Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

"Don't touch that ..., you don't know where it has been!": Computer Assisted Personal Interviewing and Question Recycling.

by N. Graham Hughes, Martin Bulmer and Julie Gibbs

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

"Don't touch that ..., you don't know where it has been!": Computer Assisted Personal Interviewing and Question Recycling.

N. Graham Hughes[1], Martin Bulmer[1], and Julie Gibbs[1]

Abstract

Past survey instruments, whether in the form of a paper questionnaire or telephone script, were their own documentation. Based on this, the ESRC Question Bank was created, providing free-access internet publication of questionnaires, enabling researchers to re-use questions, saving them trouble, whilst improving the comparability of their data with that collected by others. Today however, as survey technology and computer programs have become more sophisticated, accurate comprehension of the latest questionnaires seems more difficult, particularly when each survey team uses its own conventions to document complex items in technical reports. This paper seeks to illustrate these problems and suggest preliminary standards of presentation to be used until the process can be automated.

Key Words: Documentation, Complexity, Comprehension, Context effect, CAPI.

1. Introduction

1.1 Origins of the Question Bank

Thinking back to 1996 when the Economic & Social Research Council (ESRC) Question Bank (QB) was first set-up it is amazing to note the extent of the changes that have taken place in the world of survey research. In particular, the development of computer technology has had an enormous impact on the way that such research is conceived, planned, carried-out and analysed. It is often easy to be swept along in the current of new developments and to be persuaded that each new technical possibility must be an improvement on what went before because it allows researchers to do something now that previously they could not do. However, it may be wise from time to time to step back for a moment and review our practices to make sure that they do indeed match our assumptions about their effectiveness.

Back in 1996 we recognised the potential of the internet (or the "World Wide Web" as we called it then) to disseminate information for very little cost to either the producer or user of that material. And, in the application of survey research, many of us thought that if we could spread knowledge about the ways in which major public research was carried-out, specifically the questions that were being asked, then it should be possible to raise the quality of such research in general by making it easier for others to identify and adopt best practices. This was the primary driver behind the establishment of the QB. In simple terms, we hoped that by providing free access to a variety of significant social survey questionnaires we would enable a levelling-up process to occur, raising quality, saving time and starting a trend towards harmonisation.

Looking back now, some 12 years later, it can be seen that much has been achieved although some of today's practices of survey research would have been hard to predict even as recently as 1996. So, while we focussed our attention on how to use the new technology to display the questionnaires to a wider public, maybe we failed to foresee how the new technology would also change the way in which those questions would be asked. Looking back now, with the help of the QB's resources, we can see that in Britain the General Household Survey changed from a paper interview document (PAPI) in 1993/4 to a computer assisted interview (CAPI) in 1994/5, and the British Social Attitudes Survey similarly changed mode between 1993 and 1994. Of course these questionnaires would only

[1] N. Graham Hughes , Martin Bulmer and Julie Gibbs. ESRC Question Bank, Department of Sociology, University of Surrey, Guildford, Surrey, GU2 7XH, UK

have become available to us a year or two after the fieldwork had been completed, so the change-over was not really apparent until after the QB had been set-up.

1.2 Documenting CAPI questionnaires

Now the crucial point here is that while a PAPI questionnaire is its own documentation (Kent & Willenborg, 1997), a CAPI program has to be interpreted and edited before it becomes intelligible to anyone other than a programmer. The QB was founded just at the end of the PAPI era and inevitably its initial concept was formed by that era. The plan was to scan the paper questionnaire documents in order to turn them into digital images, and then to make these image files open to public view as Portable Document Format (PDF) documents. Those designing new questionnaires and analysts searching for understanding of data collection processes would benefit through relatively easy access to such materials. These new users would see exactly the same texts and layouts as the interviewers who had conducted the interviews, albeit without the training sessions that would have preceded fieldwork but without the constraints of completing a live interview within the limited patience of the respondent. As CAPI began to be adopted it was understandable that most users expected to see a document that looked something like the old PAPI questionnaires and so that was what the survey agencies produced.

