Question Bank FACTSHEET 4

Evaluating Survey Questions

Introduction:

Our intention here at the Question Bank is to provide you, the user, with plenty of examples of survey questions in order to help you to ask better questions in your own survey instruments. This evidently begs the question "What makes one question better than another?" The purpose of this factsheet is to try to answer that question in as constructive a manner as possible.

It is probably not possible to define a "good" question other than to say that it is one that collects accurate and informative data.

It may be somewhat easier to set out a typology of "bad" questions, which are likely to fail to collect accurate and informative data, than to describe all of the possible ways in which a question may succeed. So the recommended approach is to be critical, while accepting that perfection is probably impossible. This approach has the disadvantage that it will be difficult to provide examples of the problems to be discussed from within the materials in the QB, since we hope that these are drawn from the best current practice.

Data collected may be accurate but not informative if the question fails to generate responses that differentiate sufficiently between respondents. If a very high proportion of respondents give precisely the same answer to any one question then it will be very difficult to deduce anything about the causes or relationships of that variable with respect to any other variables in the dataset. So it is of fundamental importance that each question is capable of being answered in a realistic variety of ways by honest and fair-minded respondents.

Data that is not accurate will provide misinformation and will lead analysts to draw false conclusions. If a survey generates data that appears to contradict the expectations of the researcher, or a wider public, it may be important to re-examine the key questions involved in that data collection instrument to confirm that they were sufficiently robust in order to pre-empt likely criticism of the whole process.

Even when a question appears to be very good at the time the survey instrument is designed, there is no substitute for carrying out a thorough pilot study to test the way the instrument works in the field. This is the best form of evaluation because it calls real respondents into play. All theoretical analysis of draft questions suffers from the disadvantage that it is limited by the knowledge of the researcher who is probably an 'outsider' to the field being researched. In the comments that follow, the first step in question evaluation should be for the evaluator to attempt to put themselves in the position of a likely participant in the research who is trying to answer the question. The second step is to analyse the data obtained from a pilot study by looking for evidence that these sorts of problems may have arisen.

Causes of inaccurate data collection:

Some of the well-known causes of inaccurate data collection are as follows:

- question wording that is misunderstood by respondents
- questions requiring data that is too difficult to remember, calculate or estimate
- questions that upset or annoy many respondents
- addressing too many issues in a single question
- inclusion of double negatives in the question and responses
- the use of specialist terms or words unfamiliar to the respondents
- the use of unevenly balanced scales in attitude questions
- allowing response categories to overlap at their boundaries

For a wide-ranging discussion of question development, including 8 criteria for assessing questions and 19 principles for designing them, see Dillman (2000). Note that he points out that it is all too easy for the solution to one particular problem to infringe another principle, so compromises often have to be made. Dillman himself refers to the work of Payne (1951) who produced a "Check-List of 100 Considerations", so there is no shortage of advice available in this area.

Wording that is misunderstood by respondents:

It is clearly of crucial importance that everyone involved in the research, including the researchers themselves, the interviewers, respondents and analysts interpret each question and concept in the same way otherwise the data may be meaningless. Misunderstandings can arise through the use of language that is too technical, assumes too much prior knowledge, or is ambiguous.

It would seem to be desirable to phrase questions as far as possible in the words that respondents might use in their day-to-day speech, but the problem with this is that speech idioms vary from place to place, and time to time, so it may be unwise to take such an approach too far. Also, where an interviewer is quite obviously not a member of a particular community s/he may lose credibility if they are made to speak in the terms of that community when these are apparently alien to them. So the balance has to be struck by finding words that carry the desired meaning in a neutral way, without using academic concepts or acronyms. One solution to this problem may be seen in the lists of equivalent names used in surveys of drug misuse, where a series of slang names has been used for each drug. (Example 1)

Example 1– Smoking, Drinking and Drug Use Among Young People, 2005, Main Questionnaire

Q34 The next set of questions are about **Cannabis**, also called **Marijuana, Dope, Pot, Blow, Hash, Skunk, Puff, Grass, Draw, Ganja, Spliff, Smoke, Weed, Wacky Backy.**

It is also important to avoid using long and difficult words while trying to keep the overall length of each question as short as possible. The more words there are in a question the more likely respondents are to lose their way in it, and then they may answer the question that they think they heard rather than the one you tried to ask.

Data that is too difficult to remember, calculate or estimate:

There is a good chance that your survey will collect inaccurate data if you ask apparently factual questions that are just too difficult to answer correctly. The topic that interests the researcher may be utterly trivial to many respondents and so they may not have noticed the last time they had whatever experience it is you are asking about, or they forgot about it within a few moments. Maybe you are asking about something that they would prefer to forget, because it was embarrassing or distressing. In these circumstances the question wording will need to be understanding and sensitive, and should somehow encourage the respondent to overcome the difficulty and not to make a hasty guess.

It should be easier, with computers at your disposal, for you as researcher to perform arithmetic calculations than it is for the respondents to do mental work. So, complex matters might well be broken down into separate steps, each of which places a modest burden on the subject. These can then be combined in a programmed calculation and the answer fed back to the respondent for confirmation that it seems reasonable to them from their perspective. (see Example 2).

