

Using SAS® Special Missing Values to Enhance Display and Analysis of Survey Research Data

Francine Cannarozzi,
Abt Associates Inc., Chicago, IL

Most survey research instruments have a number of internal skip pattern and sample universe logic that result in blank or missing values. Because analysis of missing values and non-response is an important part of survey research methodology, it is important to distinguish real missing data, [i.e. data that should have been collected but was not], from legitimately missing data, [i.e. data for questions that the respondent was not asked], in reports, data files, and data file documentation. This paper will detail the ways to use SAS® special missing values to simply and effectively report these values, and how they can relieve unnecessary coding prior to many PROC steps. The use of special missing values has merit in many fields, not just survey research; therefore many general examples will be provided to showcase the broad utility of SAS special missing values.

WHAT IS MISSING DATA?

Absence of data. No data. Missing data. Non response. These are all terms that survey researchers use to describe blank fields in a data set. Blank fields are naturally expected in survey research data because they are built into the design. For example, consider the question:

"Do you have any children under 18 living with you in the household?"

Of course, if your response was, "no", the next question would not be asked, which is:

"How many children are living in your household?"

The follow up question would be blank for that portion of the survey respondents who answered "no" to the initial question. Simple tables might appear like this:

Q1. "Do you have any children under 18 living with you in the household?"

	<u>Freq</u>	<u>Cuml. Freq</u>
1-YES	123	123
2-NO	67	190

Q1a. How many children are living in your household?

	<u>Freq</u>	<u>Cuml. Freq</u>
1	79	79
2	28	107
3	10	117
4+	6	123

Frequency missing = 67

The mysterious **frequency missing** is the subject of this paper. It is simple to understand from the example that the cases classified as missing are all due to the skip logic of the questionnaire. It is not a perfect world and more often, tables from survey data really look like this:

Q1. "Do you have any children under 18 living with you in the household?"

	<u>Freq</u>	<u>Cuml. Freq</u>
1-YES	120	120
2-NO	65	185
6-DK	2	187
7-REF	1	188

Frequency missing = 2

Q1a. How many children are living in your household?

	<u>Freq</u>	<u>Cuml. Freq</u>
1	78	78
2	24	102
3+	15	117

Frequency missing = 73

Now the interpretation is not as simple. There are more missing cases than can be explained by the skip pattern.

There are many reasons a field can be blank. The most common reason is due to an error – for whatever reason the data was not collected. Many other valid reasons will result in a blank field, but without deep knowledge of the project methodology, it all looks like missing to the analyst who gets the data file.

USING SAS® SPECIAL MISSING VALUES

When working with numeric data, a missing value is normally represented by a simple (.) as shown in the first examples. By assigning a special missing value to each of the missing cases, you can distinguish between the reasons the data are missing in your dataset, and show a complete picture for each and every table. In this way, you will answer all questions about missing data before they arise!

SAS allows the use of 27 special missing values, each letter of the alphabet plus the underscore (_). A special missing value is represented in numeric data by two characters: a decimal point followed by either a letter or the underscore. Hopefully you would not have 27 different types of missing data, but you have the capacity for such. Let's look at our sample tables again, after assigning the special missing values and assigning formats to them:

Q1. "Do you have any children under 18 living with you in the household?"

	<u>Freq</u>	<u>Cuml. Freq</u>
F-In follow up	1	1
M-Missing in error	1	2
1-YES	120	122
2-NO	65	187
6-DK	2	189
7-REF	1	190

Q1a. How many children are living in your household?

	<u>Freq</u>	<u>Cuml. Freq</u>
F-In follow up	1	1
L-Legitimate skip	68	69
M-Missing in error	4	73
1	78	151
2	24	175
3+	15	190

This is a much different story. Each case is clearly accounted for and not assumed to be simply missing. All tables consistently total to the total number of cases. At the department in which I work, we have a number of standard defined missing values that are used for all survey research data by all SAS programmers. Our clients have found it very useful to review tables in this way and find they are able to understand each data point.

We use special missing values to define these conditions always encountered in our survey research data.

.N Not in universe (sample logic), i.e., respondents skip *entire* sections of questions based on eligibility criteria, e.g., age or demographic characteristics, or membership in a special sample sub-category.

.M Missing in error, i.e., a response should have been captured for this question, but was not.

.L Legitimate skip (question logic), i.e., respondents skip one or more questions *within a section* because of an answer selected for a root question.

.P Partially completed case, i.e., question not answered because the respondent broke off the interview prior to completing this question. Different projects have their own definition for a partially completed case, based on how far they got in the interview.

.A Added question, i.e., this question was added after the start of data collection and the respondent was interviewed before the question was added to the interview.

.B Break off, i.e., to indicate no interview was complete but other data may have been gathered.

