

PORTABLE SOPs FOR CONTRACT SAS® Programmers and Consultants

Bill Calvert
Service Research, Inc.

ABSTRACT:

As a contract SAS programmer/consultant one can encounter a wide range of conditions. From one place to the next, there may be very little similarity in the standards and fashions of getting the work done. To go beyond technical competence with SAS software to attain functional expertise, consider adopting a set of portable Standard Operating Procedures (SOPs). SOPs are good for everyone, but they are critical for temporary workers. Some basic SOPs and related data management concepts are presented considering the needs of temporary workers (and those who work with contract help). Strategies for implementation of portable SOPs are described to help overcome some common problems temporary workers may encounter. Additionally, the need to maintain flexibility in your SOPs and adjust to the local resource profile is discussed. With the ideas and suggestions presented here, you will be better equipped to promote improved data management wherever you find yourself.

INTRODUCTION:

The terms used to refer to contracted SAS programmers vary widely. 'Consultant' is one of the terms that is widely used in this country. This term is used in this paper for convenience to refer to workers who are hired temporarily for a limited time or for project specific tasks. Consultants generally have three or more years of experience doing the specific kind of work for which they are contracted and have appropriate skills at the intermediate to advanced level.

The consulting context can be very different from client to client (and consultant to consultant). Clients frequently hire consultants because they have problems such as too much work, not enough time, or management situations that lead to the above mentioned problems. Consultants often confront the problems of a new assignment with little support or knowledge of the situation. 'Consultants are experts, or we wouldn't hire them.' Sometimes consultants work for a placement firm or broker. The placement firm makes it easier to get work, but often does little to improve the working arrangement. The uncertainty of consulting can generate severe pressures that strain even the best of people.

The strain is greatest with a new client. You need to make a good impression from the beginning. There are a number of details to attend to the first day, when you start working with a new client. Attention to important detail from the beginning, usually makes a good impression and establishes the working relationship on a firm footing.

There are administrative details:

- who authorizes payment? ...
- who assigns the work? ...
- how are hours (time) recorded? ...
- who handles administrative matters, etc.?

There are functional matters to attend to:

- security procedures...
- logon procedures...
- where's the bathroom...

There are task specific details to learn:

- where are the specifications...
- where is the data...

And finally there are the more global professional issues. Going beyond the first day, the professional issues become more important. You were hired for your professional capabilities.

For most information workers who are using SAS, data management is or should be a large component of the professional skill set. The focus of this paper is on data management and related issues of programming style. The intent is to provide just a little help to consultants (and possibly their employers) so they can attain the highest level of professional proficiency in data management. Over time these skills and concepts will help you demonstrate your worth and keep the clients coming back to you.

Strategies for improved data management require information and dogged effort. Information about data management can be found scattered in a wide range of sources, such as government publications, computer science texts and journals, clinical trials and pharmaceutical industry journals, and various software vendor publications (e.g. SAS® User Group Proceedings). If one is patient, it is possible to piece together the experiences of others to solve the problems at hand. It is more likely that one will learn about data management from experience and/or association with other data management professionals.

To improve data management, the consultant can play a special role by setting a good example, providing pertinent articles and publications, and mentoring or advising the numerous people in their working environments.

Part of the mentoring process is in knowing what should be, and why it should be. It is rarely possible to get changes made in an organization of any size

without the benefit of social graces. It is also true that one rarely achieves anything of value without striving for well thought-out goals. The categories of data management procedures presented here are proposed as a substantial start for a portable set of SOPs.

SOP CHECKLIST:

The twelve categories listed below provide representative examples and ideas of what might be Standard Operating Procedures in most circumstances. Specific procedures must be matched to the local working environment and tested in practice. As a consultant it is important to determine if there are existing standards in these categories. If they exist, try to fit in and possibly improve on them. If they don't exist, try to establish them, if only in your own work.

The areas for SOP development are listed in approximate order of importance:

1. Management procedures

One key element of good data management is the operational management of the information processing unit or project team. Sufficient time must be allowed to observe standards and maintain the records. Often the short term demands over-run the long term necessities. This should be resisted. The payoff for good data management is fewer major mistakes and a consistently higher level of quality. A secondary benefit is the ability to absorb staff changes and turnover more easily. Associated with this is the tendency for staff to stay with a good, well-run operation.

A consultant is often in a position to advise management with substantial credibility. As an outsider, personal gain is less likely to bias the recommendations one makes.

