QUEST 2003: proceedings of the 4th Conference on Questionnaire Evaluation Standards, 21-23 October 2003

Prüfer, Peter (Ed.); Rexroth, Margrit (Ed.); Fowler, Floyd Jackson Jr. (Ed.)

GESIS ZUMA

NACHRICHTEN

Spezial Band 9

QUEST 2003

Proceedings of the 4th Conference on Questionnaire Evaluation Standards

21 – 23 October 2003

Peter Prüfer, Margrit Rexroth,

Floyd Jackson Fowler, Jr. (Eds.)

Zentrum für Umfragen, Methoden und Analysen (ZUMA)
ZUMA ist Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS). Die GESIS ist eine Einrichtung der "Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz" (WGL).

Vorsitzender des Trägervereins ZUMA e.V.: Prof. Dr. Jan v. Deth
Direktor: Prof. Dr. Peter Ph. Mohler

Hausanschrift	**Postanschrift**
B 2, 1	Postfach 12 21 55
68 159 Mannheim	68 072 Mannheim

Telefon	0621/1246 - 227	
Fax	0621/1246 - 100	
E-Mail	pruefer@zuma-mannheim.de	
Internet	http://www.gesis.org/	*GESIS*
	http://www.gesis.org/zuma/	*ZUMA*

ISBN 3-924220-27-1

Druck: Verlag Pfälzische Post GmbH, Neustadt/Weinstraße.
Gedruckt auf chlorfrei gebleichtem Papier.

Contents

Appendix

INTRODUCTION

In 1997, QUEST began. A group of 17 researchers from 8 countries representing 13 organizations gathered in Örebro, Sweden to talk about what the minimum standards for testing questions should be before they are used in a survey.

The meeting in Örebro was convened by three men who constituted the original planning committee: Hans Akkerboom, (then with Statistics Netherlands), Hakan Lindstrom (Statistics Sweden) and Alan Gower (Statistics Canada). The idea that questions should be "pretested" before they were used had been part of survey research practice for decades. The conception of what question testing might include was considerably expanded by publications in 1984 and 1987 that introduced the idea that cognitive testing in laboratories should be added to the traditional field pretest. (Jabine et al., 1984; Hippler, Schwartz, and Sudman, 1987). Government statistical agencies around the world, including those represented on the original planning committee, had begun to do some cognitive testing. However, which questions were tested and how they were tested varied widely between and within statistical agencies. Moreover, within all the statistical agencies, there was disagreement about how important question testing was and whether or not it was justifiable to use time and other resources to cognitively test questions.

The original acronym for the Örebro meeting was proposed to be MIST, which stood for **MI**nimum **S**tandards in Questionnaire **T**esting. The acronym was clearly a stretch, and it did not last, because MIST in German means garbage. However, the name accurately reflected the original goals of the gathering: To discuss what various statistical agencies were doing to test questions, what was known about the value and effectiveness of various approaches to testing questions, and to move toward some agreement about what might be a reasonable set standards for testing questions before using them in surveys.

The first goal, to discuss what agencies were doing, was certainly achieved. Participants learned that activities and practices for instrument testing differed widely. They also learned that everyone at the conference was having difficulty defending the importance of question evaluation, particularly cognitive testing. One critical reason was that there was a complete absence of empirical studies that documented how best to test questions or that cognitive testing did any good at all. For that reason, among others, it was implausible that minimum standards for question testing could be set at that meeting. Instead, what emerged was a shared commitment to begin to do the empirical studies

needed. The participants also agreed to meeting periodically to review how they were doing and exchange ideas.

The group reconvened in London in 1999, in Washington D.C. in 2001, and in Mannheim in 2003. For the London meeting, there were 23 participants from 9 countries and 16 organizations. The cast of participants change some from year to year, but the composition of the group has stayed fairly constant in that:

1. The core of the group consists of members from federal statistical agencies
2. The individuals who attend are themselves personally involved in question evaluation, so they all have first hand experience to share about how to evaluate questions and instruments
3. The group is limited to fewer than 25
4. The organizations represented at the original meeting are all still potentially represented, sometimes by different individuals. “Guests” from other organizations have been included from year to year, but there is a core group of organizations represented with only two additions (from two federal statistical agencies that were not represented in Örebro) since the inception of the meetings.

