Share this article:

How Clean Is Your Data? Understanding Survey Editing and Cleaning Options. (January 2009)

These days online surveys and other electronic forms of data collection seem to be all we talk about. Electronic data collection methods are usually designed to prevent certain types of data errors. Less forgiving are paper surveys which remain a viable methodology for certain research applications. Regardless of the data collection method employed, it is essential to pay attention to data on an individual case basis. Individual data records become the foundation of all further data analysis.

We all know that respondents don't always complete surveys in the manner we expect. Sometimes the data we are working with remind us that, despite our best efforts, survey research can be a less-than-perfect process. We often need to take steps to balance the integrity of the data, while being careful not to alter responses in a way which may introduce unintended biases. In order to meet this objective, we must determine which tasks are best suited to humans, and which should be left for computers to handle.

- **Manual survey editing**

  Editing, a manual review of paper surveys prior to and in preparation for data entry can be a tedious but important step in the survey process. This is especially important for self-administered surveys that are often filled out in nonconforming ways. Prior to editing, check-in procedures should be followed to verify counts of documents received from field sources and assign unique ID numbers and batch control codes. These steps enable quick location of individual documents as needed during subsequent data processing.

  In the editing phase itself, the primary goal is to prepare the surveys for data entry. Development of a clear and thorough editing guide (sometimes called an "edit master") is an excellent reference tool for both the researcher and the data processing group. Unfortunately, it is a step often skipped in the survey process. Ideally, this should be prepared by the researcher to instruct the survey handlers what types of cleaning should be done prior to data entry. Whether or not you provide formal edit instructions, be sure to clarify any ambiguities and always express any special concerns you may have to the survey processing group.

  Examples of common editing practices include:

  - Invalid surveys should be identified and weeded out -- these are often most easily spotted manually. Sometimes respondents send back written explanations of why they were not able to complete the questionnaire. It may be partially completed, but the explanation may provide important insight to whether it should be included in the final dataset. Certain types of complicated survey forms, i.e., diaries, can be visually inspected to get an overall sense of whether the respondent understood and completed the tasks involved in a usable fashion.
  - Vague responses stated in terms like "most of the time" when asked for a numeric response should be resolved by rules applied in the editing process. Some common practices, such as entering midpoints for ranges like "10-20%," can be handled in either the editing or data entry stages.
  - Stray comments written in the margins of a survey may have important consequences on the interpretation of responses and should be dealt with during the editing review. Sometimes action items, such as "please call me about my bill" or "change my address," need to be immediately communicated back to the sponsoring company. Over the years we've received a wide variety of non-survey related correspondence including such things as investment checks and medical records sent along with survey responses!

The focus of manual editing is on human interpretation within the context of the individual respondent's pattern of answers. Editing should be limited to tasks which require thought, not repetitive actions which are best left to the computer to handle. There is also some overlap of which issues are best dealt with before versus during data entry. As a general rule, we instruct data entry operators to "flag" problem survey documents. This allows the opportunity for later checking and determination of how best to deal with unforeseen issues on a case-by-case basis.

- **Computer-based data cleaning**

  Complicated skip patterns and instructions are usually handled in computer processing following electronic data collection or data entry. Computer cleaning instructions can be accurately and objectively applied to all surveys based on a few simple rules.

  The most common aspects of data cleaning are:

  - Checking quota fields to check data against field reports. It's always better to catch any discrepancies early -- don't wait until someone notices an inconsitency during a presentation!
  - Cleaning skip patterns to ensure consistency of responses. Consider whether specific cleaning should be done in a forward manner (i.e., blank out a response based on a previous answer), or back-cleaned (i.e., change a previous response based on a later answer).
  - Validating numeric fields is usually done automatically for electronic surveys, but will need to be done after the fact for for paper surveys. A common practice is to check fields which should add to a certain sum (i.e., 100%), and "force" them to add to a desired sum by reproportioning.
  - Reviewing "other" responses. For fields where respondents are asked to specify other answers to an existing list, the captured responses should be reviewed to determine whether new categories should be created or responses "upcoded" to prelisted categories.
  - Reviewing outliers based on initial frequency distributions. Numeric responses should be examined for outliers which can either be capped or removed to reflect acceptable minimum or maximum values.
  - Obtaining ID numbers of errant cases. Sometimes called source document cleaning, this step involves returning to the original survey forms to help clarify a respondent's intent.
  - Recoding data values from discrete numbers to range categories may facilitate later tabulation and analysis, or modifying responses from arbitrary values used to make data entry easier.
  - Generation of calculated fields may include simple calculations such as creating respondent age from date of birth, or other fields derived from survey responses (i.e., total spending, length of interview, etc.).
  - Weighting data is an option to balance the sample to reflect a known population or project to a larger universe.

Don't overcomplicate the cleaning and editing process -- just be sure it is handled in a complete and objective manner, to provide you with the confidence needed for subsequent data analysis.