2. Naming conventions

Naming conventions are a form of internal documentation. They provide a structure for assigning names that allows members of the team to find and use the work of another more easily. File names depend on the operating system, but usually include information about the account/project, the category of effort, and the sequence number and/or date. The author's initials or other identifier may also be included. Within data files, variable names are controlled by a list or dictionary for those that commonly occur across projects. Other variable names are usually constructed according to an established pattern. For example, use 6 characters to provide some clue to the content or origin of the variable. The remaining 2 characters can be used for suffix or prefix designations that indicate changes in the content from the original (source) variable. Such changes may include collapse of categories, other recoding or corrections and adjustments.

For the consultant, it is good to come into a situation with good naming conventions. It is easier to quickly learn enough to become productive. For the temporary worker, it is also good to leave behind something that is useful and of lasting value. Good names help make systems and programs useful to others.

3. Consistent programming style

Specifications for programming style improve the quality of programs and facilitate reuse and/or maintenance of code by others on the team. Style prescriptions may be for good things to do, or more often things to avoid. For example, we encourage the use of formats for recoding and discourage the use of a long series of IF statements. Programming standards can be flexible and enforced with common sense. A code formatter can also be used to keep the indentation and spacing of code uniform and easy to read.

4. Run validation procedures

Each computer run should be validated. The process may consist of:

- A. Check that there are no error messages in the SAS log.
- B. Check that the number of observations for each data set in each segment are reasonable. Be able to account for all of the original observations.
- C. Review warnings and notes in the log, be sure to understand why they occur and review the information for reasonableness.
- D. In some environments (mainframe) check for system error messages (JCL) that may be a clue to potential problems.

5. Data verification procedures

All data that is to be analyzed should be identical to the original (usually the case report forms) or map to the original data in a predictable way (as in the case of recoded data). Manual and automated means are used to confirm that the original representation of the data has not been corrupted at any point in the analysis process. Machine-readable (digital) data such as that supplied by remote sites, government agencies or insurance companies is compared to the documentation supplied from the source and any available reports or summary statistics. Verification and validation processes are often intertwined.

6. Data validation procedures

All data used in an analysis should be evaluated for reasonableness. Look for missing values and check for duplicate records. Examination of univariate statistics on each variable are helpful to observe the range and distribution of the data. Multi-variable logical checks may also be used. Additional content specific validation procedures may be defined by the principal investigator.

7. Keep extract programs

Data from large studies or claims databases used in post-market research often is obtained on tape. At a minimum, the data must be moved from tape to disk. Most often, only the data required to meet current research needs is written on disk. This extraction process should consist of one program and its documentation. Extraction of data multiple times from the same source and subsequent recombination introduces greater risk of error. Sometimes system limitations may force the data analyst to take the risk, such as in the case of very large data sets stored in a large and/or complex file structure.

8. Keep update/recode programs

Typically, working data sets are created for specific study requirements. There should be one and only one program used to generate a given working data set. Even for somewhat large data sets it is cost effective to pass the data again to make revisions and corrections. Multiple programs should be avoided; pieces can be lost, documentation and code segments can be overlooked and execution of pieces of the program series may be wasted due to a failure to understand the entire series.

9. Keep analysis programs

Analysis programs are highly variable and change according to the thread of investigation. For example, adverse events for one drug result in two outcomes, symptoms (only) or (symptoms leading to) hospitalization. The analysis thread for symptoms would have program names like SYMPT1, SYMPT2, etc. For the hospitalization thread names might be, HOSP1, HOSP2, etc. It is most useful if each thread of investigation has a name with a sequence number suffix. The analysis can then be recreated by the numbers. As with other programs the log should be kept for each run that is completed satisfactorily. Keep the output with the log too, if it isn't too large.

10. Problem resolution procedures

The individual or individuals responsible for problem resolution should be designated. In some situations a problem log is established and standards or information requirements are established to facilitate efficient investigation and resolution of problems. Logging problems and their resolution helps the team learn by experience.

For the consultant, being paid by the hour, under pressure to perform, unresolved problems make it a bad day. Being able to enlist the help of appropriate individuals, and having the support afforded by an established process can help solve the problem faster and minimize the loss of precious time.

11. Data security and disaster recovery procedures

Data management is about data security. Having a well-organized data management operation is the first step to eliminate or at least minimize the effects of entropy. This may include regular backups and virus checks, for example. Protection from destructive persons, such as disgruntled employees, can usually be attained by complying with standards for handling confidential data such as locked/secured offices, password protection on files and computer accounts, and frequent review or update of computer security provisions. Higher levels of security, such as those required for the prevention of espionage, are probably not necessary for most operations.

Disaster recovery requires provision for off-site backup of computer systems and data and a plan for the implementation and direction of recovery operations. Major computing centers usually have a disaster recovery plan. With the shift to the PC environment, individuals need to take increasing responsibility for disaster recovery planning, not only for research data, but for the essential office automation functions as well. There are horror stories about companies that cease to exist because of a failure to adequately plan to

recover from ordinary disasters like earthquake, fire and theft. Insurance may help, but money doesn't replace lost data or the time it takes to replace facilities.

