

The Difficulty of Understanding Social Survey Questionnaires from the Published Documentation

N. Graham Hughes

Abstract

The use of computer interviewing programs has contributed to a significant increase in the complexity of survey instruments. Research has shown that survey data can sometimes be affected by related matters arising during interviews. Current methods of documenting the interviewing programs show a lack of standards or conventions, making it difficult for secondary analysts to identify potential interview context effects. The solutions proposed in recent ambitious documentation projects have yet to be implemented widely. In the meantime this paper suggests some simple presentation conventions to be applied whenever a text version of a questionnaire is being prepared, in the hope that these might enable more readily understandable documentation to be published.

Keywords

Questionnaire documentation; routing instructions; control checks; presentation; conventions.

1. Why the documentation Matters

It is a fundamental assumption of quantitative social research that the answers provided by respondents in relation to each particular variable are equivalent to each other: without this there would be no meaning attachable to the mean, no standard to explain the standard deviation. Social scientists need to be able to conclude that if, for example, 72.6% of respondents agreed with an attitude suggested in a question then that explains something about the population from which those respondents were sampled. It has long been known that this is not possible if each respondent is asked a different question and so in the past great efforts were made to ensure that the same question was asked in the same way of every respondent being interviewed (Groves et al. 2004). This paper looks at the consequences of asking any one question in a wide variety of circumstances on the homogeneity of the data thus collected.

This concern has arisen through the author's work as the Content Manager at the ESRC-funded Question Bank (Qb). The Qb contains a great many survey questionnaires and makes them readily available to other researchers over the internet. Research using the questionnaires in the Qb has shown that no two questionnaires use the same stylistic conventions in their presentation of the various elements that they contain. This makes it more difficult than necessary to interpret and compare extracts from different surveys. Further research has shown that human error, in the form of inconsistencies, is apparent within the

texts. This puts into question the role of a published survey questionnaire as a “truthful” document.

Why does this matter? Well, in the years before Computer Assisted Interviewing (CAI) became widely available a survey instrument frequently took the form of a printed document given to the interviewer who, after careful training, would be expected to follow the fixed instructions contained therein. As Kent & Willenborg put it “... the questionnaire form was the documentation” (1997), so anyone else subsequently could examine the printed document and have a reasonable expectation of understanding what should have happened during the interviews. There was an effective restriction on the extent of complexity in such documents because of the need for interviewers to apply them consistently in the interview situation. The advent of CAI has removed some of those constraints and now, because the interviewer only needs to see one question at a time on the screen, it is possible to include far more complex routing instructions, which the computer can be expected to apply reliably. However at the same time, because the instrument is encoded in the software, there is no document simply available for anyone else to examine in order to learn what happened in the interview. The questionnaires that are published in technical reports and with datasets are produced through manual editing of the CAI program (Kelly, 2000). So as the instruments have become more complex they have also acquired the need for additional interpretation before they can be understood.

The situation now exists where there can be so many different possible routes through a single questionnaire that it may be possible for some respondents to have effectively unique interviews whereby nobody else answers precisely the same set of questions. In these circumstances, can the analyst still be sure that the answer data collected for any single question are sufficiently equivalent, in other words sufficiently homogeneous, to be combined and used as a simple data set? The work of Schuman and Presser (1996) and Tourangeau, Rips and Rasinski (2000) gives cause for concern over context effects in some circumstances. If there are doubts about this, is it possible for analysts to establish clearly which other questions asked earlier in the interview could have made a difference to some respondents’ interpretation of a later question? It is for these reasons that secondary analysts need clear documentation of what has occurred.

The TADEQ or Tool for the Analysis and Documentation of Electronic Questionnaires (Bethlehem and Hundepool, 2004) project attempted to develop software which would allow reporting of CAI programs in standard formats. Initially TADEQ focussed on BLAISE as the source CAI program but there was an intention that it would eventually be compatible with other major programs. For reasons that are not clear, the TADEQ project does not seem to have been realised and analysts are still seeing the use of single linear text documents as the primary explanation of CAI questionnaires.

2. Elements Typically Included in Questionnaire Documentation

Common elements

In order to examine these issues this research has attempted to identify the set of basic elements that make up the current documentation of a CAI questionnaire. The following (brief) list uses unambiguous terminology to highlight their apparent uses in examples of

documentation drawn from the Qb. The examples can be seen by following the web links listed in the Appendix to this paper.