It is not apparent that there have been any fundamental changes to CAPI program capabilities in recent years (apart maybe from enabling previously collected data in panel and longitudinal studies to be used by the program) but there does appear to have been an increase in the intensity with which the features of CAPI have been utilised. From the very earliest examples of CAPI we can see that the facilities to route interviews past irrelevant questions and to focus on details where they may be more fruitful were quickly recognised. But the extent and complexity of such routing schemes have increased considerably as the early constraints of processing power have been overcome. In the UK we are currently developing an Integrated Household Survey which brings together five separate surveys (including the Labour Force Survey, the General Household Survey and the Expenditure and Food Survey) within a single sampling procedure and utilising a common core set of questions. Differential routing schemes will take the sampled respondents down many different paths through this ‘integrated’ survey. But even the question texts are frequently varied in some surveys as the precise wording is tailored to the perceived understanding of each respondent. The combination of these two features, routing between questions and varying the wording of many questions, may have created the situation where no two respondents to a given survey are asked precisely the same set of questions. The phrase ‘may have created this situation’ is used because it would be extremely difficult to prove such an assertion, although it certainly appears to be possible.

Yet, while this step-change in the complexity of the questionnaire designs has been going on we have made no significant progress on solving the problems of rendering the CAPI program script intelligible to human readers. There have been several attempts to do this, including the Tool for the Analysis and Documentation of Electronic Questionnaires (TADEQ) (Bethlehem & Hundepool, 2004; Kelly, 2000), Blaise Automatic Documentation (BAD) and others, but to date no convincing solution appears to have been found. This has proved frustrating for the small team of staff working on the QB project at the University of Surrey. They are now conscious of the limitations of the PDF technology, which is still being used to add recently generated questionnaire materials to the website, but they are not aware of any convincing alternatives. In conversations with some staff at UK research organisations it has become apparent that the form of the questionnaire documentation currently being manually produced by the survey team staff is determined to a large extent by the requirements of the survey’s principal sponsors. And such sponsors can be extremely conservative about changes to documentation layouts and styles.

2. Problems for researchers

2.1 Uncertainty

This combination of circumstances, of increasing complexity and variety of question sequences and wording, and a lack of progress on methods to display or document such variety, is now beginning to impact on the core aim of the QB. One way of describing the activity that we hope to promote is the phrase “question recycling”. Our founding principle was to encourage researchers to re-use or recycle existing questions and response category sets in order to exploit best practice, and save time and cost that would otherwise be wasted on reinventing a wheel that already

exists. But how can our website users apply such recycling techniques when it is so difficult for them to see exactly when and how the questions were used before? This is the reason for the quotation in the title of this paper of many mothers' injunctions to their children while in the public domain – "Don't touch that ..., you don't know where it has been!". A researcher can only re-use a question with confidence when s/he can identify the full circumstances of its previous uses. Before the use of CAPI became widespread this was quite possible (if you could get your hands on a copy of the paper questionnaire) but now it is often much more difficult because the documentation is troublesome to interpret and may be inaccurate in some details.

Consider the example shown in figure 2.1-1 below, which is taken from the published documentation.

Figure 2.1-1
Extract from the English Longitudinal Study of Ageing: Wave 2, 2004/5 – Household Questionnaire

```
| | | | |
| | | | | IF (Sample status is core member at Wave 1 = Yes) OR (Sample status is partner at
| | | | | Wave 1 = Yes) [ELSA (Wave 1) = 1 OR Partner (Wave 1) = 1]
| | | | | |
| | | | | | IF reason for not living here = [Living elsewhere- relationship ended, Living
| | | | | | elsewhere- other reason] [DhWhat = 2, 4]
| | | | | | |
| | | | | | | DHARFB *@
| | | | | | | INTERVIEWER: This is a split household. Start ARF B and attempt to find follow-
| | | | | | | up address for them.
| | | | | | | Press 1 and <Enter> to continue.
| | | | | | | Range: 1..1
| | | | | | |
| | | | | | | IF (reason for not living here = [living elsewhere - relationship ended, living
| | | | | | | elsewhere - moved to a nursing/residential home, living elsewhere - other reason])
| | | | | | |  [DhWhat = 2 , 3, 4]
| | | | | | |
| | | | | | | DHMOV *@
| | | | | | | We may need to collect some more details about [^<correct first name>] but we will
| | | | | | | come back to this later.
| | | | | | | INTERVIEWER: Press 1 and <Enter> to continue.
| | | | | | | Range: 1..1
| | | | | | |
| | | | | | | END OF FILTER
| | | | | | |
| | | | | | END OF FILTER
| | | | | |
| | | | | END OF FILTER
```

