl	discfl	compfl	S1	S2	S3	agree	employ	trainreq
2	10234-00101	Online	1	1	.	1	1	4	2	1	2	1
2	10234-00102	Online	1	1	.	1	1	3	1	1	2	1
2	10234-00103	Online	1	1	.	1	3	2	2	1	2	1
2	10234-00104	Online	1	1	.	1	2	1	2	1	2	1
2	10234-00105	Online	1	.	.	.	2	1	2	2	4	4
2	10234-00106	Online	1	.	.	.	2	1	2	4	5	5

Display 3: ADPQ Data Set, Eligibility Questions

Table 1 below shows how this data is presented in the output table. In this example, there are 243 respondents, out of which 4 did not answer Question 1. These 4 respondents are marked as “Discontinued” and have a value of 3 for the variable AGREE and 4 for the variable EMPLOY. One more respondent did not answer Question 2 and has a value of 4 for EMPLOY. Another 4 respondents answered “No” to Question 1 which is a termination response and hence are determined ineligible to take the survey and have a value of 3 for the variable EMPLOY.

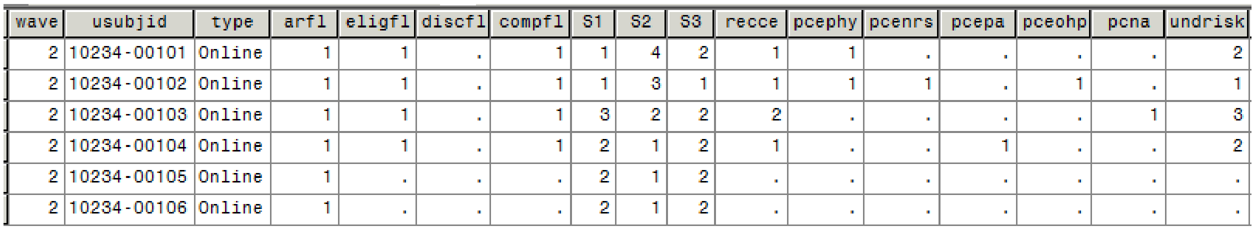
Question	**All Respondents (N=243) n (%)**
Question 1: Do you agree to participate in this survey?	
Yes	235 (96.7)
No[a]	4 (1.6)
I don’t know[a]	0
Discontinued	4 (1.6)
Question 2: Are you currently employed by UBC?	
Yes[a]	0
No	234 (96.3)
I don’t know[a]	0
Question not asked[b]	4 (1.6)
Discontinued	5 (2.1)

[a] Ineligible to participate in the survey.
[b] Question not asked because a previous question had disqualified respondent.

Table 1: Survey Participant Eligibility Results – All Respondents

HANDLING SKIPPED QUESTIONS (Q6 AND Q7)

Question 6 is a multiple response question, and a separate row for variable PCNA is added for “N/A” (Not Applicable) in the Questions spreadsheet. Question 7 will get a “N/A” under RESP3. Please refer to the previous Display 2 to see how these skipped questions are displayed in the Questions spreadsheet. In the ADPQ data set, a respondent answering “No” to Question 5, and hence having to skip Questions 6 and 7, will have value of 1 for PCNA and 3 for UNDRISK. Display 4 below shows these variables in the ADPQ data set.



wave	usubjid	type	arfl	eligfl	discfl	compfl	S1	S2	S3	recce	pcephy	pcenrs	pcepa	pceohp	pcna	undrisk
2	10234-00101	Online	1	1	.	1	1	4	2	1	1	.	.	.	.	2
2	10234-00102	Online	1	1	.	1	1	3	1	1	1	1	.	1	.	1
2	10234-00103	Online	1	1	.	1	3	2	2	2	.	.	.	.	1	3
2	10234-00104	Online	1	1	.	1	2	1	2	1	.	.	1	.	.	2
2	10234-00105	Online	1	.	.	.	2	1	2	.	.	.	.	.	.	.
2	10234-00106	Online	1	.	.	.	2	1	2	.	.	.	.	.	.	.

Display 4: ADPQ Data Set, Single and Multiple Response Questions

Table 2 below shows the 11 respondents who answered “No” to Question 5 skip Questions 6 and 7. For these respondents PCNA is given a value of 1 and UNDRISK is given a value of 3 in the dataset. From the Questions spreadsheet, it is clear that they are reported in the “N/A” rows in the table.