At the heart of the issue is the question text which is rarely made prominent in the documentation but is usually identifiable from its content and language. Example 1 shows a presentation where the question text is difficult to find. The question is almost invariably followed by the set of response categories pre-coded for analysis. Example 2 shows a clearer presentation of question text and response set, but the response codes to the right are not easily aligned with their meanings. Example 3 shows another way of presenting the response set. The question text is usually preceded by the variable name, a mixture of code and abbreviation, as in Example 4 where a lengthy presentation of variable names, section number and code, and question number and name, uses a mixture of capitals and lower case. Example 5 is rather simpler and just shows the variable name in bold face.

Frequently there will be a clue to the type of variable or question somewhere in the vicinity; for example some questions will only accept a single answer, some will accept any number of answers at the discretion of the respondent, some will be open questions with a space constraint on the number of characters of free text, and others will be calculated by the program and not actually asked aloud. These types (single, multi, open, computed) have to be inferred by the reader from instructions apparently given to the interviewer in the text. In Example 6 note “CODE ALL THAT APPLY – Multicoded (Maximum of 11 codes)”. There are other sorts of instruction for the interviewers as well, for example telling him/her to use a particular show-card, advising him/her when to prompt the respondent for more detail, highlighting some questions as very important for the sake of correct routing later, and sometimes telling him/her when a ‘Don’t know’ or ‘refusal’ response is not acceptable. Example 7 includes several prompts finishing with “EXCLUDE BBC WEBSITE”).

Routing, control checks and text-fills

Whilst the elements described so far seem to arrive in the published documentation as direct representations of text used in the program itself, the treatment of routing instructions and control checks is rather different. With these items the amount of interpretation and editing by those who prepare the documentation is much greater.

For routing (or, as it is sometimes called, “skip patterns”) several important decisions have to be taken as the documentation is prepared. These are

- i) whether to use “Go To” or “If...Then...Else” logical instructions,
- ii) how to represent nested conditions,
- iii) whether to show the conditional logic in plain language or in the algebra of variable names, values and mathematical symbols, and
- iv) how to represent loops, tables, parallel modules and sub-samples.

Examples 8 to 12 illustrate different approaches to some of these presentation decisions.

Control checks may be defined as “hard”, “soft” or “range” checks. They are used to improve the accuracy of the data collection process by identifying possible errors or inconsistencies in the recorded responses and either forcing or facilitating corrections to be made. These present the documentation editors with similar problems to those of routing over how to show the logical conditions that trigger these interruptions to the interview process. Example 13 shows one presentation of control checks. These may not always be fully reported in published documentation.

Another common feature of CAPI interviewing programs which also utilise conditional logic are text-fills or data-fills. These bits of programming cleverness can appear within the text of questions, interviewer instructions, or error messages arising from control checks. Their purpose is to make the interview flow more smoothly as a conversation by using information from earlier questions to personalise the language of subsequent items. Example 14 shows how recording the logical conditions controlling the text-fills within a question text can make the resulting document very difficult to comprehend.

The advantages for the data collection process of using sophisticated routing and control checks relate to shortening the interview by eliminating irrelevant questions and improving the accuracy of the data by reducing incorrect responses. The disadvantages are less apparent in terms of the introduction of variations in the experiences of the different respondents. For the reader of the questionnaire documentation, these elements bring additional challenges. Is it possible to work out what happened in a specific interview, or in the generality of interviews? Can context effects be identified that may have affected some respondents as they answered questions critical to the data being studied? Can it be deduced why some respondents appear to have been asked a particular question while others were not? Can analysts decide whether data collected in two different surveys, but using the same question, is really comparable or not? To deal with these challenges, clear and accurate documentation is needed.

Hopefully the range of examples used to illustrate these elements has begun to create an impression of the confusing variety of styles and conventions that have been used in these questionnaire documents in recent years. This summary has been offered, not as a criticism of the survey agencies that have prepared these materials but as an illustration of what is to be found when their forms rather than their contents are studied.