Example 2 – Expenditure and Food Survey 2003/ 4, Income questionnaire

20.110 QTJbHrsU.**DVTotHrU** (*)
APPLIES IF QTJbHrsU.EverOT = 1
Total usual hours (if work overtime)
0.00..97.00

20.115 QTJbHrsU.**AgreeHrs (*)**
APPLIES IF QTJbHrsU.EverOT = 1
Your total usual hours comes to (sum of UsuHr, POtHr & UOtHr). Is that about right, or not?
TOTAL MUST BE AGREED WITH RESPONDENT
Yes.....(1)
No.....(2)

Many people are poor at estimating numerical information about their lives so care needs to be taken when offering sets of response categories for questions, such as those about TV watching habits or time spent actively studying, that these do not influence the answers by indicating certain expectations. Dillman discusses research in which he assisted which demonstrated some substantial effects produced by using different response sets (2000 page 33). Example 3 shows a low set of time categories and Example 4 shows a higher set, both to similar questions about TV viewing habits.

Example 3 – European Social Survey 2002, Main Questionnaire

A1 **CARD 1** On an average weekday, how much time, in total, do you spend watching television? Please use this card to answer.

No time at all	00	**GO TO A3**
Less than ½ hour	01	
½ hour to 1 hour	02	
More than 1 hour, up to1½ hours	03	
More than 1½ hours, up to 2 hours	04	**ASK A2**
More than 2 hours, up to 2½ hours	05	
More than 2½ hours, up to 3 hours	06	
More than 3 hours	07	
(Don't know)	88	

Example 4 – National Adult Learning Survey 2001, Questionnaire – L, Lifestyle

```
TVWeek
"CARD L1 On weekdays during a normal the week, how much time
do you usually spend each day watching television or videos,
including programmes which you have videoed yourself.  By
watching television we mean when it is your main activity?":

(Never "I never/ hardly ever watch television or videos"
Notmuch "I do not watch television or videos every day"
Lessone "Less than one hour a day"
Onetwo "One to two hours a day"
Twofive "More than two but less than five hours a day"
Morefive "More than five hours a day")
```

Questions that upset or annoy many respondents:

It is all too easy to lose the goodwill of respondents by asking them questions that generate unhappy emotions towards the interview. If research is going to be useful in sensitive or important areas of social life then it will probably be impossible to avoid upsetting some people, but care should be taken to minimize offence for most people. This is not to say that sensitive topics should not be researched, clearly the success of surveys like the National Survey of Sexual Attitudes and Lifestyles indicate that this can be done. However care should be taken even where the topic is not expected to be particularly sensitive.

Respondents may be annoyed or upset if they come to think that their time is being wasted, for example through being asked repetitive or irrelevant questions, or if they feel that the researcher has not done sufficiently thorough preparation, for example when the questions use outdated terminology. Incorrect spelling (in self-completion surveys) or poor grammar can alienate some respondents, so if English is not your strong point you should ask someone with good English skills to check your questionnaire for these mistakes.

Addressing too many issues in a single question:

This problem is sometimes described as that of asking a double-barrelled question. If the wording of a question can be broken down into two or more separate questions, and it can be envisaged that one person might answer those separate questions in different ways then it will be impossible to analyse the responses to the combined question without ambiguity. This can sometimes arise when a researcher tries to help the respondent with some additional explanation or illustration in the question text. Groves et al. provide a useful example "do you favour legalized abortion because it gives women the right to choose?" (2004, p233). It is possible that some people might say 'yes' to the question "Do you favour legalized abortion?" and say 'no' to the question "do you think women should have the right to choose to have an abortion?" and so they would find the combined question impossible to answer faithfully.

The trap that is easy to fall into here is that of trying to provide reasons why an opinion might be held as a justification for asking about those opinions. The solution is to ask about the opinion in one question and then ask for possible reasons in another. It may be important to do this in this particular order because getting respondents to think about possible reasons might create context effects that could alter their interpretation of the opinion if it were asked afterwards.

Inclusion of double negatives in the question and responses:

Because everyday speech often contains statements and questions based around a negative phrase it is easy to include this style in a survey question. Then if the answer categories are 'yes' and 'no', a respondent may be put in the situation where in order to confirm a positive view they have to say 'no'. But, where a misunderstanding in conversation will probably be found and corrected quite easily, in the formal data collection process it may be left to stand unchallenged. So it is probably best to try to phrase all questions with a positive perspective.

The use of specialist terms or words unfamiliar to the respondents:

It is probably obvious to most researchers that some questions designed for a survey of adults will be unsuitable for use in a survey of children. However similar problems can arise when concepts are addressed directly rather than through the language by which they might be recognised by the subjects of a study. For example it will not be helpful to ask people about their 'social capital', instead the concept has to be operationalised in terms of activities that people might actually do such as helping their neighbours. It is also advisable to make sure that the general language used throughout a questionnaire is kept at a level of simplicity sufficient to ensure that almost all respondents should be able to understand the questions without further explanation.

