TOPIC 10

Data preparation

Preview

Introduction

Data preparation entails editing, coding and tabulating. The information in this topic often refers to large-scale projects, since errors in data preparation generally cause their most dramatic inaccuracies in such projects. By the same token, however, when a relatively small sample is involved, problems at this stage can be just as unfortunate. If a questionnaire has to be discarded, for example, you might lose a fair percentage of the data that was expected.

In addition to drawing your attention to ways of simplifying the preparation work for your own project, some knowledge of editing and coding procedures, in particular—and the problems associated with them—can go a long way toward contributing to your credibility when dealing with a research agency. Knowing the right questions, and understanding the answers, will not only be reassuring for you, but will give you, as well, another addition to your arsenal of convincing arguments when the time comes to justify the research findings, and possibly your decision to do the research in the first place.

Data preparation is the most 'mechanical' of the stages of a research project. Because of this, it is impossible to estimate the seriousness of mistakes that might arise at this stage. This fact should underlie the study of this topic.

Editing

The initial stage of editing is to examine the collected raw data in order to be sure that it is accurate. This is often done 'in the field' by a field editor. The completed questionnaire is checked for overall accuracy, completeness and general usability. Interviewers do some of this themselves too, finishing incomplete sentences and expanding abbreviations that only they would understand. Field editing is best done as soon as possible after the interviewing has taken place. 'Field editing' and 'central editing' together are often called 'the initial screening process'.

Central editing is done in the central office where the remainder of the editing takes place. This is best done by a single editor. If that is not possible, the best way to divide the work is not by just dividing up the stack of questionnaires among two or three people, but rather by having each person edit specific questions (e.g. one editor does the first five, another the second five, and another the last two and the demographic section).

This is important because results will eventually be presented by an analysis and interpretation of each question (among other things). By dividing up the questions, rather than questionnaires, the job will be done in a consistent manner from the outset. This can perhaps be considered an overcautious approach, but if you were paying for the research, it would be reassuring to know it was being handled this way. One couldn't help but feel—probably quite rightly—that the other stages of the research were also going to be well handled. (Though, in this case, it makes sense for the research agency too; the ‘assembly line’ system is more productive than dividing up the questionnaires.)

The editing that takes place at this stage has been well summed up in Green and Tull’s *Research for Marketing Decisions*:

> 1 *Legibility of entries*. Obviously the data must be legible in order to be used. If an entry cannot be deciphered, and clarification of it cannot be obtained from the interviewer, it is sometimes possible to infer what it should be from other data on the form. In cases where any real doubt exists about the meaning of the entry, however, it should not be used.
>
> 2 *Completeness of entries*. On a fully structured collection form, the absence of an entry is ambiguous. It may mean that the interviewer failed to attempt to obtain the data, that the respondent could not or would not provide it, or that there was a failure to record collected data. If the omission was the result of the interviewer’s not recording the data, prompt questioning of the interviewer may provide the missing entry. If the omission was the result of either of the first two possible causes, it is still desirable to know which was the case.
>
> 3 *Consistency of entries*. As is the case with two watches that show different times, an entry that is inconsistent with another raises the question of which is correct. (If a respondent family is indicated as being a ‘non-user’ of cooking sherry, for example, and a later entry indicates that they purchased six bottles during the past month, an obvious question arises as to which is correct.) Again, such discrepancies should be cleared up by questioning of the interviewer, if it is possible to do so. When they cannot be resolved, discarding both entries is usually the wisest course of action.
>
> 4 *Accuracy of entries*. An editor should keep an eye out for any indications of inaccuracies of the data. Of particular importance is the detecting of any repetitive response patterns in the reports of individual interviewers. Such patterns may well be indicative of systematic interviewer bias or dishonesty.
>
> (Green & Tull 1978, pp. 239–40)

In regard to accuracy of entries, it takes experience, but a good editor can discover cheating by having developed a sensitivity to common, telltale patterns of responses from the questionnaires of a particular interviewer. Since that single interviewer might well be responsible for a fair percentage of the questionnaires in a particular area, the potential for biased results is extreme. (The temptation to cheat is present, in particular, in the administration of omnibus surveys, where a door-to-door interviewer must generally obtain ten interviews on a weekend, each of which—depending upon the respondent’s usage or awareness patterns of the subject matter of the questionnaire—can take 30 to 45 minutes.)

Another problem here might perhaps be called *attempt-at-humour entries*. These are not as easy to identify as might be assumed. Weiers gives a good example of this in *Marketing Research*:

> *Is the respondent a comedian?* While most individuals take seriously their role as respondent, others do not. For example, some may indicate ‘star shortstop for the New York Yankees’ as an occupation, or ‘leaping tall buildings’ as a hobby. Usually, the comedian is fairly easy to spot, and his or her questionnaire can safely be thrown out. However, some responses that seem at first to be nonsensical may actually be the result of a serious effort by the respondent. For example, consider one of the projective questions …
>
> Person A: ‘The Defensive Driving Course is being offered at the plant next month. Are you going to sign up’?

