Planning, Preparing, Documenting, and Referencing SAS Products

RoJeanne W.L. Liu, U.S. General Accounting Office
Arthur D. Foreman, U.S. General Accounting Office

ABSTRACT

This guide discusses quality control and documentation methods individuals can use to plan, prepare, document and review SAS products. It is intended for audit and program evaluation applications. The environment used here in this discussion is the U.S. General Accounting Office (GAO). The GAO is an independent, nonpartisan agency which assists Congress in congressional oversight of the executive branch of the federal government.

INTRODUCTION

Because properly preparing, documenting, and referencing SAS products can be intricate and demanding in an audit environment or on a special project, this guide will assist SAS users to develop products that conform to GAO quality control and work paper standards.

The sequence of topics discussed follows the normal order of tasks in an audit assignment or general project;

*planning work that may involve SAS;
*ensuring correctness in SAS work;
*entering data into SAS from raw data formats, SAS data files, other software formats, and data bases;
*transferring SAS data between computers;
*documenting SAS work;
*referencing or reviewing SAS work; and
*storing SAS work papers and files.

This guide complements SAS training and reference manuals and supplements the Government Auditing Standards and federal information data processing standards.

PLANNING

Using SAS to support an audit assignment requires careful planning. To correctly decide if you should use SAS and, if so, how, you must have explicitly defined the assignment's objectives and the specific questions to be answered. Planning to use SAS involves

* defining the analysis techniques you will use,
* determining the source of data, estimating staff proficiency in SAS and statistics,
* determining available computer resources, and
* determining available support material.

The result of this planning effort is an initial SAS work plan. This plan, developed after defining audit objectives and specific information needs, should be included as part of the overall audit plan or evaluation design.

Deciding to Use SAS

Selecting the most appropriate package(s) depends on matching the assignment's needs with a software's strengths. The following questions will help you decide if SAS is the most appropriate.

* What methodology techniques do you need?
* How is agency data stored?
* Is the staff proficient in using SAS?
* Does the assignment team have access to computer resources capable of processing SAS programs?
* Does the assignment team have access to SAS support material?

SAS Work Plan

Your assignment team needs to decide on an initial SAS work plan. This plan should be a part of the overall assignment plan or evaluation design. The SAS work plan is important since it outlines the methodologies, programs, and procedures that are needed and shows what data will be captured and verified. It logically ties together the information needed, programs, data files, and reports.

The SAS work plan may include a process flow chart or diagram supported by a written narrative explaining the relationship among major programs, procedures, windows, data files, manual procedures, and reports. As a minimum, the plan must describe the following:

* what the assignment objectives are,
* how data will be gathered, transferred, or converted to a machine-readable form,
* how data will be checked,
* when, if ever, a permanent SAS data file will be created,
* when and how data will be transferred between SAS and other programming packages,
* what major analysis procedures will be used and how they relate to the assignment objectives,
* how programs will be tested and reviewed,
* what major reports and data files will be produced, and
* who will be responsible for the analysis and review.

CHECKING YOUR DATA AND SAS PROGRAMS

Quality evidence and sound analysis are necessary to support findings, conclusions, and recommendations. To ensure the quality of your evidence and the soundness of your analysis in SAS-based work you must use data-checking techniques and program verification procedures that are tailored to SAS. This section displays a variety of such techniques and procedures.

Checking Your Data

Reasonableness checks

- You can identify invalid categorical data and the observations from which the data came with the IF... THEN PUT ... statement and the IN function. Code the request in the following manner:

```
IF CATVAR NOT IN('VAL1','VAL2','VAL3') THEN
        PUT CATVAR= GAOID=    'Unacceptable value';
```

where CATVAR is the categorical variable being checked,
VAL1, VAL2, and VAL3
are the acceptable values, and
GAOID is the identifier for the observation.

