Monitoring and Improving Quality of Data Handling

The purpose of this document is to:

(a) Maximise the quality of the research process once the question has been formulated and the study designed.
(b) Ensure easy access to study information.
(c) Allow continuity if study co-ordinator is unavailable.

This document is **NOT** about which data analyses should be conducted, or how this should be approached.

I. BEFORE DATA COLLECTION

Before data collection, a protocol should be developed. Issues that need to be addressed in this protocol are:

- developing clear hypotheses/ rationales for the research
- selecting measures, including, if possible, getting the original measures and manual
- developing measures: reliability and validity
- a clear idea of what analyses will be conducted, and what the potential results could demonstrate
- power calculations and random sampling

II. COLLECTED DATA

The following will assume that a reasonable hypothesis has been formulated, and enough (quantitative) data collected to provide sufficient statistical power to tackle the hypothesis. From this stage to the interpretation of any analyses, it will (usually) be necessary to go through the following stages of data handling:

- data coding
- data entry
- data checking
- data cleaning
- data description
- manipulation of the data set
- statistical analyses
- drawing inferences and presenting conclusions
- archiving

DATA CODING

All (questionnaire) data should be coded prior to data entry. To facilitate this, a coding scheme should be agreed, which includes:

(i) directions for coding both "expected" responses, as well as unorthodox responses (e.g. where two categories have been selected when the questionnaire asks for one; where responses are marked between two offered categories).
(ii) the principles behind these directions, to guide future decisions that need to be made about coding.
(iii) how “missing” data are to be treated, and any distinctions felt necessary for different types of missing data.

When dealing with large or complex data sets, a coding manual should be devised, which includes explicit directions for dealing with coding. This manual should be updated on the basis of any problems that emerge from subsequent coding.

The location of any "awkward" data should be noted as coding progresses, and a policy decided after group discussion: to ensure reliability in coding, no decisions should be taken by one individual on an "ad hoc" basis. Reliability of coding should usually be statistically assessed using kappas.

All coding should be checked by a second coder to ensure reliability. There are two kinds of reliability:
(i) whether or not an event occurred
(ii) what kind of event it was.
Any disagreements can be referred to the mutually agreed coding manual.

DATA ENTRY

To reduce the possibility of confusion and errors later, all entered data should have appropriate variable and value labels. With SPSS-Win it is also possible to enter some descriptive text for each variable, to elaborate on the eight-characters allowed for variable labels.

To minimise mistakes when entering data, data should be entered twice: two different individuals should do this if possible to reduce the chances of systematic biases. The second entry should enter data in an identical way, using the same variable names, but with one character consistently changed (e.g. the variables "date", "anxiety" and "language" could become "datez", "anxietyz" and "languagz"). Care should be taken that each case has a unique identification number, which is common to both versions of the data entry.

If the dataset is large, approximately 10 percent should be entered twice to check on reliability. After the two separate data entries have been compared, a decision should be made about the acceptability of the errors uncovered, on the basis of a decision about the extent to which the rate and the type of errors would substantially affect any inferences drawn.

DATA CHECKING

Once entered, some preliminary checks must be made to ensure no "impossible" scores have been entered. A simple frequency printout for all variables is often sufficient. Similar checks should be made after the construction of any aggregated scales.

If data has been entered twice as suggested above, the first stage of data checking is fairly straightforward. The two files can then be merged, using the unique identification number. A simple subtraction of the two versions of each variable (e.g "date" and "datez") can then be used to select out those variables which are not equal (and therefore have been entered with different values).

Alternative procedures include looking at frequencies, ranges and boxplots to identify outliers.

DATA CLEANING

Where clear errors of data entry have been identified, it should not present too many problems to alter any erroneous values with the correct ones. Where errors are based on differences of interpretation, this probably reflects a deficiency in data coding, and a policy for dealing with such differences should be agreed and recorded.

DATA DESCRIPTION

The first step in any analysis is to produce tables of descriptive data, to enable the researcher to 'eyeball' the data.

MANIPULATION OF THE DATA SET

Data are often altered before analysis (e.g. variables are summed to create a composite scale or recoded to collapse several categories into fewer categories). This should **always** be done on a secondary data file, not the original systems file in which the data were entered.
The most important consideration at this stage is **never** to lose data. New variables should be created from data transformations, **never** from alterations to existing variables. All researchers concerned should make the decisions about how to treat missing data, especially for newly created scales.

Variables should be given sufficient variable and value labels to enable a researcher unfamiliar with the dataset to confidently work with them. Copies of these should be kept in the appropriate study file.

Variable distributions should be checked and normalised. When variables have been transformed, it should be noted that the direction of scales, and therefore observed relationships, may be reversed. A note should be kept of whether these changes in direction occur.

Attempts should be made to continually refer back to the descriptive statistics derived from the original data set. This will ensure that no gross errors of interpretation are made, and will facilitate understanding of the data.

The overall aim should be to develop a double-checking culture at every stage.

APPROPRIATE ANALYSES

We need to justify any analyses we have conducted. Even when a statistician's advice has been sought, we should ourselves still be confident that **we** understand and could justify ourselves (e.g. to a conference audience), as the ultimate responsibility for ensuring that the appropriate analyses have been conducted rests with us.

Other means of ensuring we have used the most appropriate analysis are to check with other people either within or outside the group, and statistical textbooks. Again, the responsibility for the use this advice is put to is the individual researchers themselves, who should ensure they understand what the purpose, drawbacks and limitations of each statistical procedure used are. Training may be necessary here.

A plan of analysis should be constructed, including the what, why and how of the analysis. This should be constructed at the stage of designing the study, and agreed BEFORE beginning the analysis.

Where complex analyses are appropriate (e.g. ANOVAs with many different levels), simpler analyses should be conducted first, to aid an understanding of what is occurring. The more complex analyses can then be "built up to": univariate analyses precede multivariate analyses.

Where possible, attempts should be made to replicate statistical results using different statistical techniques. This should increase our confidence in the robustness of our findings. Always check with means, frequencies, etc, to ensure the results "make sense" in terms of the raw data.

A thorough record of all analyses should be maintained, so queries at a later date can be handled with a minimum of trouble. This record should be as "transparent" as possible, to allow people other than the individual who carried out the analysis to adequately understand it. When individuals leave the group, a "handing over" of projects should include a familiarisation of those concerned with the layout of an analysis, preferably supplemented by written information.

III. CARING FOR DATA

Data or data analysis files should not be stored in the same directory as software files, as this creates problems when the data is re-installed or when software is upgraded.

When any data process is being conducted, attention should be paid to regularly backing up all work. Regular hard drive back-ups should be made as work is progressing. Back-ups should also be made onto floppy disks periodically, in case of computer failure. Use of different, clearly labelled back-up disks for each large data file should be considered to avoid accidentally overwriting data files.

Data should NEVER be deleted because of lack of computer disk space. It should always be copied to floppy disks and kept securely. A LABELLED copy should also be kept in a filing cabinet drawer.

IV. DOCUMENTATION/ KEEPING A RECORD

A log book of analyses should be kept. This should include a record of what analyses were conducted, when they were conducted, and where the computer records of these analyses are stored.

Every study should have a file kept up to date by the study co-ordinator. It should be clearly labelled and include a copy of the protocol, study materials, coding manuals, progress summaries, minutes of meetings, summaries of computer files, detailed procedural guidelines, etc. A disk should be included containing all study files. Written and computer documents should be regularly updated. The purpose of the study file is to inform those unfamiliar with the study in sufficient depth to be able to take over the running of the study without needing extra information or to obtain information at short notice when the study co-ordinator is not available.