
Article:
Priede, C., Jokinen, A., Ruuskanen, E. et al. (1 more author) (2014) Which probes are most useful when undertaking cognitive interviews? International Journal of Social Research Methodology, 17 (5). pp. 559-568. ISSN 1364-5579

https://doi.org/10.1080/13645579.2013.799795

Which probes are most useful when undertaking cognitive interviews?

Camilla Priede, Anniina Jokinen, Elina Ruuskanen, Stephen Farrall*

*Camilla Priede is a Programme Director at The Institute for Lifelong Learning, Sheffield University. Elina Ruuskanen and Anniina Jokinen are researchers at the European Institute for Crime Prevention and Control, affiliated with the UN. Stephen Farrall (corresponding author) is a professor of criminology in the School of Law, Sheffield University. All have been working on the FP7-funded Euro-JUSTIS project, designing new survey measures of citizens' confidence in the criminal justice system. We would like to acknowledge the assistance of Ian Brunton-Smith for help with the multi-level modelling, and to the reviewers and editors, whose comments helped improve the clarity of our paper in places. Faults remain our own, of course.

Abstract

This paper reports the use of verbal probes in cognitive interviews undertaken to test the usefulness, validity and reliability of survey questions. Through examining the use of probes by three interviewers undertaking interviews as part the piloting of a cross-national crime survey we examine which of the various types of probes used in cognitive interviews produce the most useful information. Other influences on interview quality are examined, including differences between interviewers, and respondents themselves. The analyses rely on multi-level modelling and suggest that anticipated, emergent and conditional probes provide the most useful data. Furthermore, age, gender and educational levels appear to have no bearing on the quality of the data generated.

Introduction

Cognitive interviews (hereafter CIs) aim 'to understand the thought processes used to answer survey questions and to use this knowledge to find better ways of constructing, formulating and asking survey questions' (De Maio and Landreth 2004:90). The technique is used to test the reliability of survey questions by asking respondents about the thought processes that they use to answer the questions. This approach reveals issues around the quality of responses (in terms of whether the participant is answering the question in the way intended) and any difficulties that the participant has when answering the questions (Beatty and Willis 2007). In recent years cognitive interviewing, in particular that which uses verbal probing, has become more commonly used as a survey pre-testing technique (Thompson 2008, Willis 2005).

There is, however, little research which has examined the benefits and drawbacks of the different types of probes which might be used in cognitive interviews. That research which has been undertaken is anecdotal and reports impressionistic data from interviewers about respondent behaviour and the quality of responses. Our contribution to this literature is to provide initial answers to the following questions:

- Do certain types of probes produce data which is more useful in understanding people's responses to survey questions than other types of probes?
- Do some respondents produce more useful data than others?
- Do different interviewers produce more useful data?

Verbal Probing as a method of Cognitive Interviewing

Within the Verbal Probing technique of cognitive interviewing respondents are asked survey questions, which are followed by a series of probe questions to understand more about their thought processes as they answered the questions. These probe questions can either be asked after each survey question (concurrent probing – sometimes called 'immediate retrospective'), or at the end of the whole survey (retrospective probing) (Willis 2005). The study that is reported here used concurrent probing, within a researcher as investigator paradigm where the researcher is 'guided by intuition, experience and flexibility' (Beatty and Willis 2007) and allowed to adapt probes and produce new ones through the course of an interview.

Four different types of probes used within this technique, with regards to the way that they are *presented by the researcher*: anticipated probes; spontaneous probes; conditional probes (Conrad and Blair, 2009) and emergent probes. For a fuller description of these probe types, see Priede and Farrall (2011) and Beatty and Willis (2007). In addition to this there interviewers also use Functional Remarks (Beatty et al, 1997) - These are remarks uttered by the interviewer in order to keep the respondent talking (for example "ah-ha", "I see" or "that's interesting"). Beatty et al (1997) also outline three categories of probe depending on the *type of information* that the interviewer is seeking from them:

Cognitive Probes- these are probes which centre on people's understandings of questions; the information they drew on when answering the question; the time depth they

considered when answering the question; the level of difficulty they found answering the question; how they interpreted the terms within the question.