Question	**Respondents (N=238) n (%)**
Question 5: Did you receive counseling/education on the serious risks associated with MADEUP treatment?	
Yes	227 (95.4)
No	11 (4.6)

Question	Respondents (N=238) n (%)
Question 6: Who provided the counseling/education on the serious risks of MADEUP treatment? (Select all that apply)[ab]	
Physician	123 (54.2)
Nurse	26 (11.5)
Physician Assistant	64 (28.2)
Other healthcare professional	70 (30.8)
N/A (Answered "No" to Question 5)	11
Question 7: Did your healthcare provider, when counseling/educating you on the serious risks of MADEUP treatment, ask if you understood the risks?[b]	
Yes	215 (94.7)
No	12 (5.3)
N/A (Answered "No" to Question 5)	11

[a] Percentages may not add up to 100% because more than one response can be chosen.

[b] Percentages are calculated based on the sample presented with the question.

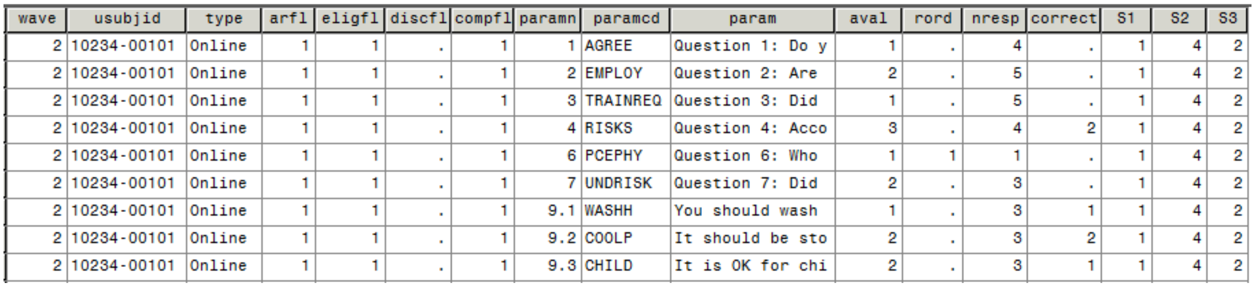
Table 2: Responses to Questions about Counseling/Education

ADTQ (ANALYSIS DATA SET – TRANSPOSED QUESTIONNAIRE)

ADTQ.sas7bdat is an analysis SAS data set with one record for each response per respondent. The table output generally requires counts for each response to a question. For ease of calculation, the question numbers are included in the ADTQ data set. The ADPQ data set is transposed and merged with the QUESTIONS data set. ADTQ contains the question numbers, question text, sort order variable, possible number of responses, and correct response code. The verbatim text responses are excluded from ADTQ. The following code joins the appropriate variables from ADPQ and QUESTIONS and transposes the result to have one record per respondent per response:

```
proc sql ;
   create table adtq as
      select ttq.wave, ttq.usubjid,  ttq.type, ttq.arfl, ttq.eligfl,
             ttq.discfl, ttq.compfl, ttq.varname,
             ttq.col1 as aval,
             q.qnum as paramn, q.rord, q.varlabel as param, q.nresp,
             q.correct,
             ttq.s1, ttq.s2, ttq.s3,
         from ttq left join questions as q
      on ttq.varname eq q.varname
      order by wave, usubjid, qnum, rord ;
quit ;
```

Display 5 below shows the ADTQ data set, with a record for each respondent and each PARAMCD which was transposed from all of the single and multiple response question variables in the ADPQ data set.



wave	usubjid	type	arfl	eligfl	discfl	compfl	paramn	paramcd	param	aval	rord	nresp	correct	S1	S2	S3
2	10234-00101	Online	1	1	.	1	1	AGREE	Question 1: Do y	1	.	4	.	1	4	2
2	10234-00101	Online	1	1	.	1	2	EMPLOY	Question 2: Are	2	.	5	.	1	4	2
2	10234-00101	Online	1	1	.	1	3	TRAINREQ	Question 3: Did	1	.	5	.	1	4	2
2	10234-00101	Online	1	1	.	1	4	RISKS	Question 4: Acco	3	.	4	2	1	4	2
2	10234-00101	Online	1	1	.	1	6	PCEPHY	Question 6: Who	1	1	1	.	1	4	2
2	10234-00101	Online	1	1	.	1	7	UNDRISK	Question 7: Did	2	.	3	.	1	4	2
2	10234-00101	Online	1	1	.	1	9.1	WASHH	You should wash	1	.	3	1	1	4	2
2	10234-00101	Online	1	1	.	1	9.2	COOLP	It should be sto	2	.	3	2	1	4	2
2	10234-00101	Online	1	1	.	1	9.3	CHILD	It is OK for chi	2	.	3	1	1	4	2

Display 5: ADTQ Data Set

The ADTQ data set is primarily used in the table programs. Question numbers are assigned as values for macro parameters to make the programs more robust and efficient. Response text is derived from formats that are programmatically created as described in the following section.

