Paper 85-26

SAS Formats: Making the Best of a Bad Situation

Michelle Pritchard, The Lewin Group, San Francisco, California

ABSTRACT

A common problem for SAS programmers when working with categorical data is how to utilize SAS format capabilities that associate labels with the values of a variable. It is often hard to decide whether to permanently associate the format with the variable. Sometimes you want to work with the formatted values and sometimes you want to work with the underlying numeric values. Also, this decision could have implications for quality control. Other issues to consider are: 1) Should the original value be included in the format? 2) When is it appropriate to reorder categories using formats? 3) How can format libraries be transferred to other operating systems? and 4) What issues arise when using formats with SAS/GRAPH? This paper will discuss these common considerations, with examples, and provide recommendations.

INTRODUCTION

The SAS formats and user defined formats allow programmers to change the look of SAS data. This is somewhat scary, but potentially extraordinarily helpful! So as not to wreak havoc on our databases, it is best to make sure that we know what we are doing when we use SAS formats.

SAS version 6.12 was used throughout this paper.

THE PERMANENT FORMAT DILEMMA

Working with formatted values in certain SAS procedures can provide useful clarification. For example, formats enable us to match a numeric value to a text description when presenting frequencies.

(A) NOT AS HELPFUL:

SATISFY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	30.0	3	30.0
2	2	20.0	5	50.0
3	3	30.0	8	80.0
9	2	20.0	10	100.0

(B) MORE HELPFUL:

SATISFY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Very	3	30.0	3	30.0
Somewhat	2	20.0	5	50.0
Not at all	3	30.0	8	80.0
No response	2	20.0	10	100.0

(Note that in the FREQ procedure, the ordering is internal by default)

Using formats in example B helps with the interpretation of the frequencies. When in production mode for an extensive analysis, it is tempting to assign variables a permanent format, thereby freeing the programmer from having to specify the format in each procedure. This also ensures that the same formats are used each time. But what happens if a programmer wants to return to the original value to do additional data step programming? Depending on the existing study documentation, returning to the original value of a variable may or may not be a straightforward task. You can always match original values and formats using a Proc Contents and an examination of the format catalog using the following code:

```
libname library "path";
proc format fmtlib lib = library;
 select <format name>; (this is a helpful
optional statement)
run;
```

Which produces:

FORMAT NAME: SATNHF LENGTH: 11 NUMBER OF VALUES: 4
MIN LENGTH: 1 MAX LENGTH: 40 DEFAULT LENGTH 11 FUZZ: TD

START	END	LABEL (VER. 6.12 30JUN00:14:05:10)
1	1	Very
2	2	Somewhat
3	3	Not at all
9	9	No response

This can be tedious when there is an abundance of variables. An alternative way of creating formats is to include the original value next to the text description. Not only does this provide a simple snapshot of the format and original value, it also keeps the internal and formatted ordering the same.

(C) MOST HELPFUL:

SATISFY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1: Very	3	30.0	3	30.0
2: Somewhat	2	20.0	5	50.0
3: Not at all	3	30.0	8	80.0
9: No response	2	20.0	10	100.0

Analytic procedures often use the numeric (internal) value; therefore it is important to know that value to be sure the analysis is correct. For example, “neutral” values (such as not applicable or unknown) often appear as the highest value rather than in the middle.

When formats are permanently associated with a variable, it can be difficult to perform quality control. For example, if some values are reordered and the format is not removed then the format values will be incorrect. It is also more difficult to confirm that code is referring to the proper numeric values.
It is always nice to work with dates that have been permanently assigned appropriate date-time formats.

REORDERING CATEGORICAL VARIABLES

It is often helpful to examine the same variable in different ways. For example, neutral values such as no response or unknown may occasionally be reordered or set to missing. If you are routinely creating different flavors of the same variable, it is probably best to create completely different variables rather than using formats to differentiate the flavors.

TRANSFERRING FORMAT LIBRARIES

When transferring datasets it is particularly important to include the format catalog if the formats have been permanently

assigned. To transfer to a different operating system, use Proc Cport to convert the format catalog into transport form. The following code can be used:

```
filename xportf "path\xportf.dat";
libname  library "path";
proc cport catalog = library.formats
file=xportf;
run;
```

The receiving party can then use Proc Cimport to transfer the catalog back into a SAS catalog on their system.
Without the format catalog, permanently formatted data will not be accessible unless the formats are stripped off. The formats can be removed using the following format statement within a data step:

```
data temp02;
  set temp01;
  format _all_;
run;
```

Whether the transferred dataset is permanently formatted or not, it is a good idea to send along a format catalog for documentation purposes. In fact, perhaps the best way to send formats is to provide SAS code for Proc Format to create all the necessary formats together with a simple data step to associate the format with the variable. If the SAS dataset does not have permanent formats assigned, this approach allows the recipient to assign the formats permanently or not as desired.

FORMATS AND SAS/GRAPH

SAS/GRAPH allows a large amount of labeling flexibility. Programmers can override the variable values and labels within the axis statement. As a result, it is best to work with the underlying value of the variable instead of the formatted value. Keeping track of the underlying value, the formatted value, and how they both relate to the user specified axis statements is too overwhelming.

CONCLUDING RECOMMENDATIONS

At the onset of any analysis, write a program to create one permanent format library that will be updated throughout the duration of the analysis. The program can be entitled Formats.sas and it will create the format library Formats.sc2. For each additional format added to the catalog, update the Formats.sas program by including the new format. Existing formats should not be written over. At the end of the analysis, the log and output from Formats.sas can be used as study documentation.

It is recommended that formats contain a practical ordering, which may depend on the task at hand. This makes reading SAS output easier. When appropriate, the format should contain the original value that links to the original source document. For example, the variable Health has values ranging from 1 – 5. An appropriate format would be:

1 → “1: Excellent”
2 → “2: Very Good”
3 → “3: Good”
4 → “4: Fair”
5 → “5: Poor”

Similarly, for regression analyses, formats should be created so that the reference cell represents the desired group. For instance, the variable Asthma is coded to either 1 (having asthma) or 0 (not having asthma). So that “No Asthma” is the reference group (Order = Formatted may or may not be required in the first line of the PROC depending on the default), a suggested format would be:
0 → “N: No Asthma”
1 → “A: Asthma”

It is strongly advised that original variables not be assigned a permanent format. However, in some circumstances, this recommendation can necessarily be violated (e.g., an annotated paper form does not accompany data file). In the event that this situation occurs, formats should be created so that they:

- Are one-to-one (i.e., categories are not combined.)
- Contain the original value within the format itself
- Are not reordered

In addition, any assignment of permanent formats to original variables should be done in one program.

CONTACT INFORMATION

The author welcomes questions and comments. She can be contacted at:

Michelle Pritchard
The Lewin Group
490 2nd Street, Suite 201
San Francisco, CA 94118
Work Phone: 415-495-8966
Fax: 415-495-8969
Email: michelle.pritchard@lewin.com