Perhaps the most important effect of these meetings so far emerged from the London meeting in 1999. At that time, it was concluded that one of the biggest barriers to the spread of effective question testing was the absence of a set of publications. Most of the studies of cognitive testing had been qualitative, and the number of cases involved was usually small. Such studies are hard to publish in statistical journals.

The American Statistical Association (ASA), usually in partnership with numerous other professional groups, has sponsored a set of methodological conferences over the past two decades. Topics have included telephone interviewing, measurement error, nonresponse, and computer-assisted data collection. The QUEST group decided that it was time for such a conference focused on question and instrument evaluation. The purpose of the conference would be to provide a forum for publishing empirical studies that had been done and to stimulate new research on these topics. Jennifer Rothgeb of the United States Bureau of the Census took the lead to put together a planning committee and apply to the ASA for sponsorship. In the fall of 2002, the conference on Question Design, Evaluation and Testing (QDET) was held in Charleston, S. Carolina, USA. A book containing a set of selected papers from the conference will be published by Wiley early in 2004. In this way, one of the highest priorities of the QUEST participants at the initial meeting has been achieved.

The 2003 meeting in Mannheim, hosted by ZUMA, was similar to its predecessors. There were 26 attendees from 9 countries representing 14 organizations. As at previous workshops, attendees made brief, informal presentations of their recent work and interests. Sessions included 2 or 3 presentations, following by extensive discussion. The presentations are not intended to be formal papers. Rather the goal of the presentations is to present ideas and to stimulate thought and discussion. These presentations are reproduced in this volume.

At the end of the workshop, the group held an extended discussion focused on two subjects. First, the group discussed the current state of knowledge about how to do question testing. Second, the group discussed what new research was needed to help inform question and instrument testing. The floor discussion was recorded, and a summary of the main points from those two discussions is presented in the volume after the presentations.

The editors would like to use this opportunity to thank Christa von Briel , ZUMA, for her help in producing this volume. We also wish to thank Patricia Lüder, ZUMA, for her support in planning the conference.

Editors

Peter Prüfer

Margrit Rexroth

Floyd Jackson Fowler, Jr.

References

Jabine, T. B./Straf, M. L./Tanur, J. M./Tourangeau, R. (eds.). 1984: Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press.

Hippler, H. J./Schwarz, N./Sudman, S. (eds.). 1987: Social Information Processing and Survey Methodology. New York: Springer-Verlag.

PARADIGMS OF COGNITIVE INTERVIEWING PRACTICE, AND THEIR IMPLICATIONS FOR DEVELOPING STANDARDS OF BEST PRACTICE

PAUL BEATTY

1. Introduction

Although cognitive interviewing has emerged as a major method to identify and correct problems with survey questionnaires, researchers who employ the method seem to lack consensus on several important points. There does not appear to be a commonly accepted definition of what cognitive interviewing is; we also lack detailed knowledge about what actually happens in cognitive interviews, other than the general notion that people "think out loud" and possibly probe for additional information about the meaning of responses, how people come up with answers, difficulties they have, and so on.

A definition that seems consistent with its most common application is as follows: cognitive interviewing entails administering a draft survey questionnaire while collecting additional verbal information *about* the survey responses, which is used to evaluate the quality of the response, or to help determine whether the question is generating the sort of information that its author intends. But beyond this general categorization, cognitive interviewing potentially includes different activities that may be based on different assumptions about the type of data that is being collected and the role of the interviewer in that process. For example, the verbal material generated by such interviews could consist of respondent elaborations regarding how they constructed their answers, explanations of what they interpret the questions to mean, reports of any difficulties they had answering, or anything else that sheds light on the broader circumstances that their answers were based upon. This material could be based on explicit follow-up questions (or probes) from an interviewer, or based on general instructions to "think out loud" as

much as possible. The interviewer herself could range from a relatively unskilled data collector to an expert investigator; the interview could be based on a scripted protocol, semi-scripted, or largely improvised based on the issues that emerge from discussion. Analysis may be based on systematic review of interview transcripts, or entirely from notes taken during the interview. The various permutations of these activities and their underlying assumptions could lead to quite different products, but all could apparently fall under the rubric of "cognitive interviewing."