SAS FORMAT CATALOG CONTAINING ALL SURVEY TEXT

Programmatically producing a SAS format catalog directly from all of the information in the Questions spreadsheet will save time and effort and keep the question and response text in one original location. The FORMAT procedure can be used to build the format catalog, but first there are some necessary steps to prepare the data for this procedure.

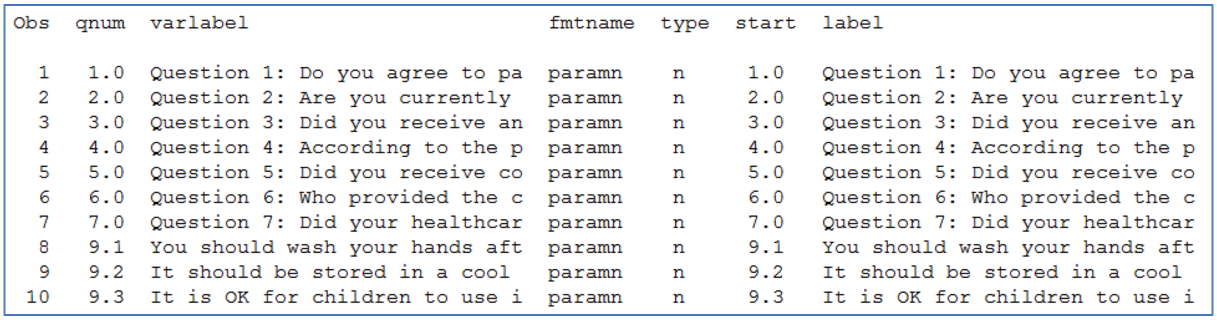
CREATING THE FORMATS

Format for Question Text

A useful format to have in your format catalog is one that contains all of the question numbers and corresponding question text. First, generate a temporary SAS data set containing each unique question number and decode. To do this, use the imported QUESTIONS data set and run a PROC FREQ step on QNUM*VARLABEL, where QNUM is the question number and VARLABEL is the question text. Next, use the resulting data set to further assign all variables that are needed to create the format. Examples of these two steps are shown below:

```
proc freq data=orglib.questions ;
   table qnum*varlabel / out=paramnf (keep=qnum varlabel) noprint ;
run ;

data fmtquest (keep=qnum varlabel fmtname type start label) ;
   length fmtname $14 label $500 ;
   set paramnf ;
   fmtname = 'paramn' ; ** NAME OF FORMAT-TO-BE **;
   type    = 'n' ;
   start   = qnum ;
   label   = strip(varlabel) ;
   output ;
run ;
```

The resulting data set FMTQUEST contains 6 variables. QNUM contains each question number, and VARLABEL contains the question text which originated from the “Label” column in the Questions spreadsheet. The new variable FMTNAME is the name of the format-to-be, which in this example is the numeric format name “paramn” with TYPE set to “n”. A character format would include a leading ‘$’ in FMTNAME and TYPE would be set to “c”. START is assigned to match the question number QNUM, since the starting row in the format will be the first question number. LABEL is assigned to match the question text QLABEL. Display 6 below shows the FMTQUEST data set using a PRINT procedure (where the variables VARLABEL and LABEL are formatted to a length of $30 for printing purposes only).