3. Proposal for Standard Conventions

Whilst the ambition of projects like TADEQ and DDI is admirable, it does not seem likely that they are going to be implemented quickly or universally. In the meantime, until they are widely implemented, it will almost certainly be necessary to continue to use manually edited linear documents. The modest proposal here is to suggest a standardised set of conventions that could be used by almost all questionnaire documentation teams whenever they are asked to prepare a text version of a questionnaire, without requiring any new technology or software. The simple goal is to see common ground across a whole range of survey questionnaires so that users can devote their energy to understanding the events of whatever survey they are studying rather than having to first figure out how to read the document. Such a step might mark significant progress towards future automation of the process by which survey metadata is retrieved from these documents.

A wish-list for such conventions would be as follows:

1. Use a fixed order or sequence for the basic elements within each question 'event' in the documentation. (E.g. Question number, variable name, variable type, routing applicable, question text, calculation rule, response set, show-card used, interviewer permitted prompts, range checks, hard checks, soft checks, explanatory material.)

2. Use standard font characteristics to help identify each element above. (E.g. shape of brackets, italics, capitals, bold-face and inverted commas.) This need not restrict the actual font style so that surveys could maintain some individuality of appearance.
3. Use a common structure for routing instructions. Probably the ideal from a user's perspective would be the "Ask if" format with algebraic style code of both previous variables and relevant responses followed by full plain English text, placed in every question event.
4. Avoid using short-cuts such as tables and common response sets referenced by asterisk. Instead try to show the full questions in standard form.
5. Show the conditional logic for all checks in algebraic style code form only (for brevity) with the relevant error message in full text.
6. If computations, conditional logic or routing instructions use data from a source outside the current questionnaire (such as a previous interview or wave of the survey) list all these variables and their associated response sets in a clearly identified section near the start of the document.

Illustrations and examples

In order to add some substance to these proposals, here is a suggestion for a basic set of stylistic conventions which might be useful as the starting point for a discussion.

Suggested questionnaire style example:

Question # **Variable Name** { Variable type: single/multicode/text/calculated etc }

ASK IF: [Variable1 = # AND Variable2 <> #], [Plain language statement of routing logic]

"Question text, including (data-fill alternative1/datafill alternative2) and ending in a question mark?" (Data-fills determined by Variable#)

Or put calculation rules here if it is a derived variable.

1. First response
2. Second response
3. Third response etc.

SHOW-CARD # / READ OUT RESPONSES LIST

Interviewer instructions, in italics, "with any text to be spoken by the interviewer (including data-fills) in bold and between inverted commas". (Data-fills determined by Variable# & Variable#).

CONTROL CHECKS:

Type: Hard/Soft/Range

[Logical condition as algebraic expression only]

Instructions for interviewer, including "any text to be spoken"

Explanatory notes.

In this suggestion bold font is reserved for the variable name and spoken text only, the use of capital letters is kept to a minimum, logical expressions are enclosed within square brackets, text-fills and data-fills are enclosed within curved brackets, and instructions for interviewers are shown in italics. The routing is shown in both algebraic form and plain language, and is

placed after the variable name but before the question text, in order to be as unambiguous as possible. Not shown, but also suggested, is a separating line between each question to indicate when a fresh screen would have been shown to the interviewer.

Below are two illustrations of this style suggestion as applied to existing questionnaire material. The first is from the British Social Attitudes Survey 2005, shown as Example 6. Here the proposed presentation is shown first with the current version as taken from the Qb materials beneath it. Because this example is of a multicoded variable it is suggested that the variable name associated with each response is placed to the left of the response code. The explanatory notes are derived from information referenced by the \$ 1 in the original (an introductory note and footnotes).

Reworked Example 6 – British Social Attitudes Survey 2005 – Main Questionnaire

Q565-575 **WDisFW** {Multicoded – maximum 11 codes}

ASK IF: [Version = A, C or D], [All of sub-samples A, C and D]

“People have different ideas about what it means to be disabled. Which of the people on this card would you think of as a disabled person?”