- SAS will help you select a sample of data for you to compare for consistency with your source file. For example, you may want to select 10 observations from the beginning of the file, 10 observations from the end of the file, and 5 percent randomly from a SAS file of 10010 observations. You can use an IF statement as follows:

```
IF (_N_ LT 10) OR
   (_N_ GT 10000) OR
   (RANUNI(0) LT .05);
PROC PRINT;
```

where _N_ is a SAS variable to count observations and
RANUNI is a SAS random number generator.

Missing data checks

- SAS will check for missing data in categorical variables every time you use the PROC FREQ procedure. In the following example, all values, including missing values, will be identified. Code the request in the following manner:

```
PROC FREQ;
        TABLES CATVAR / MISSING;
```

where CATVAR is the categorical variable being checked.

- SAS will check for missing data in numeric variables with the PROC MEANS or PROC UNIVARIATE procedure. These procedures will report the number of observations with missing values. Code the request in the following manner:

```
PROC UNIVARIATE;
        VAR NUMVAR;
```

where NUMVAR is the numeric variable being checked.

Record error checks

- SAS can help you locate unexpected duplicate observations. You can locate duplicates by first sorting the file with a PROC SORT procedure and then using an IF ... THEN ... PUT statement with the LAST and FIRST modifiers on the variable. Code the request in the following manner:

```
PROC SORT;
        BY GAOID;

DATA; SET;
        BY GAOID;
        IF NOT (LAST.GAOID AND FIRST.GAOID) THEN
        PUT GAOID= 'Possible duplicate';
```

where GAOID is the variable where a duplicate value is possible,
FIRST.GAOID is a variable indicating the GAOID is the first observation in a group with the same value of GAOID, and
LAST.GAOID is a variable indicating the GAOID is the last observation with the same value of GAOID.

If duplicate observations exist, they will not be both the first and last observation with a particular value. For the above example, every duplicate will be listed on the SAS log.

- When an individual observation is structured as multiple records or lines, you can resynchronize the input after a missing record or a missing line by using the LOSTCARD statement. The LOSTCARD statement reports unexpected changes in the record identifier and is used in conjunction with the IF ... THEN PUT ... statements. The input file must be ordered by the record identifier. In the following example, the input data file has two records per observation:

```
DATA;
        INFILE RAWDATA;
        INPUT GAOID1 NUMVAR1 #2 GAOID2 NUMVAR2;
        IF GAOID1 NE GAOID2 THEN
            DO;
                PUT GAOID1= GAOID2= 'Record error';
                LOSTCARD;
            END;
```

where NUMVAR1 is a variable,
NUMVAR2 is another variable,
GAOID1 is the identifier for the observation on the first record, and
GAOID2 is the identifier for the observation on the second record.

Checking Your Program

A program is reliable if it is performing as expected. SAS programs depend on several simple practices to enhance program reliability and error detection and to reduce the time to correct errors.

Before writing a SAS program, you should establish its specific purpose. Include a general description of your program and the

techniques you will use to verify that the program is working correctly. These details will assist in selecting the appropriate SAS statements and procedures, as discussed below.

Program structure

- Preprogrammed procedures and built-in formats are more likely to be error free than programs and formats you write. If a preprogrammed procedure or built-in function exists and fits your analysis needs, use it. For example, the descriptive statistics produced by the PROC MEANS procedure can be duplicated using a DATA step, however, the PROC MEANS procedure takes less time to program and gives reliable results.