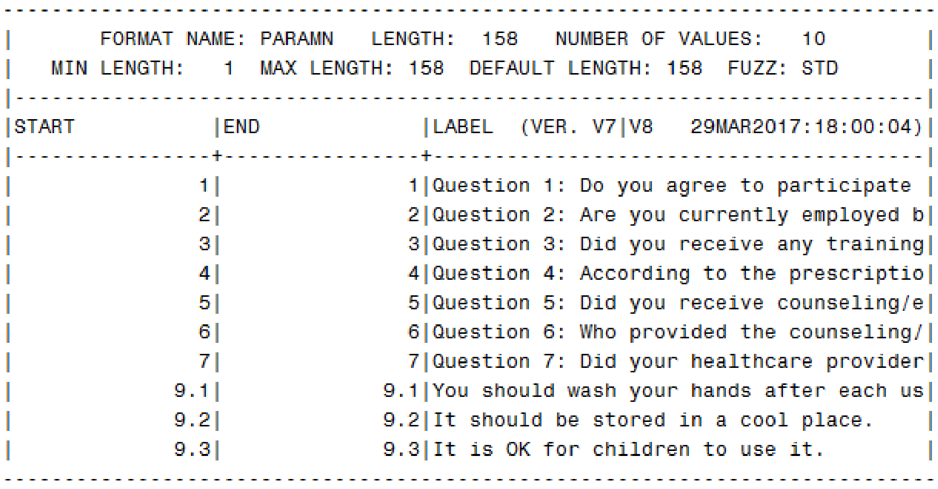
```
Obs  qnum  varlabel                        fmtname  type  start  label

  1   1.0  Question 1: Do you agree to pa  paramn    n     1.0   Question 1: Do you agree to pa
  2   2.0  Question 2: Are you currently   paramn    n     2.0   Question 2: Are you currently
  3   3.0  Question 3: Did you receive an  paramn    n     3.0   Question 3: Did you receive an
  4   4.0  Question 4: According to the p  paramn    n     4.0   Question 4: According to the p
  5   5.0  Question 5: Did you receive co  paramn    n     5.0   Question 5: Did you receive co
  6   6.0  Question 6: Who provided the c  paramn    n     6.0   Question 6: Who provided the c
  7   7.0  Question 7: Did your healthcar  paramn    n     7.0   Question 7: Did your healthcar
  8   9.1  You should wash your hands aft  paramn    n     9.1   You should wash your hands aft
  9   9.2  It should be stored in a cool   paramn    n     9.2   It should be stored in a cool
 10   9.3  It is OK for children to use i  paramn    n     9.3   It is OK for children to use i
```

Display 6: PROC PRINT of Data Set Used to Create Format PARAMN

Next, begin building the SAS format catalog QFORMATS.sas7bcat using PROC FORMAT to add the format “paramn” using the data set above. Additional formats can be appended to the catalog with a similar PROC FORMAT statement:

```
proc format library=adslib.qformats cntlin=fmtqlab ;
run ;
```

Display 7 below shows the contents of the resulting format named PARAMN that can be used to decode the question numbers into question text.



```
----------------------------------------------------------------------------
|       FORMAT NAME: PARAMN   LENGTH:  158   NUMBER OF VALUES:   10        |
|   MIN LENGTH:   1  MAX LENGTH: 158  DEFAULT LENGTH: 158  FUZZ: STD       |
|--------------------------------------------------------------------------|
|START           |END             |LABEL  (VER. V7|V8   29MAR2017:18:00:04)|
|----------------+----------------+----------------------------------------|
|               1|               1|Question 1: Do you agree to participate |
|               2|               2|Question 2: Are you currently employed b|
|               3|               3|Question 3: Did you receive any training|
|               4|               4|Question 4: According to the prescriptio|
|               5|               5|Question 5: Did you receive counseling/e|
|               6|               6|Question 6: Who provided the counseling/|
|               7|               7|Question 7: Did your healthcare provider|
|             9.1|             9.1|You should wash your hands after each us|
|             9.2|             9.2|It should be stored in a cool place.    |
|             9.3|             9.3|It is OK for children to use it.        |
----------------------------------------------------------------------------
```

Display 7: Format PARAMN for Question Text

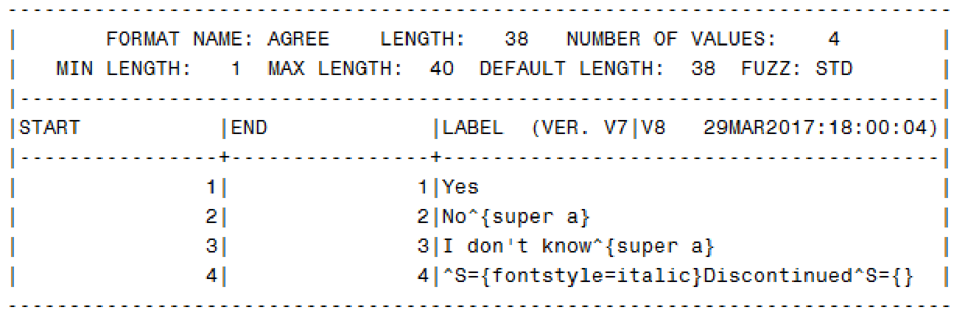
Format for Each Single Response Question