Given such variety, it may be difficult to understand what someone means when claiming to have conducted cognitive interviews. Furthermore, a lack of consensus on objectives, procedures, or even general terminology can inhibit methodological developments. The major goal of this paper is to lay the groundwork for continued discussion about cognitive interviewing methodology by reviewing what it is, where it came from, and where it may be going. One potentially useful approach is to consider cognitive interviewing as falling into one of two major paradigms. The first of these paradigms has apparent roots in psychological laboratory methods, while the latter is more strongly rooted in the tradition of "intensive interviewing." After reviewing these paradigms, it will be possible to critically evaluate cognitive interviewing in its various guises – what it accomplishes, what it does not, and some of the key assumptions that its practitioners make when employing the method. The discussion will touch on numerous components of cognitive interview practice, including the role of the interviewer, selection of interviewees, data collection procedures, and evaluation of cognitive interview data.

2. Emergence of Cognitive Interviewing

By the mid 1980s, survey researchers had accumulated considerable knowledge about survey questions, but significant limitations remained – for example, they had limited knowledge about the mechanisms involved in response effects. Specific guidance for writing survey questions was still largely dictated by common sense and the experience of individual researchers. Basic field pretests appeared to be the most commonly used method for evaluating draft survey questions.

A seminar known as the first "CASM" meeting (for Cognitive Aspects of Survey Methodology) assembled survey researchers and cognitive psychologists in 1983 at St. Michaels, Maryland. The influential report from this meeting, *Cognitive Aspects of Survey Methodology: Building a Bridge Across Disciplines* (Jabine/Straf/Tanur/Tourangeau, 1984) proposed a number of important interdisciplinary collaborations. The report also introduced a four-stage model explaining how respondents are likely to answer survey questions: they must comprehend the question, recall relevant information, judge

the appropriateness of the information available to the particular question, and respond in the format provided (Tourangeau, 1984). This model, now widely adopted by methodologists (and expanded in Tourangeau, Rips and Raskinski, 2000), offers considerable help for both researchers and questionnaire designers, who can evaluate how well questions work with regard to each of these components. While these developments were taking place, researchers at the Center for Surveys, Methods, and Analysis (ZUMA) in Mannheim, West Germany, were considering similar ideas and held a conference of their own, with largely new participants. The resulting volume, *Social Information Processing and Survey Methodology* (Hippler/Schwarz/Sudman, 1987), presented some of the first substantive findings from collaborations between survey methodology and cognitive psychology.

Back in the U.S., the National Science Foundation (NSF) had funded several new collaborative research projects in the wake of the CASM meeting. One such project explored how principles of cognitive psychology might be applied in a laboratory setting toward the development and evaluation of questions on the National Health Interview Survey, which is conducted by the National Center for Health Statistics (NCHS) (Lessler/ Tourangeau/Salter, 1989). Additional funds from NSF were used to establish a "cognitive laboratory" at NCHS. Staff of this facility would evaluate and pretest questionnaires on a regular basis, in addition to investigating question-based response errors. Soon thereafter, similar laboratories were established at the Bureau of Labor Statistics and the Census Bureau, (Dippo/Norwood, 1992), to be followed by laboratories in academic and commercial research organizations (Forsyth/Lessler, 1991; Sirken/Schechter, 1999). Cognitive interviewing is the most common activity conducted in these laboratories.

3. Paradigms of Cognitive Interview Practice

It may be useful to distinguish between two general paradigms of cognitive interviewing. One involves a cognitive interviewer whose role is to encourage participants to verbalize their thoughts as they come to mind, but to intervene as little as possible in generating this verbal information. The other involves an interviewer who asks additional, direct questions about responses and who may assume greater responsibility for guiding a discussion about the basis for responses. The former paradigm relies almost entirely upon the think-aloud procedure, in which interviewers encourage participants to verbalize thoughts while answering questions (e.g., "tell me what you are thinking... how are you coming up with your answer to this?"). The latter paradigm relies more heavily on follow-up probes administered after the participant has answered the question (e.g., "can you tell me in your own words what that question was asking?"), although think-alouds may be encouraged to provide supplemental information. Below, I consider the origins and general parameters of both of these paradigms.

The "pure" think-aloud and non-intervening cognitive interviewer

The original paradigm of cognitive interviewing was explicitly psychological. Loftus (1984), elaborating upon ideas presented at the first "CASM" meeting, proposed that a technique known as *protocol analysis* could be adapted as a pretesting methodology for survey questions. The blueprint for this technique was developed by Ericsson/Simon (1980; expanded in 1993) and relies heavily upon *think-aloud* reports. Think-aloud reports were used to yield insights into the thought processes involved in participants' completion of certain tasks in a laboratory setting. Loftus reported that protocol analysis of think-alouds yielded information about how participants tended to retrieve memories of medical visits, and this information was used to develop question wordings that reflected these retrieval strategies. For example, she suggested defining the reference period of recall questions from a past date up to the present, rather than from the present backwards.