Readability

- Put a program description near the beginning of every SAS program and put a description of the SAS data step or procedure step before every major logical grouping. At the beginning of the program, you can create a comment block with the following information: title of the assignment or project, programmer/analyst's name, program name and date prepared, program description or purpose, and data source. For example:

```
/*************************************************
    Review of Medical Costs
        <123456>

    Programmer: J.D. Programmer

    Program : MEDCST01.SAS
        August 31, 1992

    This program edits the cost file
    for bad entry codes. Bad codes
    are listed on a print file for
    manual validation.

    Source: MEDCT000 Cost File
*************************************************/
```

- Put global titles near the beginning of the program. The first title line, TITLE1, will contain the assignment name; the second line could contain other identification information. For example, at the GAO, if would be the identifying assignment code. These titles will remain constant throughout the SAS program. Every procedure that produces output will have an associated title which explains the output. For example:

```
TITLE1 'Review of Medical Costs';
TITLE2 '<123456>';

PROC PRINT;
        WHERE SSN EQ '' AND FY EQ 91;
        TITLE4 'Records Lacking Social Security Numbers';
        TITLE5 'for Fiscal Year 1991';
```

- Align items for readability. The following two program examples look exactly the same to SAS but the latter is more readable.

```
DATA; INFILE RAWDATA; INPUT @17 VAR1 $5. #2  @3
VAR2 MMYYDD8. @15 VAR3 5.2; IF VAR1 NE VAR2 OR VAR1
NE VAR3 THEN VAR1=VAR3; RUN; PROC PRINT; RUN;
```

```
DATA;
        INFILE RAWDATA;
        INPUT       @17 VAR1 $5.
            #2      @3  VAR2 MMYYDD8.
                    @15 VAR3 5.2    ;
    IF VAR1 NE VAR2 OR VAR1 NE VAR3
        THEN VAR1=VAR3;
RUN;

PROC PRINT;
RUN;
```

Logic and statistical correctness

- When defining categories or groupings for SAS procedures or programs, make sure you have explicitly defined the categories as mutually exclusive and collectively exhaustive. When using the IF and SELECT statements, make sure all possibilities are stated and the entire range of values is covered. Be careful that the end points of the categories reflect assignment needs and do not overlap.

Be explicit about all conditions, even those that may be unusual. Use the ELSE statement with the IF statement. Use the OTHERWISE statement with the SELECT statement. Use the special range names - HIGH, LOW, OTHER - in the FORMAT procedure. In the following example you are distinguishing between gender and allowing for unusual responses.

```
DATA FEMALE MALE ERROR; SET;
        IF GENDER EQ 'F' THEN OUTPUT FEMALE;
        ELSE IF GENDER EQ 'M' THEN OUTPUT MALE;
        ELSE OUTPUT ERROR;
```

Program testing

- To ensure that each program is logically correct, review the program as soon as possible after it is written. Use a program "walk-through" inspection. If the program is large or quite complex, ask your supervisor to "walk through" the program in detail. If your supervisor needs assistance, a knowledgeable colleague can help.

- Test complex SAS procedures with small test files. Compare the SAS results with known results. Include bad, missing, and valid data in your test files. Test results should be documented and included in the work papers. Precisely explain the choice of options and modifiers.

Efficiency

- SAS programs may be more efficient if they temporarily limit the number of variable being processed. This is accomplished with the DROP or KEEP options attached to the SET statement. For example:

```
DATA; SET OLD(KEEP=VAR1);
        IF VAR1 GE 3 AND VAR1 LE 6;
```

- Avoid repeating complex and time-consuming calculations and file sorting. After making a calculation that you intend to use in a later program, save the calculation as a variable in the SAS data file. Avoid repeatedly sorting a file by saving the file in the order

that will be most useful in future programs. Indexing may also be efficient.

DATA ENTRY AND TRANSFER

The method you use to read data into a SAS data set will depend on the data source. In every case, you must verify that the data were entered correctly. This section explains how to implement those requirements with SAS.

Raw Data

In the data step, the INPUT statement is used for reading raw computer data. Most data from agencies, questionnaires, and data collection instruments are converted to SAS with the INPUT statement. SAS can read any file and record structure, no matter how complex. The following points facilitate the input of the data.

- To reliably enter data, you need a complete description of the data file, including a file structure, record layout, and data descriptions for at least the critical data elements.

- The reliability of the data entry program is critical. Compare the resulting values of the SAS data set with raw data to ensure that data conversion is reliable. If the data entry program is at all complex test it with a test file and document the results.