> Person B: 'No, I … … … … … … … … … … .'
>
> One response, 'No, I gave at the office', was initially thought to have come from a frustrated comic. In a subsequent focus group interview, however, the respondent volunteered that his answer was not intended to be comical, but reflected his personal view that taking the course would be a 'donation' of his time and would not offer him appreciable self-improvement benefits. While this was admittedly a minority view, it represented the individual's feelings toward the course. Without the added insight offered by the focus group setting, it would have been discarded as just one more prankster response.
>
> (Weiers 1984, p. 382)

The purpose of editing, simply put, is to prepare the questionnaires for coding and tabulating. Perhaps hundreds of people have been asked the same questions. The replies—for the open-ended questions—tend to adopt a variety of phrasings, even though the meanings of many of them may well be the same. Therefore, before lumping them together under the same heading in order to count them, it is necessary to ensure that they do, in fact, mean the same thing. Then they can be categorised for coding. (Texts differ as to whether or not this categorising is properly part of 'editing' or 'coding'. Keep this in mind if you wish to do any further reading on the subject.)

Coding

Sekaran remarks that 'the data have to be categorized under broad headings, and errors in categorization may result from misinterpretation' (1992, p. 276).

Luck and Rubin (1987) give an excellent list of the principles that should be followed in setting categories to facilitate coding:

1 *Convenient number of categories*. The number of categories should be substantial enough so that differences in the data can be revealed, yet not so few as to hide important information. At the initial stages of data analysis, class intervals should be rather narrow so that significant tendencies in the data are not lost within the intervals.

2 *Similar responses within categories* (intraclass homogeneity). Responses classified in a particular category should be similar with respect to the characteristic being studied.

3 *Differences of responses between categories* (interclass heterogeneity). Given the characteristic under study, the differences in responses between categories should be dissimilar enough to reveal substantial differences between the responses.

4 *Mutually exclusive categories.* Categories should not be overlapping. They should be constructed so that any response can be placed in only one category.

5 *Exhaustive categories.* The construction of categories should provide that all responses be included in a category. This may include categories, where appropriate, of 'Don't know' and 'No answer' as responses.

6 *Avoid open-ended class intervals.* Open-ended class intervals should not be used, because lack of specified interval limits obscures the extremes of distribution and precludes computing the average value of the observations in such intervals.

7 *Class interval of the same width.* Class intervals should be of the same width, where possible, rather than of varying widths. Disregard of this principle may lead to situations where the intervals lack a consistent spread. However, the unequal breadth of categories may be acceptable when categories apparently contain relatively small proportions of the total response characteristic, and finer categories may be uninformative.

> 8 *Midpoints of class intervals.* If the respondents are likely to have broadly estimated the answers they gave, stating them in round numbers, the class intervals should be designed so that major round numbers fall at the midpoints of the class intervals. For instance, income earners may be prone to report their incomes to the nearest hundred dollars. An income category of '$4000 to $4099' would conceal the fact that those classified in this bracket had tended to give their lower limit; thus '$3950 to $4049' would be a better interval to use, because it places the central tendency in its middle.
>
> (Luck & Rubin 1987, p. 348)

Examples of 'class intervals' would be the 'age in years' or the 'number of years in this organization' categories in Sekaran's questionnaire (1992, p. 278).

Two useful terms regarding this subject are 'precoding' and 'postcoding'. To 'precode' is simply to set up numbers or letters before each of the answer choices on a questionnaire, as in the Sekaran example, or to choose a scaling system where the respondents' circles or 'x's would be equally simple to record. 'Postcoding' is more difficult. It requires setting up categories *after* the questionnaires have been completed.

Here is a good example of postcoding from Weiers.

> **Question**
>
> 'When I see a Porsche automobile, it makes me think of … '
>
> **Responses**
>
> 1 'how much fun I'd have if I owned one'.
>
> 2 'how unfair our social system is that only a few people have enough money to afford a car like that'.
>
> 3 'racing'.
>
> 4 'small cars and how dangerous they are'.
>
> 5 'the U.S. balance of payments'.
>
> 6 'what a ball it would be to drive'.
>
> 7 'my brother, because he's a sports car nut'.
>
> 8 'how much the insurance must cost to own one'.
>
> 9 'rich people'.
>
> 10 'how well I like my Datsun 280ZX'.
>
> 11 'all those Pittsburgh steelworkers who are laid off'.
>
> 12 'what a pain they must be to work on'.
>
> 13 'my wife fainting if I drove one home'.
>
> 14 'going to a movie'.
>
> 15 'sticking out my thumb and hitching a ride'.
>
> After selecting categories that account for all responses, we might postcode the preceding responses as follows:
>
> Desire to drive or own one, responses 1, 6, 13, 15
>
> Negative social/economic comment, responses 2, 5, 9, 11
>
> Undesirability or ownership disadvantage, responses 4, 8, 10, 12
>
> Other, responses 3, 7
>
> Irrelevant or comic, response 14