Early papers on cognitive laboratory methods (e.g., Royston/Bercini/Sirken/Mingay, 1986) suggest that initial cognitive interviews were based heavily, if not exclusively, upon thinking-aloud. In practice, this meant that cognitive laboratory participants would be asked to report what they were thinking while answering survey questions, and interviewers would simply remind respondents to continue providing such information as necessary. Think-aloud responses were therefore the unique data products of the interviews, and interviewer behavior was constrained accordingly.

An alternative paradigm: interviewers asking direct questions about responses

At some point, an alternative paradigm of cognitive interviewing emerged that expanded upon the use of "pure" think-alouds – in particular, allowing for the addition of direct probing by the interviewer. Apparently, the distinction between true think-aloud interviews and "intensive interviews" (which had been used to evaluate questionnaires prior to the CASM meetings – see DeMaio, 1983) became blurred, with both eventually falling under the header of "cognitive interviewing." It is easy to imagine how this could have occurred, especially if the intensive probes used could be construed as "cognitive" – addressing how terms were interpreted, how participants remembered certain facts, whether answers fit into available response categories, and so on.

This paradigm seems to have emerged gradually. Most descriptions of cognitive interviewing in the early 1990s (Forsyth/Lessler, 1991; Bercini, 1992) focus on think-alouds as the dominant component of cognitive interviewing, mentioning the possibility of probing for supplemental purposes – although Willis/Royston/Bercini (1991) suggest that both think-alouds and probing could both be viable alternatives. Later, Willis (1994)

proposed putting a greater emphasis on probing. Several other articles suggest that the trend toward acceptance of such activities continued: Gerber/Wellens (1997), for example, noted that cognitive interviewing had seemed to evolve from its original form to include "more probes and probes about meaning than was originally intended" (p. 35). Willis/ DeMaio/Harris-Kojetin (1999), noting that cognitive interviews are often called "think-aloud interviews," recommended that the latter term should be used more sparingly because think-aloud protocols were not necessarily the dominant component of cognitive interviews as currently practiced. And, Beatty (in press) concluded that cognitive interviews conducted in a study at NCHS relied heavily on interviewer adaptation to particulars that emerged in individual interviews.

Taken together, these works suggest that some practitioners of cognitive interviewing adopted a paradigm allowing for the collection of additional verbal material other than "pure" think-alouds, and at least in some cases, empowering the interviewer to guide the content of interviews. This is not to say that the use of think-alouds was abandoned, since virtually all descriptions of cognitive interviewing mention think-alouds as one possible component. It is also not to say that this paradigm completely replaced the alternative one firmly rooted in think-alouds, as Conrad/Blair/Tracy (2000), among others, clearly favor the original approach. Rather, another paradigm emerged, one that in practice owed relatively little allegiance to the procedures for verbal protocol analysis proposed by Ericsson and Simon.

The goal under both paradigms is to generate verbal information that is usually unseen in a survey interview in order to evaluate how well the questions are meeting their objectives. This puts both paradigms on an important common ground. Yet they are carried out differently and are based on some different assumptions, which may have implications regarding the data that they generate. It is therefore worthwhile to consider the rationales offered for both paradigms.

An assessment of the two paradigms

Advocates of the original paradigm propose that is has several advantages. One advantage proposed by Conrad/Blair/Tracy (2000) is that relying only on thinking-aloud avoids problems with artificiality that can arise due to interviewer probing. It is virtually indisputable that inserting probes into the middle of a survey interview alters the content and flow of the interaction. Such alteration of the realistic flow is the major reason why Oksenberg/Cannell/Kalton (1991) proposed that probing should be used after only a few questions per interview.