Here the strings of vertical 'pipe' characters indicate the existence of filter patterns to restrict certain questions so that they are only asked of respondents for whom they are relevant. In this example there are up to seven such filters applying simultaneously and it would take considerable effort to trace each of these back to its origin, maybe several pages earlier in the script, to establish the exact circumstances in which such a variable would have been activated. However the skills required for the interpretation of this questionnaire may not be useful in many other cases because of the variety of styles in which this sort of information is currently presented across the range of surveys.

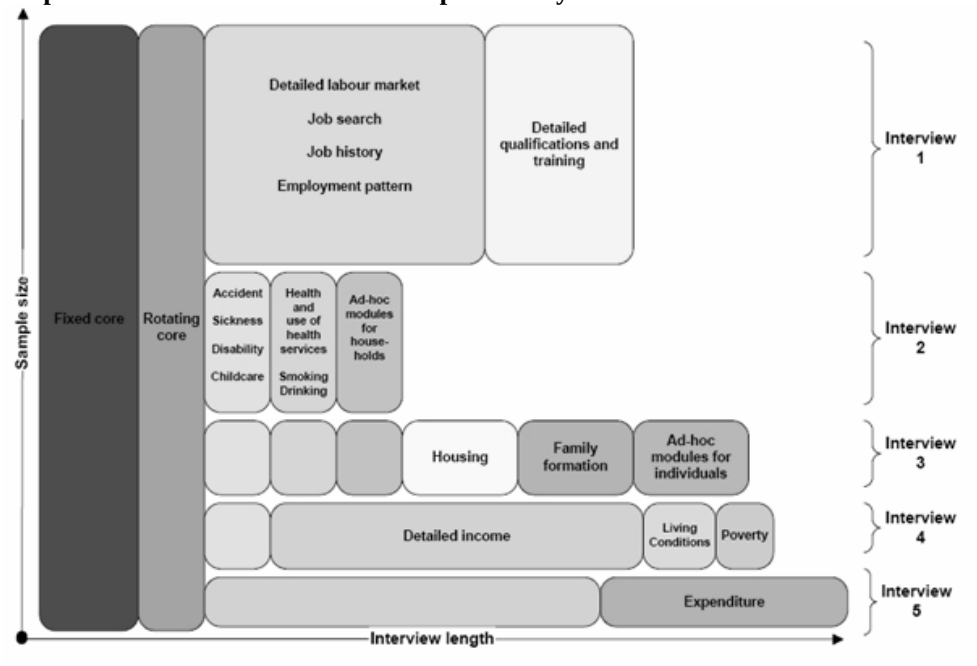
2.2 Context effects

There is, maybe, another problem arising from this combination of complexity in the program and the inadequacy of the documentation, and that is the possibility of unidentified context effects arising in the primary dataset (Schuman & Presser, 1996; Tourangeau Rips & Rasinski, 2000). Clearly this is not a problem at the extreme end of the variability scale where only a few respondents are asked a particularly unreliable set of questions because their biased responses will probably have little effect on the full data set. However, as the possibility of context effects is multiplied by the increasing use of multiple modules, analysts may need to devote more time and energy to looking for such distortions, and in particular to enquire whether the modules were always asked in the apparent sequence in

which they have been listed in the documentation. There is a legitimate concern that in some instances programmers or maybe even interviewers may decide to cover some modules in a different order and this may not be adequately documented.

Figure 2.2-1 (below) shows an extract from an Office for National Statistics (ONS) planning document for a complex survey which illustrates the potential problems. The shaded boxes represent the volumes of data to be collected in each module. The vertical axis gives an indication of the proportion of the sample which would be asked questions within each module stream, while the horizontal axis gives an indication of the number of questions to be asked in each module. The challenge for those creating and using the questionnaire documentation for such a survey is to find a way of showing, or knowing, the exact order in which the questions were asked because the CAPI programmers might decide that some interviews would flow better in a different sequence.