Confirmatory Probes- to check that the information given by the respondent is thus far correct.

Expansive Probes- which are used to get additional details from the participant about their experiences, beliefs and morals etc.

Within a cognitive interview all four types of probe may be used from the three categories, along with functional remarks, and may be used in varying quantity depending on the nature of the interview. There is debate as to how much to probe, what types of probe to use, and when to stop probing (Beatty and Willis 2007, Conrad and Blair 2009), and little in the way of generic guidelines currently exist for this.

Advocates of verbal probing suggest that the technique works well as researchers can focus attention on pertinent issues, and thus ask about aspects of the question which they are particularly keen to investigate (Beatty and Willis 2007). It is therefore possible to ask cognitive probes about people's experience of answering the question, and more general thematic questions about their understanding of certain concepts. Despite this it has been suggested that if a researcher is not careful with their use of probes, problems can be created (Wills 2005). In order to avoid this it is necessary for researchers to find a balance in their probing technique between not revealing enough information about respondent's understandings of questions, and over-probing and thus compromising the quality of the information collected by creating false problems.

Cognitive interviewers tend to produce probes in response to individual survey problems, or utilise probes that have been used before successfully, meaning that "Cognitive interviewing approaches evolve somewhat independently across organisations reflecting the preferences in each" (Beatty, Schechter, and Whitaker 1997). Whilst this is not in itself a bad thing it does mean that lessons learned in one context are not transferred to others, and until the publishing of Willis' (2005) book, there was little in the way of a starting point for researchers wishing to undertake cognitive interviews.

In this work Willis also provides some guidance and examples of certain probes for certain questions, in terms of what elements of them require testing. Problems with the questions are revealed, but it is not indicated whether any of the probes are themselves problematic, or how useful they are. He also presents a useful table describing probes which can be used to assess questions for different potential problems, using a question appraisal system where potential problems are revealed.

Willis (2005:115-127) outlines a series of guidelines that he suggests can be used to reduce the amount of false problems that are discovered by verbal probing, that is items which are reported as being problems with the survey questions which are not real. Some of these act as guidelines more generally for the good practice of verbal probing. The items described by Willis are quite loose, and largely serve to guide the researchers towards things to avoid when composing probes. There are, however, few studies that exist which provide the researcher with a list of 'dos' for verbal probing.

Previous Attempts to Explore the Relative Usefulness of Probing Styles

Very few studies have attempted to unpack whether or not certain probing styles produce data which is more useful in developing an understanding of respondents' orientation towards the questions. However, some studies do exist.

Beatty, Schechter and Whitaker (1997) examined how much interviewer behaviour varied between interviewers, and whether this affected the results of the interview. They examined 17 interviews which each consisted of nine survey questions, which were undertaken by four different interviewers (analysis included the work of only three of these interviewers). Concurrent probing was used, and there were scripted probes but interviewers were free to adapt or ignore the suggested probes. The analysis focussed on how interviewers deviated from the script and whether this had any effect on people's responses to CI questions. Analysis looked both at the amount of interviewer activity, and the type of probing used. Mean utterances per interviewer ranged between 34.6 and 37.3, and that the majority of these utterances were either expansive or confirmatory probes. All interviewers used a relatively small number of traditional cognitive probes (means ranged from 2.3 to 5.5 per interview). The number of probes asked per question was also examined. Whilst this study did examine the method of verbal probing in more detail than is usually found, it did not address which of the strategies employed was the most successful in terms of information obtained.

Although limited to the examination of scripted probes Foddy's (1998) study is more thorough. This study explored eight survey questions using a retrospective verbal probing

technique[1]. Foddy – who measured probe effectiveness in terms of the "avowed purpose of the particular probe" (p116) - concluded that 'the least successful probes are also either the least specific or the least direct of the probes used' (Foddy 1998:129). Probes such as 'what were you thinking about?' were less successful than those which asked about specific concepts ('what do you understand by the term 'unwell'?'). Foddy also suggested that less well educated respondents were less able to answer some probes than were the better-educated. Whilst this study is useful, and provides some interesting insights into ways of measuring the success or failure of certain probes there are some issues with the ways that success was measured.