Similar to the above, create a separate format for each variable representing a single response question. The following code shows how to create the formats for single response questions. There are a total of 5 possible responses across all questions in the QUESTIONS data set (variables RESP1 through RESP5). The resulting formats (one per variable name) will contain values 1 through the maximum number of responses for that particular question.

```
data fmtsing (keep=varname nresp fmtname type start label) ;
   set orglib.questions (keep=varname nresp resp: rord where=(rord eq .)) ;
   length fmtname $14 label $500 ;
   type = ‘n’ ;
   array resps {4} $ resp1 - resp5 ;
   do i = 1 to nresp ;
      fmtname = strip(varname) ;
      start = i ;
      label = strip(resps{i}) ;
```

```
         output ;
      end ;
   run ;

   proc format library=adslib.qformats cntlin=fmtsing ;
   run ;
```

Display 8 below shows the contents of the format AGREE that was created for the single response Question 1, “Do you agree to participate in the survey?”



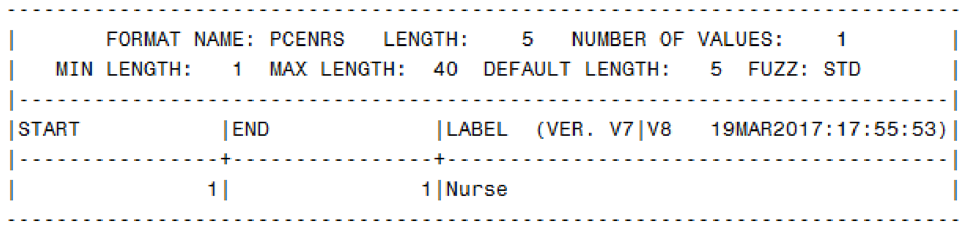
```
----------------------------------------------------------------------------
|       FORMAT NAME: AGREE    LENGTH:   38   NUMBER OF VALUES:    4         |
|   MIN LENGTH:   1  MAX LENGTH:  40  DEFAULT LENGTH:  38  FUZZ: STD        |
|--------------------------------------------------------------------------|
|START           |END             |LABEL  (VER. V7|V8   29MAR2017:18:00:04)|
|----------------+----------------+----------------------------------------|
|               1|               1|Yes                                     |
|               2|               2|No^{super a}                            |
|               3|               3|I don't know^{super a}                  |
|               4|               4|^S={fontstyle=italic}Discontinued^S={}  |
----------------------------------------------------------------------------
```

Display 8: Format AGREE for Question 1 Responses

Format for Each Multiple Response Question

Similar logic can be applied to create a format for each of the multiple response questions. The difference is that each format will have a single value, because in a multiple response question, each response is either denoted as selected or not selected.

Display 9 below shows the contents of the format named PCENRS that correspond to the multiple response Question 6 and the individual response “Nurse”.

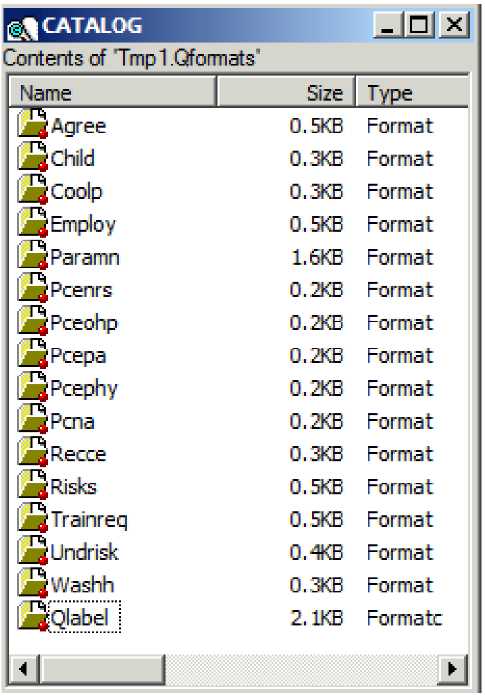


```
----------------------------------------------------------------------------
|       FORMAT NAME: PCENRS   LENGTH:    5   NUMBER OF VALUES:    1         |
|   MIN LENGTH:   1  MAX LENGTH:  40  DEFAULT LENGTH:   5  FUZZ: STD        |
|--------------------------------------------------------------------------|
|START           |END             |LABEL  (VER. V7|V8   19MAR2017:17:55:53)|
|----------------+----------------+----------------------------------------|
|               1|               1|Nurse                                   |
----------------------------------------------------------------------------
```

Display 9: Format PCENRS for Question 6 “Nurse” Response

THE FINAL FORMAT CATALOG

Display 10 below shows the entire format catalog QFORMATS.sas7bcat that resides in the ADSLIB library, and it contains a format for each of the items mentioned above.