(Weiers 1984, pp. 388–9)

Responses 9 and 13 were considered, quite rightly, to be a bit of a problem. They were placed in their categories tentatively. Afterwards, discussion with respondent number 9 showed that he was referring to the fact that he would like to be wealthy, and saw the Porsche as a status symbol. This cleared up the question of whether or not of not the word ‘rich’ had been meant to convey anything derogatory.

Response 13 did not appear to the editor to be quite so problematic, and was assumed to mean that the desire for the car was there, but the cost would be well beyond his reach. Response 14 was not listed in the ‘other’ category because it was deemed irrelevant; note that that is not same thing as ‘other’.

This should give you some idea of the ‘categorising’ abilities required when editing data. It gets more complex than this too. Many open-ended answers include more than one point, or idea, in the single response. The editor, or coder, in this case, has to decide whether to split them, or if the two (or more) points together mean a single thing.

In setting up these open-ended code frames, it is obviously best for one person to work on one question. If it is necessary for more than one person to code the same open-ended responses, they must at least get occasional looks at each other’s work, and discuss particularly difficult responses.

Even with computers available, many coders prefer to do the initial work in pencil on large sheets of paper, finding it more effective that way to draw arrows, circle chunks of material, cross things out, and so on. If you picture such a sheet of paper, you will see that it is the visual counterpart of the open-minded attitude required for such work, reflecting a necessary readiness to change one’s mind in an instant if new information betrays an earlier wrong categorising decision. This sounds melodramatic, perhaps, but it can be a very tough mental exercise, even after a good deal of experience.

Coding questions that are not open-ended, for companies that specialise in doing large surveys, requires much self-discipline for the long periods of time required. This can be more laborious than challenging, but here too an experienced coder keeps an eye out at the same time for falsified entries and anything else that might be of interest to the research company. There is seldom much to find, though, and one would like to think that these larger firms are at least varying their coders’ work fairly regularly in order to keep up their morale.

So, if you are buying research, you would want to know that your qualitative work was being done by one person, or by a few in constant contact, and that your quantitative work was being done by someone who liked their job well enough to take an interest in it, or at least had a basic level of experience. (These two points are often mutually exclusive.)

Tabulating

Tabulating is the final step in data preparation. It simply means counting the number of responses in the various data categories. For your research, if you choose to do something like this, you will probably have relatively few categories, a small sample and limited analysis. Manual tabulation would almost certainly be best.

‘Cross-tabulation’ is an important process in research, but it would be more appropriate to consider it in the next topic, under ‘data analysis’.

If you ever need to use an outside firm for tabulating data, the following guidelines regarding the chosen firm will be helpful:

- It should specialize in tabulating for the research industry.
- It should be able to write custom programs for the data.
- It should proof all tables before delivery to clients.
- References should be obtained from previous users of the service.
- It should not be selected on the basis of ‘cheapest’ service.
- The service should have some input into the questionnaire’s design since the design affects the tabulation format.

(Kress 1988, p. 233)

Review

Summary

This topic has covered the important aspects of data preparation. It has offered some general guidelines to follow, and has attempted to give an idea of what is involved in the editing, coding and tabulating of data. Each of these stages, the definitions of which generally overlap, has its pitfalls for the unwary. Much care is required here.

It is recommended too that you take the effort when reading research reports to consider the human factor behind the numbers and categorisations; it might well help you to catch inaccuracies that slip past others. With computers being used now, to some extent, at every stage, serious number problems can infiltrate a study that in the past would probably have been picked up at an early stage as obvious mistakes.

References

Green, P. E. & Tull, D. S. 1978, *Research for Marketing Decisions*, 4th edn, Prentice Hall, Englewood Cliffs, NJ.

Kress, G. 1988, *Marketing Research*, 3rd edn, Prentice Hall, Englewood Cliffs, NJ.

Luck, D. J. & Rubin, R. S. 1987, *Marketing Research*, 7th edn, Prentice Hall, Englewood Cliffs, NJ.

Sekaran, U. 1992, *Research Methods for Business: A Skill-Building Approach*, 2nd edn, John Wiley, New York.

Weiers, R. M. 1984, *Marketing Research*, Prentice Hall, Englewood Cliffs, NJ.