A number of different types of analysis were undertaken on the dataset from the interviews. The information given from the probes was coded using an inductive approach and was coded in terms of the responses mentioned by respondents (so, for example, if they were asked 'which emergency service were you thinking of?' different codes would have been given to fire brigade, police, ambulance), and suggested that one measure of success was the number of individual different answers that were given to a probe. This is not a valid assumption, as a probe may be successful even though it only reveals one or two different answers as these represent the full range that respondents are actually thinking of when answering the question. Foddy also looked at the number of non-responses to

[1] In retrospective CIs, all of the survey questions are delivered and answers elicited. The interviewer then asks the respondent a series of probes about these questions, taking each question or question set in turn.

different probes, and survey questions (so when they answered ‘don’t know’, or could not answer the probe), and whether the probe could either be judged to be successful or unsuccessful in terms of the information given. They also tested probes which asked the respondent to summarise the question in their own words, and examined whether the response categories given were apt. Again, the analysis of this must be questioned, as if a respondent failed to criticise response categories this was coded as ‘probe failure’ rather than ‘question success’.

A more useful technique which was used from this study was the coding and rating of comprehension probes, where they examined how many times particular comprehension probes did the job they were intended for. Foddy suggested that ‘the least successful probes are also either the least specific or the least direct of the probes used’, so probes such as ‘what were you thinking?’ were less successful than those which asked about specific concepts. They found that a similar picture emerged when probes which looked at people’s perspective (broadly similar to ‘expansive probes’ in other studies) when answering questions – more specific probes gave more useful answers. The relationship between the successfulness of probes and education level was also examined and it was discovered that the less well educated respondents were less able to answer both comprehension probes and perspective probes. Foddy does, however state that interviews should not just be undertaken by asking specific probes to clever people, whilst this will reliably get you a large amount of information, it will not necessarily result in all problems being found.

Methods

Our research was conducted as part of a wider project (Euro-JUSTIS, a European-Commission funded project designed to provide EU institutions and Member States with new indicators for assessing public confidence in justice). One key aspect of this work – and which we draw upon herein – was the design and cognitive testing of measures of trust and confidence in the criminal justice system. This work was led by a team based in England and took place in four countries (England, Italy, Bulgaria and Finland)[2] (See Farrall et al 2012). All countries completed 15 to 30 interviews (in line with suggestions from Beatty and Willis, 2007:296). The questions in these interviews were largely attitudinal, examining people's perceptions of fairness in the criminal justice system.

In all, 21 questions were selected for examination. All were translated from English to Finnish by the interviewers prior to being tested. The interviews were carried out using standardised probes, which participants were asked after each item, or battery. Our anticipated probes had been developed by researchers during a previous round of Cognitive Interviewing of related questions (see Farrall et al, 2012). Interviewers were also allowed to

[2] However, herein we rely on data collected from just England and Finland. The process of coding in such detail and checking equivalence of coding decision across all four countries was felt to be too onerous to be conducted on all countries' data. In the UK 25 and in Finland 24 CIs were completed.

use discretionary (emergent and spontaneous) probes to discover more about issues which may arise during the interview process. The interviews were recorded and then analysed. Interviewers were trained social scientists (many to doctoral level) and received additional training in the conduct of cognitive interviews.

Analytic Orientation

Each utterance from the interviewer that prompted a response was coded, using a framework set out by Beatty and Willis (2007). Responses to the probes were also coded in terms of their usefulness. This coding scheme ranged from 1 (not very useful) to 3 (very useful):

1. Not very useful: Little is revealed about how the respondent understood the question or produced their answer. A further probe was required to get the required information.

2. Useful: Sufficient information is provided about how the question/coding scheme was understood and on what information the respondent based their response. The data was of sufficient quality for the researchers to assess whether there was a problem or not with that question and what the cause of the problem may be.