WDisNone	0	None of these
WDisArth	1	A person with severe arthritis
WDisAIDS	2	A person who has HIV/AIDS
WDisSchi	3	A person who has a diagnosis of schizophrenia
WDisDepr	4	A person who has a diagnosis of severe depression
WDisDown	5	A person who has Down’s Syndrome
WDisCanc	6	A person who has cancer
WDisOldH	7	An older person who cannot hear without a hearing aid
WDisBlin	8	A blind person
WDisWhlc	9	A person who uses a wheelchair most of the time
WDisBrok	10	A person with a broken leg, using crutches while it heals
WDisFacD	11	A person with a severe facial disfigurement
	97	All of these

SHOWCARD E1

Interviewer: code all that apply and probe: “Which others?”

Explanation: Variable WDisFW does not appear in SPSS file, the 12 variables derived from it, and shown above, do appear with values 0=not mentioned, 1=mentioned.

Example 6 - British Social Attitudes Survey 2005 – Main Questionnaire – Disability

VERSIONS A, C AND D: ASK ALL
 Q565- [WDisFW] \$¹
 Q575 CARD E1
 People have different ideas about what it means to be disabled.
 Which of the people on this card would you think of as a disabled
 person?
 PROBE: Which others?
 CODE ALL THAT APPLY
 Multicoded (Maximum of 11 codes)

0	(None of these)	[WDisNone]
1	A person with severe arthritis	[WDisArth]
2	A person who has HIV/AIDS	[WDisAIDS]
3	A person who has a diagnosis of schizophrenia	[WDisSchi]
4	A person who has a diagnosis of severe depression	[WDisDepr]
5	A person who has Down's Syndrome	[WDisDown]
6	A person who has cancer	[WDisCanc]
7	An older person who cannot hear without a hearing aid	[WDisOldH]
8	A blind person	[WDisBlin]
9	A person who uses a wheelchair most of the time	[WDisWhlc]
10	A person with a broken leg, using crutches while it heals	[WDisBrok]
11	A person with a severe facial disfigurement	[WDisFacD]
97	(All of these)	

Next is shown a reworking of Example 14 to illustrate how routing and text-fills might be clarified.

Reworked Example 14 – Offending, Crime and Justice Survey 2004 – CAPI Questionnaire

V1vehS {single code}

ASK IF: [V1veh = 1], [Someone living here has owned or regularly used a car, van, motorbike or other motor vehicle in last 12 months]

“Since the first of (month) 2003, (has anyone who lives here had their / have you had your / have you or anyone who lives here had their) motor vehicle stolen or driven away without permission, even if (they / you) later got it back?” (Data fills determined by month of interview, respondent aged 16 or more, number of people in household aged 16 or more).

1. Yes
2. No
3. Don't know
4. Refused

Example 14 - Offending, Crime and Justice Survey 2004 – CAPI Questionnaire – Victimization

V1vehS [ASK if V1veh=1]
 Since the first of [MONTH] 2003, [IF L1age<16: has anyone who lives here had their/ IF L1age>15 AND ONLY ONE PERSON 16+ IN HOUSEHOLD: have you had your/ IF L1age>15 AND 2 OR MORE PERSONS 16+ IN HOUSEHOLD: have you or anyone who lives here had their] motor vehicle STOLEN OR DRIVEN AWAY WITHOUT PERMISSION, even if [they/ IF L1age>15 AND ONLY 1 PERSON 16+ IN HOUSEHOLD: you] later got it back?

1. Yes
2. No
3. Don't Know
4. Refused

Documents using these conventions could still be prepared using standard word-processing software and then be converted to PDF for publication on the internet. However it would be quite possible subsequently to convert such documents to XML and apply standard content tags in order to make them more flexibly useful to users with appropriate software. This can be seen as a half-way step towards the DDI model. In the meantime the use of strong conventions in this way should help analysts and researchers to identify more clearly just how the data they are using was originally collected.

Appendix: Examples – Web-Links and Screen Shots

Examples format: - Survey name & year – Questionnaire PDF name – Bookmarked link. (The web-link will take you to the start of the PDF document and you should then click on the relevant bookmark, or use the Adobe page number system, to locate the specific page referred to in this paper.)