Forsyth/Lessler (1991) and van der Veer/Hak/Jansen (2002) are among those who propose an additional advantage: that think-aloud data are preferable because they are collected *during* the response process, and therefore have a certain purity that probe responses (provided *after* responding) do not. However, there is a considerable body of research beginning with Nisbett/Wilson (1977) that calls into question whether think-alouds are literal reflections of actual thought processes. More likely, they are re-constructions created after the fact. For the most part, they are likely to be reasonable reflections of actual processes (Wilson/LaFleur/Anderson, 1996) – but not necessarily literal representations. Furthermore, as Willis (in press) notes, Ericsson/Simon (1980) themselves did not insist upon exclusive use of thinking-aloud: their crucial point was that self-reported information should be in *short-term memory* (as opposed to long-term). From that perspective, reports based on probes immediately following questions are probably not much different than think-aloud reports. Both are approximations of "the real thing."

Advocates of the more probing-centered paradigm suggest that the alternative has particular advantages as well. One is that probing may provide necessary focus to cognitive interview interactions. Willis (1994) suggests that thinking aloud often leads participants to diverge onto irrelevant tangents. The most efficient way to correct this problem is through probes selected to re-focus attention on pertinent issues. Of course, doing so requires interviewer judgment. This is important, because it is not the use of probes per se that regain control of the interview, but an interviewer skilled at using the "right" probes. What Willis is really advocating is interviewer discretion in guiding the interview content.

Another potential advantage of probing is that it may have less of an impact on the response process than thinking-aloud. Although the actual content of probe responses is probably quite similar to that of think-alouds (see above), procedures for obtaining them are different; in the think-aloud case, participants at least attempt to provide some verbal information prior to responding to the question. According to Ericsson/Simon (1980), thinking-aloud should not interfere with the response process. However, Russo/Johnson/Stephens (1989) found that thinking aloud did have an impact on the accuracy of various mental computations; furthermore, Willis (1994) argues that thinking aloud increases the effort spent on creating a response, which has an unknown impact on the response process.

Perhaps the strongest justification for the more probe-based paradigm is that it generates verbal material that questionnaire designers find useful, but that may not emerge unless a cognitive interviewer specifically asks for it. As Willis (in press) observes, think-aloud

procedures were originally proposed by Loftus (1984) to shed light specifically on *retrieval* processes. However, the cognitive model proposed by Tourangeau (1984) also addresses comprehension, estimation/judgement strategies, and selection of particular responses. Even if think-alouds are effective at illuminating retrieval processes, they may not consistently generate information about these other issues. For example, participants might not explicitly consider the meanings of words or phrases while answering questions, and might not recognize their own misunderstandings. Also, Conrad/Blair/ Tracy (2000) note that think-alouds alone sometimes suggest a problem with a question without providing enough information to diagnose what the problem is. Probe responses might help to fill in this gap.

Today, the original distinction between the paradigms (thinking-aloud vs. probing) is probably not of central importance. Advocates of the original paradigm (Conrad/Blair/ Tracy, 2000) have conceded that probing makes important contributions, and advocates of the alternative (Beatty, in press) have acknowledged that probing can shape interview content in some undesirable ways. Most conceptualizations of cognitive interviewing today include some degree of probing. The primary distinction is now between an *unobtrusive* cognitive interviewer, who relies on standardized think-aloud protocols and possibly scripted probes, and an *active* cognitive interviewer, who is given more latitude to explore topics as they emerge within interviews. The practical decision has moved from whether or not to allow probes, to *how much* probing is appropriate, whether this probing should be standardized or determined by interviewer judgment (or to what extent), and how researchers should select the most appropriate probes for various purposes. The remainder of this paper considers how researchers might make these and other decisions in applying cognitive interviewing methodology.

4. Toward More Specific Practical Guildelines

While Forsyth/Lessler (1991), Willis (1994), and DeMaio/Rothgeb (1996), among others, have contributed significantly to establishing general parameters of cognitive interviewing, the literature is almost silent about many specifics regarding the design, implementation, and analysis of studies based upon this method. Guidance regarding the probing decisions mentioned above would be useful, as would guidance about the ideal background and training of interviewers, how many interviews are required to adequately test a questionnaire, and how participants should be selected.

Cognitive interviewers as data collectors vs. investigators

Tucker (1997) – in a position largely consistent with the original paradigm discussed earlier – calls for much greater standardization of cognitive interview procedures. Without this, he argues that "effective manipulation [of variables] will be impossible... the notion of falsifiability has no meaning... [and] the conditions necessary for generalizability will be absent" (p. 72). Conrad/Blair (1996) similarly argue that "rigorous experimental methods that rely on objective quantifiable data" are preferable to "largely impressionistic methods" that they suggest are generally used in cognitive laboratories (p. 8). Under this perspective, creative contributions from interviewers that lead to non-standardized behavior are undesirable. Given that interviewers would be constrained against improvising, the investigative burden is clearly on the front-end, meaning that researchers would need to determine in advance any issue they wished to probe about.