3. Very useful: as 2 above, but with additional insights into the question/coding scheme revealed.

Ensuring Equivalence of Coding

All of the UK information was coded by one researcher. In order to check that the coding was appropriate, a second researcher listened to and scored five per cent of the data. The results of a Cohen's Kappa inter-rater analysis were 0.875 (p= <0.001), indicating a high level of agreement between raters. To ensure that they were coding along similar lines, the two Finnish researchers both coded three interviews, compared results and discussed discrepancies. In certain cases these were discussed with the UK interviewer who was responsible for the development of the coding framework. Although not tested statistically there was a high level of agreement between the Finnish researchers in their coding of probes and rating of responses, and in practice they regularly consulted each other during the coding.

In order to check between the consistency between the coding of the UK and Finnish datasets two separate exercises were carried out. First, one of the Finnish researchers and one of the UK researchers carried out a test rating of a random 10% sample of the UK dataset. The Cohen's Kappa was .811 (p=.000) again suggesting a high level of agreement between raters. As the UK researcher did not have working knowledge of Finnish, it was not possible for the same approach to be taken to the examination of the coding of the Finnish dataset. Instead, the UK researcher rated the usefulness of probes in one randomly selected Finnish interview which had been translated into English. These were compared to the Finnish results and there was a substantial agreement (Cohen's Kappa=.656, p=.000). Thus using a range of triangulation methods, we tested the extent to which our coding across interviews was equivalent, and found that to a large degree it was.

Results and Discussion

In total 2,955 probes were delivered in 49 interviews. The number of probes per interview ranged from 39 and 80, and the interview length between 20 and 50 minutes. The interviews were open and expansive in nature and the length of interviews tended to vary because of the length of discussion about various questions. Similarly the number of probes in a cognitive interview can vary because of the amount of prompting (or clarification) which is required during the interview. In terms of the distribution of probes, it can be seen from Table 1 that anticipated probes dominated.

Table 1: Distribution of Probe Styles

	N	%
Anticipated	1,561	(53)
Spontaneous	10	(<1)
Conditional	237	(8)
Emergent	1,004	(34)
Functional Remarks	143	(5)
TOTAL	2,955	(100)

The key focus of our work was the usefulness of each type of probe. In order to assess this, we coded (see above) the usefulness of the data generated. Table 2 reports on this. As can be seen below, the majority of responses were judged to be 'useful' that is that the probe worked as intended, and revealed enough information without it being in some way groundbreaking or new to the research. The missing values equate to times when a probe was asked which produced no useful information.

Table 2: Usefulness of Probes

	N	%
Not very useful	620	(21)
Useful	2,086	(71)
Very useful	81	(3)
Missing	168	(5)
TOTAL	2,955	(100)

Modelling the Usefulness of Probe-generated Data

Our dataset contains almost 3,000 separate observations on the usefulness of the data generated by the various probes asked. In order to analyse the data, ordinal multi-level modelling was employed since our data on the usefulness of the probes was 'nested' with respondents (n = 49) whilst both probe type and interviewer were used as fixed effects in the model.

Multi-level modelling is a statistical technique employed when one or more observations share some common source. For example, siblings share parents, school-children share classes, schools and education authorities, whilst employees share an employer. Under such conditions the assumptions of independence of observation are violated, and correctly attributing observed variances to the individual (school child) or collective (class) levels becomes impossible. Multi-level modelling solves this conundrum by approaching observations as being nested with in higher levels or groups. In our case, as each respondent had answered several different probes, so answers were nested within probe types which were nested within respondents. Respondents themselves were nested within three different interviewers.

To explore the impact of differing types of probes on the usefulness of the data generated, we estimated a series of increasingly complex models. First we estimated a variance components model (M1), which serves as a useful comparison against subsequent models. This gave us an idea of the percentage of variance in the usefulness of the probes at the individual (i.e. respondent) level. All of our models used "very useful" as the reference category of the dependent variable (hence, strictly speaking we are modelling *un*usefulness).

Following M1, we estimated a model (M2) which includes information about which interviewer carried out the interview (using Interviewer 1 as the reference category). This tested the extent to which different interviewers may (inadvertently via the interviewing style) have been associated with greater levels of usefulness of the probes.