- Fixed-format data entry should be used in preference to free formatted and named data entry. Free-formatted data entry can lead to misread and skipped data. When free-formatted data entry is the only alternative, you should count the number of data elements read for each observation.

- Be careful with short records. When necessary, use the MISSOVER modifier on the INFILE statement. Use the LOSTCARD statement to synchronize data on multi-card input.

SAS Data

Some agencies maintain data in a SAS file format. SAS can read these files without error. However, be aware of the following points:

- You must have a complete description of the data file and data descriptions for the critical data elements. You must also get a copy of agency written SAS formats. Use the PROC CONTENTS procedure for a description of the file and how it was created.

- You must know the SAS version used to create the data file and how it was exported.

Data Formats From Other Statistical Software Packages

The file you need may be in the native format of another statistical software package. While SAS can read these files, you should keep the following points in mind:

- Be careful that the conversion maintains consistency between the value of the variable and its format. Be cautious of variable name conversions, truncation of long character variables, redefinition of missing data, changes in date conventions, and changes due to format inconsistencies. These potential inconsistencies are documented in the language and procedure manuals.

- Have a complete description of the agency's data file, including a file description and data descriptions of all critical data elements.

Data From Data Base Packages

Some agencies maintain data using data base packages, such as dBASE. SAS can read these files with the PROC DBF procedure for dBASE files and with special procedures for other data bases. To convert data from data base packages, you should be aware of the following points:

- Test and document the reliability of the data conversion programs. Sources of error may be mistakes in record segment pointers and incorrect SAS record structures.

- Have a complete description of the agency's data file, including a file description and data descriptions of all critical data elements.

SAS-Generated Data

SAS can generate data internally in data and procedure steps. Data generated in this manner must have the same level of reliability as data from external sources. A couple of points deserve special mention.

- GAO has approved the use of the SAS RANUNI random number generation function and subroutine to generate uniform random numbers for sample selection. When you need more than one stream of random numbers, use the subroutine rather than the function.

- Numeric data created from mathematical calculations are subject to rounding errors. The ROUND and FUZZ functions will eliminate rounding errors at specified rounding units.

On-Line SAS Data Entry

You can manually enter data directly into SAS data sets using the CARDS option, direct variable assignment, and full screen products, such as PROC FSEDIT. Data entered in this manner must have the same level of reliability as data provided by other methods. Also, you must be careful to create a solid audit trail, for example:

- The PROC FSEDIT procedure does not have a log file of the changes made to the data base. Use the PROC COMPARE procedure to compare values between the old and new data files, and review the output report.

Transferring and Moving SAS Data Sets

Transferring and moving SAS data sets is reliable, but specific methods must be used to retain the integrity of the files. Several methods are available. One method, for example:

- When copying files between the remotely linked computers running SAS, use the PROC UPLOAD procedure or the PROC DOWNLOAD procedure.

PROCESSING AND DOCUMENTING THE RESULTS

Documentation standards for SAS programs are similar to those for manual work papers; that is, SAS work papers should be complete, accurate, clear, neat, relevant, and understandable. The term "understandable" needs some clarification when dealing with SAS programs. SAS programs must be understandable to people with a general knowledge of SAS. However, SAS documentation need not be a complete explanation of SAS procedures and statements since this information is available in the SAS language manuals. Documentation of a SAS program helps ensure the quality of the product and provides the reviewer and referencer with an essential audit trail. The proper time to document your SAS program is during program development since documentation also helps you ensure that your program is working correctly.

Documentation includes the original SAS program, the SAS log file, and the output. The documentation can also include, but is not limited to, flow charts, comments and titles within the programs, and explanations of messages in the SAS log file.

The level of documentation suggested in this guide is based on Federal Information Processing Standard Publication 38, Guidelines for Documentation of Computer Programs and Automated Data Systems. Most of your work involves special-purpose programs which will not be used repetitively. Documentation of this type allows replication but is not designed for efficient long-term program maintenance.