Figure 2.2-1
Proposed modular structure for a complex survey



Once again we have not had sufficient resources in the QB to investigate this possible problem, and maybe the datasets which might be most affected by it are not yet publicly available, but this is a plea to methodological researchers to consider this as a line to investigate with the new multi-module surveys.

3. Proposals

3.1 Future developments

Maybe we should make some positive suggestions about what we would like to see being done about these problems, from the perspective of users of the metadata for complex surveys. A promising development that we believe will be most helpful in the future is the adoption of the Data Documentation Initiative (DDI 3) as the basic setting for questionnaire documentation. For this to be really effective we would also want to see the development of some tools for interrogating and reporting the XML files containing this metadata. This has the potential for creating a significant advance in documentation quality to match the increased complexity of the questionnaires. When users can use a software tool to specify the particular elements of a survey questionnaire that they want to see and the style in which they want them to be displayed then they will have broken free from the grip of the survey sponsor that tends to limit the development of innovative documentation solutions. The sponsor can have his own report tailored

from the XML in the style to which he is accustomed, and other users will be able to design their own reports to be quite different.

3.2 Short term standardisation

In the meantime, so long as linear documents continue to be used based upon editing a text output from the CAPI system in a word-processing program, maybe we could recommend a limited amount of standardisation of layout. For example the most effective way of reporting routing that we have seen in recently prepared PDFs in the QB has been that of stating with each variable the set of conditions which have to be satisfied for that question to be asked (rather than indicating a "Go to" command beside some response categories for a previous question). For the maximum benefit of readers these routing conditions should be expressed both in plain language and in variable codes with arithmetic operators.

For example (adapting an instruction from the English Longitudinal Study of Ageing)

> Variable EXMOVHA:
> "Asked if respondent is not resident in an institution – IF IAskInst<>1"

Here the first part conveys the general sense of the routing logic, while the second unambiguously shows which previous variable and response govern the current question. It is unreliable to use the mathematical operators in the plain language element because many people are unsure of the precise meanings of ">" and "<".

But more information is not always an improvement on less. Consider figure 3.2-1, an extract from a questionnaire exactly as it is published in its technical report.

Figure 3.2-1
Extract from the Offending Crime and Justice Survey 2004

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

Here the reader's comprehension of the question is severely hampered by the inclusion of both the various alternative text substitutions used, and also of the logic rules governing the selection of those substitutions. This is probably going too far and a better result could be achieved by stating the simple alternatives and separately identifying which variables governed the substitution rules, as in the following suggestion:

> **"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?"** (Text substitutions determined by month of interview, respondent aged 16 or more, number of people in household aged 16 or more).

This leaves the reader with the task of working out how the identified variables would have been used to select each text substitution.

There is insufficient time and space in this paper to suggest a comprehensive set of standards to fully define this sort of documentation template, but the idea should be apparent. And this is only a temporary requirement to improve the accessibility of questionnaire documentation in the short-run before the DDI tools become available to take these issues into another dimension altogether.

4. Conclusion

This paper has argued that it is time we developed the skills of questionnaire documentation editors in order to match the enthusiasm shown by the CAPI programmers for complexity. It is no longer sufficient to continue documenting questionnaires in the ways we have used for the last 12 years because these ways no longer adequately reflect the reality of the instruments they purport to represent.

In the meantime we would like to draw the attention of all users of current survey questionnaire documentation to the problem of the limitations inherent in that documentation. These days it may be only a partial representation of what has occurred.

And finally we would like to suggest that methodological research be carried out into the possibility of context effects in the highly complicated multi-module surveys now being launched. It seems possible that the data collected may be more revealing than the questionnaire documentation itself if some groups of respondents have received different interview experiences.

References

Bethlehem, J. and Hundepool, A. (2004). TADEQ: A Tool for the Documentation and Analysis of Electronic Questionnaires. *Journal of Official Statistics*, 20, 2, 233-264

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*.

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*, Thousand Oaks CA: Sage.

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.