Our third model (M3) added to this characteristics of the respondent (such as their age, gender and educational status) to assess the extent to which such factors influence the usefulness of data produced by the probes. Being male, aged 16-24 and having completed only compulsory schooling were the respective reference categories for these variables. Finally, our fourth model (M4) added to M3 data relating to the nature of the probe itself (whether it was a spontaneous, conditional, emergent or a functional remark), using anticipated probes as the reference category. Table 3 reports on these models.

Table 3: Modelling Usefulness of Data Produced from Different Types of Probes

Variables	M1	M2	M3	M4
Interviewer 2 (ref: Interviewer 1)	-	.411(.249)	.536(.238)	.542(.239)*
Interviewer 3	-	.632(.193)**	.626(.221)**	.607(.222)**
Gender (ref: male)	-	-	.082(.170)	.073(.170)
Aged 25-40 (ref: 16-24)	-	-	.279(.292)	.242(.293)
Aged 41-55	-	-	.016(.274)	-.015(.275)
Aged 56-65	-	-	.194(.304)	.179(.305)
Aged 66 and above	-	-	.664(.617)	.615(.619)
Edn non-compulsory (ref: only compulsory schooling)	-	-	.268(.441)	.272(.443)
Edn 1st degree	-	-	-.172(.418)	-.175(.419)
Spontaneous (ref: anticipated)	-	-	-	1.212(.599)*
Conditional	-	-	-	-.808(.173)***
Emergent	-	-	-	.019(.089)
Functional remark	-	-	-	.230(.187)
% variance	9	7	6	6

* = significant at the $p < .05$ level; ** = significant at the $p < .01$ level; *** = significant at the $p < .001$ level.

The percentage of variance refers to the proportion of variance that is attributable to between respondent differences. For the base model (M1) this is 9% of the variability. The

fact that the inclusion of more terms leads to the proportion of variance to go down in smaller increments is not surprising, and would be the same in any model (if one entered each variable separately, one would be likely to see them each make a larger contribution). Additionally, in M4, the variables entered are measured at the question level (not the respondent level). Given that the inclusion of these will explain variability between questions, it is perfectly possible for the proportion of the remaining unexplained variance attributable to differences between respondents actually to go up a little. In short, despite the decreasing proportion of variance that is attributable to between respondent differences, the models are robust. Taking the inverse log of the coefficient gives us the odds ratio. For the significant variables in our final model this means that spontaneous probes are 3.36 (ln of 1.212) less useful than anticipated probes (although due to the frequency with which these probes were used, there is a high standard error and accordingly a large confidence interval around the estimate for the spontaneous probes). Similarly, conditional probes are a lot more useful than anticipated probes, with the odds of the probe being un-useful more than 50% lower for the conditional probes. The effects of interviewers 2 and 3 when compared to interviewer 1 were similar (1.7 and 1.8 times less useful). What this suggests is:

1) That spontaneous probes were generally less useful than anticipated probes, whilst conditional probes were more useful than anticipated probes.

This is in line with expectations. Firstly, spontaneous probes were asked when an interviewer had a hunch that there was a problem with the respondents' answer to a

question. Whilst such probes may sometimes be useful (as a whole new issue or way of understanding the ways in which a question operates may be revealed), more often than not spontaneous probes do not reveal any new insights. As well as this, the small number of spontaneous probes used may attest to their relative uselessness in the case of these questions- quite simply the interviewers did not ask spontaneous probes, as there was no need to in this case. In cases where a different set of anticipated probes is used, it may be that there is the need to ask more spontaneous probes, which may be found to be of more use.

Conditional probes are likely to be of more use than other probes as they are only asked when a respondent has given a certain answer or exhibited certain behaviour, and therefore are more likely to produce information than anticipated probes which are not tailored to the respondent's individual situation and thus may not be of relevance to them. This is not, however to suggest that other forms of probing should be abandoned in favour of conditional probes, to do so would be to miss the richness and variety of information that comes from a varied probing strategy. It may however be wise, other than in any pilot interviews undertaken, to avoid doing too much in the way of spontaneous probes to maintain the focus of the interview and keep the respondent engaged in the topic in hand.