Contents of 'Tmp1.Qformats'

Name	Size	Type
Agree	0.5KB	Format
Child	0.3KB	Format
Coolp	0.3KB	Format
Employ	0.5KB	Format
Paramn	1.6KB	Format
Pcenrs	0.2KB	Format
Pceohp	0.2KB	Format
Pcepa	0.2KB	Format
Pcephy	0.2KB	Format
Pcna	0.2KB	Format
Recce	0.3KB	Format
Risks	0.5KB	Format
Trainreq	0.5KB	Format
Undrisk	0.4KB	Format
Washh	0.3KB	Format
Qlabel	2.1KB	Formatc

Display 10: Final Format Catalog

TABLE PROGRAMS AND OUTPUT

KEY RISK MESSAGE (KRM) ANALYSIS TABLES

For analysis purposes, certain survey questions are often grouped together as part of a Key Risk Message (KRM). In general, a KRM is a statement that reflects one or more objectives of the KAB survey. A KAB study may have more than one KRM. A patient survey may have a KRM such as: “It is important to follow the conditions of safe use of MADEUP treatment.” Questions related to the patient’s understanding of the importance of safe use may be included as part of this KRM. The analysis tables typically summarize each KRM separately, including the corresponding questions and responses associated with the KRM.

APPLICATIONS OF THE FORMAT CATALOG

The formatted values in the format catalog will automatically inherit any special formatting, RTF code, etc. that was originally entered into the Questions spreadsheet then imported into the Questions SAS data set, which was then used to programmatically create the formats. Furthermore, if any text updates are needed while programming the analysis tables, the updates only need to be made in the source location, the Questions spreadsheet. Subsequently, the new spreadsheet can quickly be reimported (to generate an updated Questions data set), and the SAS format catalog regenerated. The formats will then reflect the new updates and will be ready to use in table programs to appropriately format the question and response text.

For Single Response Questions

The following code in a table program shows how to use the ADTQ data set to calculate counts and percents, and then apply the formats from the format catalog. The steps are run separately for each question using a macro %DO loop. Formats from the format catalog are applied using the PUTN and PUT statements.

```
%macro mcounts ;
   ** LIST OF SINGLE RESPONSE, NON-ELIGIBILITY QUESTIONS **;
   %let mcqs = 4 5 7 9.1 9.2 9.3 ;
   %do x = 1 %to 6 ;  ** DO-LOOP FOR EACH QUESTION **;

      %let dovar = %scan(&mcqs,&x,' ') ; ** GET VAR NAME FOR QUESTION **;

      ** GET COUNTS **;
      proc freq data=adtq (where=(paramn eq &dovar)) ;
         table paramn*paramcd*aval / out=cnts (drop=percent) noprint ;
      run ;

      data cnts (rename=(aval=statord)) ;
         set cnts ;
         if count gt 0 then pct = round((count/&pop)*100, .1) ;
         length prntname statname $500 ;
         statname = putn(aval, paramcd) ;
         prntname = put(paramn, paramn.) ;
      run ;

      ** COMBINE INDIVIDUAL QUESTION COUNTS **;
      %if &x = 1 %then
      %do ;
         data mcnts ;
            set cnts ;
         run ;
      %end ;
      %else
      %do ;
         data mcnts ;
            set mcnts cnts ;
         run ;
      %end ;
   %end ;  ** END OF DO-LOOP FOR EACH QUESTION **;

%mend mcounts ;
%mcounts
```

For Multiple Response Questions

The following code shows how similar steps are applied for multiple response questions. The main difference here is that a macro %DO loop is not needed, since all of the multiple response question variables have values of 1; therefore, the PUTN statement can reference ‘1’ for all of the questions.

```
proc freq data=adtq (where=(prntord eq 6)) ;
   table paramn*rord*paramcd*aval / out=ocnts (drop=percent) noprint ;
run ;

data ocnts (rename=(rord=statord)) ;
   set ocnts (drop=aval) ;
   if count gt 0 then pct = round((count/&pop)*100, .1) ;
   length prntname statname $500 ;
   statname = putn(1, paramcd) ;
   prntname = put(paramn, paramn.) ;
run ;
```

The resulting data sets OCNTS (for multiple response questions) and MCNTS (for single response questions) are then SET together to further calculate summary statistics for the table output.