12. Maintain documentation

Documentation is essential to any well-founded data management operation. All of the items previously mentioned rely in some part on the existence of usable documentation. Documentation is essentially a means of communication that helps multiple people work together effectively through time and space. As a component of data management standards, it serves a dual function: It prescribes what should be done and provides a mechanism for obtaining the information necessary to monitor whether the prescribed actions are being done. A documentation system may consist of several elements:

A. Guidelines - guidelines are maintained for: what documentation is to be kept and the required content and format, names and naming conventions, issues of programming style and various data management operations.

B. Logs - sequential records or diaries of project activity may be tracked in a project log that contains information about the project's data resources, personnel, hardcopy and machine-readable information. It serves as a guide to project resources and supports essential administrative functions. Another very useful log is a trouble log or problem log; it helps avoid making the same mistake twice, or at least minimizes the time required to recover from a problem if it has been encountered before. Other logs may be used depending on the complexity of operations.

C. Program Documentation - programs are documented internally (see programming style) and externally. External documentation at a

minimum consists of an entry in the project log.

D. Data Documentation - data are of two main types - original data, and working data sets; each has its own documentation requirements. Original data is usually provided with documentation generated by the 'source'. If the data originated locally, minimum documentation should be produced to include file specifications and a description of the file structure. Working data sets are generally documented by the programs used to produce them, (Comments are important here) and through reference to the original data source documentation.

E. Analysis Documentation - analyses are documented by putting the SAS log and output of properly completed runs chronologically in a binder with any applicable notes. This facilitates repetition of analytical threads as necessary.

F. Problem Reports - problem resolution strategies are formally recorded to suggest the best approach to gaining a solution to data management difficulties and concerns. Sometimes these may be recorded in a log, as described above. In more complex environments, or where the development of more general solutions are desired, the nature of problem reports may exceed the sequential confines of the log format. For long running projects, for which issues of responsibility and accountability are recorded or updated as a matter of policy, a combination of a summary document with an historical log can be useful.

G. Process Documentation - Don't forget to record the big picture. Often the details are easier to handle if the overall process is described in a general way so the pieces can

be put together. Process documentation can also be very helpful for troubleshooting conceptual problems with the system.

Documentation can make an otherwise difficult assignment easier for the temporary worker. Reading and reviewing documentation for a few hours can often save days of trial and error explorations. Consultants sometimes get a bad name when they don't return the favor of good documentation to those who come behind them.

CONCLUSION:

Standard Operating Procedures (SOPs) for data management are written guidelines which help establish a working environment conducive to higher data quality. Generally applicable data management procedures and practices have been presented here. The categories of procedures described as potential elements of SOPs for various types of data management operations are provided as guidelines for consultants. The rationale and potential contribution of each element as it relates to higher data quality is discussed. Good data management broadly consists of establishing and maintaining communication. This communication must work between and among programmers, statisticians, writers, and other information workers. Communication is especially important to temporary workers who may be part of the team for only a short time. If good verbal and written communication can be established and maintained, the resulting coordination and understanding throughout the information processing team produces better data quality.

The people and situations will differ (especially for the consultant), consequently the optimal operational procedures will differ. Standard Operating Procedures become Proper Operating Procedures when they work. This transition requires 'people' skills from consultants and all other members of the team. SOPs alone will not assure better data quality.

Engineering better data quality requires motivated competent people who know what the SOPs are, how they are applied, and why they exist. Technology and situations change, so must SOPs. Periodically, the data team will need to be re-trained to maintain their effectiveness as the SOPs grow and adapt. As long as SOPs are flexible and 'alive' to the present, they can be the basis of effective communication and functional understanding for good data management operations.

Consultants have special value in keeping data management practice at optimal levels. Due to their broad experience with systems, organizations and people, the consultant can often set the example of the best way to do it, even in the face of severe production pressures. To fulfill this role, the consultant typically needs to be ready for anything, work well under any circumstances, smile when kicked in the shin (or some place worse), and maintain the highest professional standards all the time. Having a set of portable standards of operation developed from the concepts presented here can help the consultant consistently meet the high expectations of clients while working on data intensive projects.

REFERENCES:

1. Calvert, Rao; "A SAS® Code Formatter". SAS Users Group International Conference Proceedings: SUGI 17, Cary, NC; SAS Institute Inc. pp.158-162 April 1992.
2. Calvert, Floor; "Engineering Better Data Management." Invited presentation at Pharmasug '95 discussing standard operating procedures in pharmaceutical research organizations.
3. Calvert, Swartz; "Using features of SAS software to improve your documentation". NESUG Conference Proceedings: October 1995.

◀ ■ ▶

SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.