An alternative perspective is that interviewers themselves may serve as investigators. For example, Willis (1994) compares cognitive interviewers to "detectives" who rely at least partially upon improvisation in looking for clues about questionnaire problems. In subsequent work, he draws an analogy between cognitive and clinical interviews, which may be guided by intuition, experience, and flexibility (Willis, in press). This perspective forgoes consistency across interviews in favor of freedom to explore issues that emerge in discussions with participants. Presumably its major advantage is that it allows interviewers to explore issues that emerge within the interview, were not anticipated in advance, and might be missed through more scripted interviews. While this perspective does not preclude identifying some issues to be watchful for in advance, it does place considerable trust in the interviewer's ability to notice potential problems and to conduct *emergent probing* in ways that shed light on the sources of these problems. Thus the interviewer takes on some of the role of investigator as well as data collector.

These two perspectives might call for very different sorts of cognitive interviewers, with potentially different skills, backgrounds, and training. The skills necessary for data-collection cognitive interviewers might not be much different than those of survey interviewers – e.g., they would require training in general procedures, but not in the subject matter being investigated (Fowler/Mangione, 1990). They would not have to know why think-alouds or probes were being administered – only to recognize when participants were providing adequate think-aloud or probe responses.

For investigative cognitive interviewers, such skills would be necessary but not sufficient, since they would at least partially determine the content of the interview. In doing so, such interviewers might need to draw upon knowledge of the objectives of the questions, potential types of cognitive or communicative errors that could affect the accuracy of

survey responses, and familiarity with various options for eliciting useful verbal material from participants. Strangely, the literature on cognitive interviewing does not seem to address the appropriate background of such cognitive interviewers. A solid grounding in survey methodology would probably be useful, since interviewers would generally be working in an environment geared toward producing survey statistics, and presenting their conclusions to professionals in that field. It would be useful for advocates of this type of cognitive interviewing to think clearly about what other skills or background are most desirable for identifying effective interviewers.

What to ask: the selection of probes

As discussed earlier, most recent conceptualizations of cognitive interviewing involve probing to some degree. If the interviewer is also an investigator, then she may select some of these probes herself; if a data collector, then the probes may be selected for her. But either way, someone must choose what probes are used. Although cognitive interviewing literature provides many examples of possible probes, it provides little guidance regarding which probes are likely to be most effective for various purposes. A few basic guidelines are available: for example, Willis (1994) notes that probes should not suggest a "correct" answer, a principle that also applies to survey questions. Foddy (1998) concludes that specific probes such as "what does [term] mean to you?" are more effective than general ones such as "what were you thinking when you first answered the question?" In another recent study, Beatty (2002) found that participants answered probes about the meaning of terms differently when they were administered alone than they did within the context of a particular survey question. However, these sort of recommendations appear to be uncommon, and almost completely missing from published literature in this area.

Cognitive interviewers may be able to obtain some guidance about how to choose "good" probes from literature on qualitative interviewing, which may include lessons on what to ask, how to ask it, and how to make sense of narrative data. For example, Weiss (1994) suggests that interviewers generate narrative by asking about specific events rather than generalized experience. Holstein and Gubrium (1995) encourage interviewers to be on the lookout for "confusion, contradictions, ambiguity and reluctance" as signs that "meanings are being examined, reconstituted, or resisted" (p. 79). In the case of cognitive interviews, such instances might call for additional probing. Variants of qualitative interviewing are also employed by anthropologists, and some guidance may be obtainable from that field as well. For example, Gerber (1999) notes that anthropologists might explore whether terms are "culturally inappropriate" for a particular population. But rather than simply asking a participant what a term such as *self-reliance* means, an anthropologist might

explore its meaning in different contexts, e.g., with regard to child rearing, older family members, or welfare recipients. This might suggest that general cognitive interview probes such as "what does this term mean to you?" might be less effective than specific ones exploring how a term is used in a participant's life.