2) That the characteristics of respondents do not appear to be related to the quality of the data they can provide.

Our model tested the impact of gender, age and education level on the quality of the data generated by the probes. We found that none of these were statistically significant factors in explaining the quality of the data generated by the probes. In other words, all sectors of society were equally likely to provide data of use in understanding how survey questions operated and, hence could be improved (or left unaltered if no problems were found with the question).

3) That one of our interviewers produced more useful data than the other two.

One of the Finnish interviewers produced more useful data than the other two interviewers. The reason for this cannot be found in any difference in the probing strategy employed. It was not that this interviewer used more or fewer, or different probes to the other two. Equally, because the Finnish interviewers coded a mixture of their own and each other's interviews it cannot be suggested that one interviewer was more generous with their rating of usefulness than the others. We cannot explain why the probes posed by this interviewer produced more useful data than those posed by the other two interviewers.

Conclusions

When commencing this study we set out to develop a 'best practice' for the use of verbal probes in cognitive interviews. Through undertaking 49 interviews in two countries we analysed almost 3,000 verbal probes to discover which types of probes were the most useful

and whether certain respondents provide more useful data. Using scripted anticipated probes as the reference, it was discovered that scripted conditional probes were of the most value in terms of the usefulness of the data produced. Spontaneous probes were of least use in this regard, whilst emergent probes and functional remarks were as useful as scripted anticipated probes.

Of course there are some limitations to this study, as there are any studies of this nature. We rely on our own assessment of the usefulness of the probes, and, of course, probes were not 'randomly' assigned to questions. Nevertheless there are still some lessons which can be learnt. Anticipate and Spontaneous probes are best used to probe *for* problems; whilst Conditional and Emergent probes are best used to try to uncover the *reasons* for the problems (and possible solutions). We would thus suggest to researchers carrying out cognitive interviews that they should not stick strictly to scripted probes, but feel free to probe further using emergent probes when they feel that it is useful to gain more information. Regarding respondents for cognitive interviews, there is no one 'type' of respondent which is more useful than any other, therefore it is desirable, as it is with any type of survey piloting, to carry out cognitive interviews with as wide a range of the survey population as possible.

References

Beatty, Paul C, Susan Schechter, and Karen Whitaker. 1997. Variation in Cognitive Interviewer Behavior - Extent and Conseqeunces. Paper read at Proceedings on the section on survey research methods, American Statistical Association

Beatty, Paul C, and Gordon B Willis. 2007. Research Synthesis: The practice of Cognitive Interviewing. *Public Opinion Quarterly* 71 (2):287-311.

Conrad, Frederick G., and Johnny Blair. 2009. Sources of Error in Cognitive Interviews. *Public Opinion Quarterly* 73 (1):32-55.

De Maio, Theresa J, and Ashley Landreth. 2004. Do Different Cognitive Interview Techniques Produce Different Results? In *Methods for Testing and Evaluating Survey Questions*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer. Chichester: Wiley.

Farrall, Stephen, Camilla Priede, Elina Ruuskanen, Anniina Jokinen, Todor Galev, Michela Arcai and Stefano Maffei (2012) Using Cognitive Interviews to Refine Translated Survey Questions: an Example from a Cross-National Crime Survey, *International Journal of Social Research Methodology,* 15(5):467-483.

Foddy, William 1998 An Empirical Evaluation of In-Depth Probes Used to Pretest Survey Questions *Sociological Methods and Research* 27(1) 103-33

Priede, Camilla, and Stephen Farrall 2011 Comparing Results from Different Styles of Cognitive Interviewing: 'Verbal Probing' vs. 'Thinking Aloud'. *International Journal of Social Research Methodology* 12(4) 271-287

Thompson, Mary E. 2008 International surveys: Motives and methodologies *Survey Methodology* 34(2) 131-141

Willis, Gordon B. 2005. Cognitive Interviewing: A Tool for Improving Questionnaire Design. London: Sage.