APPLICATIONS OF THE QUESTIONS DATA SET

Besides the questions and response survey text, the Questions data set (originating from the Questions spreadsheet) is also helpful for the table programs. The NRESP variable provides the number of responses for each question when transposing the responses in a vertical data set. The CORRECT variable makes it easy to access the desired correct response, if any, for a question through various table programs. The QNUM and RORD variables provide the sort order so that the questions are displayed in the same order as they appear in the survey.

For Creating Dummy Data Set for Zero Counts

When displaying counts and percentages of each response in the output tables, any response in the survey that is not picked by any respondent will not be displayed by a PROC FREQ. For those missing responses, a dummy data set is created from the Questions data set which includes all the possible responses for each survey question. This dummy data set is merged with the resulting data set from a PROC FREQ and any missing response in the data set is zero-filled.

```
data dummy (rename=(qnum=prntord varlabel=prntname varname=paramcd)) ;
   set orglib.questions (where=(qnum in (9.1 9.2 9.3))) ;
   length statname $500 ;
   array resps {*} resp1-resp5 ;
   if qnum in(9.1 9.2 9.3) then
     do i = 1 to nresp ;
        statord = i ;
        statname = resps{i} ;
        output ;
     end ;
run ;

data final ;
   merge dummy (in=indummy) all (in=inall) ;
   by prntord statord ;
   ** ZERO FILL ROWS **;
   if indummy and not inall then cat_1 = '0' ;
run ;
```

Table 3 below shows that the “I don’t know” response for the first sub-question under Question 9 has a count of 0. In the ADTQ data set, for PARAMN = 9.1, there is no AVAL = 3. Using the dummy data set ensures that the AVAL = 3 is included in the output table.

Key Risk Message #1: It is important to follow the conditions of safe use of MADEUP treatment.

Question	**Completed Surveys (N=234) n (%) [95% CI]**
Question 9: Please answer True, False or I don’t know for each of the following statements about MADEUP treatment:	
You should wash your hands after each use.	
True[a]	226 (96.6) [93.4-98.5]
False	8 (3.4)
I don’t know	0
It should be stored in a cool place.	

Key Risk Message #1: It is important to follow the conditions of safe use of MADEUP treatment.

Question	Completed Surveys (N=234) n (%) [95% CI]
True	6 (2.6)
False[a]	224 (95.7) [92.3-97.9]
I don’t know	4 (1.7)
It is OK for children to use it.	
True[a]	171 (73.1) [66.9-78.6]
False	18 (7.7)
I don’t know	45 (19.2)

[a] Correct response.

Table 3: Questions Linked to Key Risk Message #1 – Completed Surveys

For Summarizing Correct Response Statistics in KRM Tables

The analysis tables usually require that the correct responses for questions pertaining to a KRM have additional statistical calculations like confidence intervals. To calculate the confidence interval, the correct response for any question can be easily referred to in the Questions data set. The variable CORRECT contains the desired correct responses for the questions. A correct flag CVAL is given a value of 1 if the variable value (response chosen) matches the CORRECT value, and a value of 0 if it doesn’t. A simple PROC FREQ of CVAL (using the appropriate confidence interval calculation specified in the SAP) will give the confidence interval values. The previous example (Table 3) shows how the confidence intervals are displayed for the correct response of the KRM questions.

The following is an abbreviated form of the program to show how CVAL is derived and counts of each CVAL value are obtained, then how certain confidence intervals may be calculated:

```
data adtqci ;
   merge questions adtq ;
   by usubjid qnum aval ;
   if aval eq correct then cval = 1 ;
   else cval = 0 ;
run ;

proc freq data=adtqci ;
   by prntord statord ;
   table cval / out=cicnts noprint ;
run ;

** CALCULATE 95% EXACT BINOMIAL CIs **;
proc freq data=cicnts noprint ;
   by prntord statord ;
   table cval / binomial (level=2) alpha=0.05 ;
   weight count / zero ;
   output out=stats binomial ;
run ;

data ci ;
   length ci $50 ;
   set stats ;
```

```
     by prntord statord ;
     if nmiss(xl_bin,xu_bin) = 0
        then ci = cat("[", strip(put(100*xl_bin,5.1)),"-",
                           strip(put(100*xu_bin,5.1)), "]") ;
  run ;
```

WHEN UPDATES ARE NEEDED TO QUESTION/RESPONSE TEXT

Sometimes during the programming process, it might not be immediately obvious where there are typos and other problems in the question and response text until they are found while reviewing the output tables. When this happens, if it is determined that the problem originated in the Questions spreadsheet, it is important to make the update at the source. After any update is made to the Questions spreadsheet, it must be reimported into the Questions SAS data set, followed by the re-running of the program to generate the format catalog, then the ADTQ program, and so on. If updates are only needed to the Questions spreadsheet, the re-runs of the subsequent programs should be fairly quick. Figure 1 below shows a flowchart of the programming process, including what to do when updates are needed to the Questions spreadsheet.