Who to interview, and how much interviewing to do

Cognitive interviewing literature pays even less attention to issues of how to select samples of participants and how to determine when an adequate number of interviews have been completed. The lack of attention to this issue is likely a consequence of its association with psychological laboratory methods, which have often placed little emphasis on such matters. Cognitive interview practitioners generally acknowledge that participants are chosen by convenience and that such samples are "not designed to be representative [of any larger population], but to reflect the detailed thoughts and problems of the few respondents who participate in [cognitive interviews]" (DeMaio/Mathiowetz/Rothgeb/Beach/Durant, 1993).

Other than that, the only specific guidance that seems to be available is that some demographic variety of respondents is desirable, and that participants should include people relevant to the topic of the questionnaire being tested (Willis, 1994). One clear consequence of such sampling is that cognitive interviewing can never determine the extent of questionnaire problems in a population. Still, some sampling considerations could help to strengthen claims that a reasonably thorough effort has been made to identify the most pressing problems with a questionnaire.

For example, participants could be selected to cover as much of a questionnaire's conceptual terrain as possible. If questionnaires include skip patterns that lead to various branches, the sample should be sufficiently diverse to explore all of these different paths. Whatever topic the questions focus on (e.g., health insurance), the sample should cover a variety of circumstances relevant to that topic (e.g., people with a variety of health insurance situations). Within those parameters, it also seems desirable to select participants representing some demographic variety. Practitioners should not operate under the illusion that such diversity ensures "representativeness"; it only casts a wider net over varying circumstances, maximizing the chances that discovery will be effective. Similarly, interviewing in multiple locations could improve the variety of circumstances that are captured in testing.

As for what constitutes an adequate cognitive interview sample size, little guidance has been offered on this point either – often, literature in this area simply acknowledges that samples are small. Several researchers report that cognitive interviews are commonly

divided among several “rounds” consisting of about 10 interviews each (Willis, 1994; McColl, 2001). Such small rounds of interviews are considered sufficient to identify some questionnaire problems, at which point questionnaires can be revised and tested again in subsequent rounds. This iterative approach seems useful, but it still leaves open the question of whether researchers can determine when they have conducted enough rounds of interviews to stop the process. Some qualitative researchers make decisions regarding when enough interviews have been conducted based on the idea of *category saturation* (Strauss/Corbin, 1990). Put simply, this means that the researcher identifies groups of people most relevant to the study and conducts interviews with members of each until they yield relatively few new insights. In other words, operating under a principle of *diminishing returns* may be effective. Operating in this manner makes a very important assumption: that the most critical questionnaire problems will be revealed quickly from virtually any group of relevant participants. This assumption is most likely to hold up if participants reflect a range of experiences that a question attempts to measure, and also represent at least an attempt to obtain some demographic diversity. These are not guarantees of representativeness. Rather, they are guidelines that maximize the chances of discovering potential questionnaire problems as efficiently as possible. Still, greater attention to how participants are selected and how many of them should be interviewed could maximize cognitive interviewing’s potential to quickly home in on the most significant problems with a particular questionnaire.

Evaluating evidence from cognitive interviews

Whether cognitive interviews are conducted based on a fairly standardized protocol or with greater interviewer flexibility, the result is still verbal text that needs to be evaluated to determine whether or not a question poses a problem for respondents. One advantage of fairly standardized protocols is that they allow for more systematic analysis. For example, Conrad/Blair (1996) propose that verbal protocols be coded in a table with “types of problems” on one axis (lexical, temporal, logical, etc.), and “response stage” (understanding, task performance, and response formatting) on the other. Whenever problems were observed answering a question, they would be coded in the appropriate category. Of course, the success of this procedure (and others like it) is based on the assumption that the interviewing technique brings cognitive problems to the surface so that they can be observed. It also assumes that an analyst will be able to make enough sense of these verbalizations to code them appropriately.

As we have seen, some may counter-argue that additional probing from a skilled investigator brings enough additional material to the surface to justify the lack of standardization – that is, the above procedure would miss some important observations.