The following suggestions help ensure that SAS programs will be developed with sufficient documentation and adequate review.

- You are encouraged to use interactive programming for program development, but work papers for the final programs should be run in the batch mode or with a clean interactive run.

- SAS programs are easily modified. To accurately document the running of a SAS program you must retain the log file in the work papers. SAS output reports must never be separated from their log file. Formal run books are not needed. On mainframe versions of SAS, the job control language messages should also be retained.

- The log file must be complete, including all source file statements. When calling external programs or macros with the %INCLUDE statement, use the SOURCE2 option so that all source file statements are listed on the SAS log file. When using macros, use the MPRINT and SYBOLGEN options to show the macro and to display how symbolic variables are resolved.

REFERENCING SAS WORK

Full referencing is required for all GAO reports and testimony, and material based on SAS is no exception.

To reference SAS-based information, the referencer must be able to determine that the work papers provide sufficient evidence. Referencers who are not familiar with SAS must seek assistance from someone who is. This person, called the technical reviewer, must also be independent of the assignment.

The following guidance will help the referencer and the technical reviewer reference SAS work papers:

- The referencer must determine if the work papers include 1) cross referenced system documentation; 2) program documentation including log files; 3) data documentation and validity tests; and 4) tests of complex programs.

- The work papers need sufficient instructions so that the referencer can duplicate the work. Instead of manually verifying every computation (which, in most cases, would be impossible), the referencer must ensure that computer program logic has been tested, the data have been checked for reliability, and the report item matches the supporting SAS output. In critical situations, the figures may be independently verified using another computer program.

DISPOSITION OF WORKPAPERS

Retention standards for SAS work papers are the same as those for manual work papers. The method of archiving data and programs will depend on the amount of data and the facilities available for archiving. Microcomputer data will most likely be archived on a floppy disk. Mainframe computer data will most likely be archived on magnetic tape.

You do not have to save every data file you create in the course of a job. The files that are archived will depend on how the data were created and what files would be time consuming to recreate. The following items should be considered in archiving SAS programs and data files:

- Use only SAS procedures, such as the PROC COPY procedure, and DATA steps to make backup copies of SAS data files. Never use operating system utilities on these files.

- When saving data files, remember to include formats, windows, and indices.

BIBLIOGRAPHY

The SAS Institute publishes an extensive list of SAS software guides which are customized for various operating systems and computer environments. A few of the over 200 SAS publications are listed below.

Version 6: MS-DOS and PC-DOS environment

SAS Language Guide for Personal Computers, Release 6.03 Edition

SAS Procedures Guide, Release 6.03 Edition

SAS/STAT User's Guide, Release 6.03 Edition

SAS/GRAPH User's Guide, Release 6.03 Edition

SAS/FSP User's Guide, Release 6.03 Edition

Version 6: AOS/VS, CMS, MVS, OS/2, PRIMOS, and VMS environments

SAS Language: Reference Version 6

SAS/STAT User's Guide, Version 6

SAS/GRAPH Software Reference, Version 6, Volumes 1 and 2

SAS/FSP Software Usage and Reference, Version 6

SAS Companion for the MVS Environment, Version 6

You will find additional guidance in these publications:

Federal Information Processing Standards Publications, Department of Commerce, National Bureau of Standards, Washington, D.C.: 1970-1991.

Using Micro Computers in GAO Audits: Improving Quality and Productivity, Technical Guideline 1, General Accounting Office, Information Management and Technology Division, Washington, D.C.: 1986.

Assessing the Reliability of Computer-Processed Data, GAO/OP-8.1.3, General Accounting Office, Information Management and Technology Division, Washington, D.C.: 1990

Project Manual, General Accounting Office, chapter 10.1, Washington, D.C.: 1986.

Government Auditing Standards, 1988 Revision, General Accounting Office, Washington, D.C.: 1988