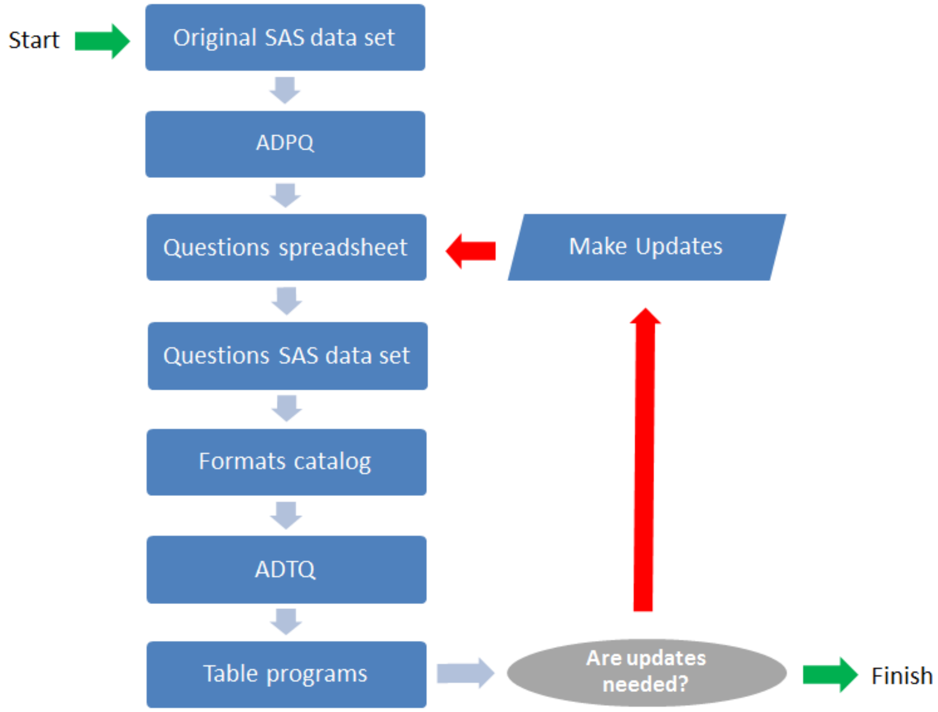


Figure 1: Flowchart of Programming Process

CONCLUSION

The seemingly complicated task of writing programs for a KAB survey can be streamlined to produce quality reports via a strategic plan to handle the characteristics of KAB survey data. Understanding the different question types and planning how to handle each one in the analysis data sets to assign variables and sort order is important. Planning ahead and entering all question and response text into a single Questions spreadsheet will achieve significant efficiency, and more importantly, will eliminate manual error and will result in consistency across tables, listings, and figures. Adding question numbers, variables and correct answers will help link the analysis data sets and conveniently allow programmers to know what to expect for each value in the data set. Programmatically creating a format catalog from all of the question and response text in the Questions spreadsheet further enhances productivity and accuracy.

All of these ideas combined lead to a smoother programming process. The upfront work in preparing the Questions spreadsheet, ADPQ and ADTQ analysis data sets, and format catalog allows the reporting programs to be more robust. Once a system is put in place, many similar table, listing, and figure programs from one study can be seamlessly used for another study with minor modifications.

REFERENCES

Base SAS® 9.2 Procedures Guide: Statistical Procedures, Third Edition. The FREQ Procedure, Statistical Computations, Binomial Proportion. Available at http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_freq_a0000000660.htm

Nair, Indu; Patel, Binal. 2014. Attain 100% Confidence in Your 95% Confidence Interval. *Proceedings for PharmaSUG Conference 2014.* Available at http://www.pharmasug.org/proceedings/2014/IB/PharmaSUG-2014-IB05.pdf

ACKNOWLEDGMENTS

We would like to sincerely thank Carol Matthews from United BioSource Corporation for her significant contributions in developing SAS programs and processes to efficiently handle KAB survey data. Her experience, expertise, and recommendations for this paper have been invaluable.

CONTACT INFORMATION

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

Cara Lacson
United BioSource Corporation, Blue Bell, Pennsylvania, USA
cara.lacson@ubc.com

Jasmeen Hirachan
United BioSource Corporation, Blue Bell, Pennsylvania, USA
jasmeen.hirachan@ubc.com