1. National Travel Survey 2004 – Individual Questionnaire – Journey to work
<http://qb.soc.surrey.ac.uk/surveys/nts/04individual.pdf> (p20/42)
2. Health Education Population Survey 2005 – Main Questionnaire – Alcohol
<http://qb.soc.surrey.ac.uk/surveys/heps/05mainqheps.pdf> (p25/55)
3. British Crime Survey 2004/5 – Main Questionnaire – Main questionnaire
<http://qb.soc.surrey.ac.uk/surveys/bcs/04mainqbc.pdf> (p9/118)
4. Family Expenditure Survey 1999 – Household Questionnaire Part5 – Purchase of Furniture
<http://qb.soc.surrey.ac.uk/surveys/fes/fes99hque5.pdf> (p24/30)
5. Health Survey for England 2004 – Household Questionnaire – Accommodation & tenure
<http://qb.soc.surrey.ac.uk/surveys/hse/04hqhse.pdf> (9/16)
6. British Social Attitudes Survey 2005 – Main Questionnaire – Disability
<http://qb.soc.surrey.ac.uk/surveys/bsa/05mainqbsa.pdf> (p57/174)
7. British Social Attitudes Survey 2005 – Main Questionnaire – Newspaper readership
<http://qb.soc.surrey.ac.uk/surveys/bsa/05mainqbsa.pdf> (p17/174)
8. Continuous Household Survey (NI) 2004 – Household Questionnaire – Tenure
<http://qb.soc.surrey.ac.uk/surveys/chs/04housechs.pdf> (p7/19)
9. ONS Omnibus Survey 2003/4 – Classificatory Questionnaire – Paid work
<http://qb.soc.surrey.ac.uk/surveys/omnibus/OMNClass0304.pdf> (p15/26)
10. People Families & Communities Survey 2005 – Questionnaire – Illness or Disability
<http://qb.soc.surrey.ac.uk/surveys/citizenship/05questcs.pdf> (p81/106)
11. English Longitudinal Study of Ageing 2004 – Private Questionnaire – Expectations
<http://qb.soc.surrey.ac.uk/surveys/elsa/04w2privatelsa.pdf> (p8/40)
12. Scottish Health Survey 1998 – Individual Questionnaire Part A – General Health module
http://qb.soc.surrey.ac.uk/surveys/ShealthS/SHS98_IQUIREa.pdf (p3/29)
13. Families and Children Study 2004 (Wave 6) – Main Questionnaire – 7. Housing
<http://qb.soc.surrey.ac.uk/surveys/facs/04mainquest.pdf> (p46/174)
14. Offending, Crime and Justice Survey 2004 – CAPI Questionnaire – Victimisation
<http://qb.soc.surrey.ac.uk/surveys/cjs/04capicjs.pdf> (p48/66)

References

- Bethlehem J and Hundepool A** (2004) TADEQ: A Tool for the Documentation and Analysis of Electronic Questionnaires. *Journal of Official Statistics*, vol.20, No.2, 233-264
- Groves RM, Fowler FJ, Couper M, Lepkowski JM, Singer E and Tourangeau R** (2004) *Survey Methodology*. Wiley, New Jersey.
- Kelly M** (2000) What users want from a tool for analysing and documenting electronic questionnaires: the user requirements for the TADEQ project. *Blaise Users Group conference paper*.
- Kent J-P and Willenborg L** (1997) Documenting questionnaires. *Research Paper No. 9708*, Voorburg: Department of Statistical Methods, Statistics Netherlands.
- Schuman H and Presser S** (1996) *Questions and Answers in Attitude Surveys*. Sage, Thousand Oaks CA.
- Tourangeau R, Rips LJ and Rasinski K** (2000) *The Psychology of Survey Response*. Cambridge University Press, Cambridge.

About the Author

Graham Hughes is the content manager of the ESRC-funded Question Bank, based at the University of Surrey. Since taking up the post in October 2005 he has added a considerable volume of questionnaires and related materials to that freely accessible web-site, all in PDF files. Prior to this appointment he completed an MSc degree at Surrey in Social Research Methods as a career change after working as a Chartered Accountant for the previous 23 years. He has worked with computers since 1984, making the principle of 'one PC per desk' a fundamental basis of the accountancy practice that he co-founded in 1987, and exploiting a variety of opportunities for installation of and training on software for his clients. He brings a combination of maturity and freshness to his current role. He can be contacted at Question Bank, Department of Sociology, University of Surrey, Guildford, Surrey, GU2 7XH; tel. 01483 682794; fax. 01483 689551; e-mail n.hughes@surrey.ac.uk.