The use of unevenly balanced scales in attitude questions:

Researchers are probably unlikely to be tempted to put unequal options for positive and negative views into a survey, but the sponsors of research are sometimes known to put pressure on for this. The justification when this happens is generally in order to achieve better discrimination between similar opinions. If we expect 80% or so of the sample to agree with a suggested attitude then it may seem to be informative to create more subtly distinct response categories on the positive side and fewer on the negative, but this could be seen as creating a distorting effect that would steer some of those who might have disagreed towards an agreement response.

In self-completion attitude questions the best advice (for example Dillman, 2000) is to have an even number of response categories, with the last one being a "Don't know/No opinion" option and a "Neutral" option as the central value of the remaining (uneven number) options. It is also best to label all of the values and not to leave some intermediate choices to be inferred from their place in the overall pattern (see Example 5). If the questionnaire is to be administered through an interview then it may sometimes be better to withhold the "Don't know/No opinion" option from the list offered to the respondent but permit the interviewer to record that if the respondent insists on that answer (see Example 6 taken from the same survey as Example 5 but in the interview section).

Example 5 – British Social Attitudes Survey 2005, Self completion questionnaire

58. Please tick <u>one</u> box on each line to show how much you agree or disagree with each of these statements:

*PLEASE TICK **ONE** BOX ON EACH LINE*	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Can't choose
a. Most disabled people should expect to work rather than rely on benefits.	☐	☐	☐	☐	☐	☐
b. Disabled people make just as good parents as people who are not disabled.	☐	☐	☐	☐	☐	☐

Example 6 – British Social Attitudes Survey 2005, Main Questionnaire

```
Q893    [TUSmooth]
        CARD (H6/H8)  AGAIN
        (Using this card, please say how strongly you agree or disagree with the
        following statement about (trade unions/the staff association) at your
        workplace.
        ((Trade unions/The staff association) at my workplace . . .)
        . . . (help/helps) make things run more smoothly at work

*       [TUNotice] tp [TUSmooth]

1       Agree strongly
2       Agree
3       Neither agree nor disagree
4       Disagree
5       Disagree strongly
```

Allowing response categories to overlap at their boundaries:

It is a common mistake to make in factual questions, such as those establishing a monetary amount like income or property value, to set bands of values as the responses with the same unique value being the upper limit of one band and the lower limit of the next. For example "£10,000 to £15,000, £15,000 to £20,000, £20,000 to 25,000 etc.". The problem is that a respondent who estimates his own value to be the same as one of these boundary points could truthfully respond with either of two values. It may appear to be pedantic to set the values instead as "£10,001 to £15,000, £15,001 to £20,000, £20,001 to £25,000 etc." but the resultant data collected will be significantly more accurate.

Suggested standards and methods of question evaluation:

It may also be useful to evaluate questions using a different approach. Groves et al. (2004 Chapter 8) identify three standards that good questions would be expected to meet. These are as follows:

1. Content standard – does the question ask about the right thing? Good questions will meet the needs of the researchers and analysts by addressing the issues with which they are concerned, and will be answerable by reasonable respondents.
2. Cognitive standard – does the question make sense to the respondents who will be asked to answer it? And will those respondents be able to answer it accurately?
3. Usability standard – can all the people involved in the process use the question easily and effectively? This includes aspects such as how easy it is to read the question aloud (for interviewers) or to grasp its meaning on a first reading (for self-completion studies).

The same authors suggest five different methods that might be employed to evaluate a draft questionnaire, or at least to inform the preparation of a draft. These are as follows:

1. Expert review – get experienced researchers and questionnaire designers to examine the draft with the benefit of their experience with similar instruments.
2. Focus groups – use these to identify the terminology used by likely participants and the sort of response categories that will be needed for closed questions.
3. Cognitive interviews – these are interviews where proposed questions are tried out on a very small sample of respondents from the target population, with detailed analysis of the processes used to generate the answers provided, either by discussing these with the respondents during the interview or by recording the interview and then analysing the transcript, audio- or video-tape.

4. Pilot studies – in this case a slightly rather larger sample of respondents is drawn, the proposed questionnaire is administered to them under normal interview conditions, and then subsequent debriefing of the interviewers and analysis of the data collected is used to identify weaknesses and problems.
5. Split-ballot experiments – in these cases two or more different versions of a question are used in a large scale survey, allocated randomly to the sample respondents, and then the data is analysed to see if any differences can be observed between the set of responses to each version of the question. It may still require a subjective judgement to decide which version provided the most accurate data.

Clearly there are different costs and benefits associated with each of these methods, and some of them are likely to be well beyond the means of small-scale research teams. They also address different aspects of the three standards mentioned before. However some major studies now include some information about their question development process in their technical reports (for example the British Crime Survey) which may be very useful material for a researcher thinking of adapting a question from that source. It should also be born in mind that many of the questions to be found amongst the Question Bank's materials have been subject to some of these processes already.

References:

Dillman DA (2007) *Mail and Internet Surveys – the Tailored Design Method 2007 Update with New Inernet, Visual and Mixed Modes Guide*. Wiley, New York

Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2004) *Survey Methodology*. Wiley, New Jersey.

Payne SL (1951) *The Art of Asking Questions*. Princeton University Press, Princeton New Jersey.