.X No recoded response, i.e., these cases did not have a recoded verbatim response for the category variable created during data processing.

Appendix A shows an example first, with no special missing values, and again, using the special missing values.

Using special missing values adds to the display of your data in a number of ways. Certainly they increase understanding of data and can aid a project manager or analyst in planning for analysis. When you use special missing values to monitor data collection and when generating reports you can quickly identify problem areas in the data.

In survey research often the data not collected is as important as the data that is collected, referred to as a non-response analysis. Especially useful in non-response analysis, special missing values immediately highlight a non-response from an operational missing value like a legitimate skip.

Survey research is not the only field where special missing values can be utilized. For instance, think of how to apply special missing values to an inventory database. Instead of blank fields, now you can classify missing data points by assigning special missing values to them. For example, a location field might be missing for those offices not carrying a specific unit; a size or color field will not be applicable to every item; real missing data needs to be differentiated from missing information.

WORKING WITH SPECIAL MISSING VALUES IN YOUR DATA

ASSIGNMENT

Using special missing values in your SAS code is facilitated by a number of conventions in the SAS language. To code a special missing value, use a standard assignment statement:

```
if school = . then school = .A;
      is the same as
if school = . then school = .a;
```

Either upper or lower can be used with coding and also when referring to special missing values. Always use the decimal point and the letter to refer to a special missing value; if you do not use the decimal point, SAS thinks you are referring to a variable.

Treat a special missing value like a standard value to call them in an assignment statement:

```
if (location= 6 and quantity = .b) then hold_flag = 1;
```

CALCULATIONS AND PROC STEPS

Even though you can have 27 different missing values, in certain circumstances, SAS treats them all as just plain missing. Most importantly is in the way they are handled in PROC statements. They are seen as missing for numerical calculations. So you have the benefit of displaying your data with full description, yet you have nothing to do in order to drop them from calculations when appropriate, SAS handles this for you.

SORTING

When sorting data that contain special missing values, they are treated as missing, that is to say, they are sorted first. If regular (.) missing also exist, then the hierarchy is

```
. (standard missing) then
.a through ._ (special missing) then
negative values then
positive values.
```

MERGING AND UPDATING

Normally a missing value in the update set will not overwrite a value in merging or updating data sets, which is the desired action. However, when merging or updating, a special missing value is treated as a true value, and will replace another value. This special handling is very useful when applying transaction datasets in your data editing or data cleaning process.

PRINTING

When printing, an unformatted special missing value appears as simply the letter or the underscore. No decimal point is printed. Also, in report printing such as PROC FREQ, you must use the "missing" option to display special missing values in your table; otherwise they are left out of the table.

FULL SCREEN DISPLAY

When viewing datasets interactively, the entire special missing value is displayed, including the decimal point.

FORMATTING

When creating user defined formats treat special missing values as any regular value and assign it a format.

```
VALUE S2Q01C
  1 = "1 - EXCELLENT"
  2 = "2 - VERY GOOD"
  3 = "3 - GOOD"
  4 = "4 - FAIR"
  5 = "5 - POOR"
  6 = "6 - DON'T KNOW"
  7 = "7 - REFUSED"
.L = ".L - LEGITIMATE SKIP"
.M = ".M - MISSING IN ERROR"
.B = ".B - BREAK OFF"
.N = ".N - NOT IN UNIVERSE"
```

SUMMARY

Using special missing values allows flexibility in the way you present data. The use of SAS special missing values enables you to display data with rich detail in tables, codebooks, and reports. Documentation is enhanced with detail in an easily readable form, making your product professional and comprehensive. Special missing values aid in analysis of trends and abnormalities in the data by allowing easy classification for various levels of missing data. Special missing values are easy to work with, and are handled intuitively by the SAS system.

APPENDIX A

Example of Frequency Table without Special Missing Values

Q3: HOURS PER WEEK SPENT WATCHING TELEVISION

<i>Response Categories</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
0-LESS THAN ONE HOUR	167	167
1-ONE TO THREE HOURS	1527	1694
2-FOUR TO SIX HOURS	9887	11581
3-SEVEN TO TEN HOURS	6855	18436
4-MORE THAN TEN HOURS	5041	23477

Frequency Missing = 20072

Example of Frequency Table with Special Missing Values

Q3: HOURS PER WEEK SPENT WATCHING TELEVISION

<i>Response Categories</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
.L-LEGITIMATE SKIP	12984	12984
.M-MISSING	83	13067
.N-NOT IN UNIVERSE	6888	19955
.P-PARTIAL INTERVIEW	117	20072
0-LESS THAN ONE HOUR	167	20239
1-ONE TO THREE HOURS	1527	21766
2-FOUR TO SIX HOURS	9887	31653
3-SEVEN TO TEN HOURS	6855	38508
4-MORE THAN TEN HOURS	5041	43549