The resulting data may be harder to interpret and it may be harder to judge whether a problem is "real." One possibility for evaluating questionnaire problems is that attempts should be made to link them to characteristics of the question. For example, consider the survey question "Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?" Schechter/Beatty/Willis (1998) concluded that respondents might have a difficult time answering this. The conclusion can be backed up by the following process of reasoning:

- *observation of a problem*: several participants could not provide codeable responses (a number between zero and thirty), even when probed.
- *studying the specifics of the problem*: some participants indicated that the question did not allow them to answer in meaningful terms ("a day is part good and part bad – you can't characterize it as one or the other"); others complained about response task difficulty ("I don't do bookkeeping on this"), especially given complicated health status.
- *identifying the question characteristic that is the source of the problem*: the question is based on the assumption that a "day" is a reasonable metric, but it may not be for people with varying-quality days.
- *determining whether this is this generalizable*: it seems reasonable that this problem could recur, e.g., for people who experience health problems that bother them at various times during the day, or people with multiple intermittent health problems.

Claiming that this process found "proof" of the problem would be overstating the evidence. However, a reasonable case could be made that the problem is likely to be found in respondents with similar circumstances, and is created by a faulty assumption about the way individuals think about their health. Note also that the evidence is not linked to the *number* of participants who report a particular problem. Whether it takes many or a few participants to construct such an argument, it needs to be evaluated based on logical merits. It is conceivable that a solid argument about a questionnaire problem could be constructed around a single case, or that such an argument might fail to materialize around several.

Error in cognitive interview analysis

Still, since the method does rely heavily upon human judgment, the possibility that cognitive interviewing could lead to conclusions that are incomplete, misleading, or incorrect must be addressed. There are several possibilities for error: cognitive interviews could identify problems that would not turn out to be "real" in surveys; cognitive interviews could fail to identify problems that exist in actual survey administration; and, cognitive interview findings might be inconsistent when conducted by independent

groups of researchers. The first two could be considered problems with the validity of the method and might be respectively classified as *errors of commission* and *errors of omission*. The third could be considered to be a problem with the reliability of the method.

Practitioners of cognitive interviewing make several assumptions to defend themselves against the possibility of errors of commission. One is that *cognitive interviewing finds problems that will carry over to actual surveys.* Unfortunately, there is often no obvious way to verify that hypothesized problems are "real." Logical arguments may have to do, and researchers will have to determine for themselves whether they find such arguments to be meritorious. However, there has been at least one attempt to verify that cognitive interview findings were borne out by field data (Willis/Schechter, 1997). They revised questions based on cognitive interview findings, and the results were more plausible than statistics from the original question – however, such efforts are expensive and the results are not always conclusive.

Another potential concern is that cognitive interviewing could *fail* to find problems that would actually appear in survey interviews. Actually, there is no reasonable way that cognitive interview practitioners could claim to have found all problems with a questionnaire. Its usefulness is based on the assumption that the most egregious problems will become evident in most groups of participants who are reasonably appropriate to the topic of the survey. Interviewing often concludes based on a subjective judgment that interviews are yielding diminishing returns. However, there is always the possibility that one additional interview could yield a significant new insight, or that an additional interviewer would be more likely to notice additional problems. By the same token, claims that a questionnaire has "no problems" are impossible – the strongest claim that could be made is that no problems have (yet) been discovered.

Finally, there is always the possibility that independent groups of cognitive interview practitioners might not reach the same conclusions about a questionnaire. However, it would probably not be unusual for different groups to discover different insights, especially if interviewers were operating under an "investigator" paradigm. Differences could be a function of different interviewer backgrounds and sensitivities to various sorts of problems. Hopefully, however, the findings of different groups would not be wildly incompatible (e.g., with one group finding a series of comprehension and recall problems, and another finding no problems even after many interviews).

On a related note, one potential view of cognitive interviewing is that its objective is to identify and eliminate all of the problems with a question, and that by doing so we can find the "ideal" question to ask. However, an alternative view is that researchers could

reach very different conclusions about the quality of a certain question, both of which could be correct. Rather than attempting to find the "right" way to ask a survey question, cognitive interviewing may be more suited to helping researchers assess the advantages and disadvantages of asking questions in a certain manner – and helping researchers make decisions based on the potential errors they are more comfortable with. Questionnaire designers actually make these decisions all the time (e.g., adding verbiage to a question might clarify its intent, but doing so could also make it longer and more burdensome). It may be that cognitive interviewing proves to be useful simply because it provides information to make such design decisions as logically as possible, some of which might not be available through other pretesting methods. In other words, cognitive interviewing may be less suited to finding the "best" questions than guiding "best informed" design decisions.

It seems unlikely that cognitive interviewing will generate "reliable" findings in the sense that survey researchers might use the term (i.e., each set of interviews identifies the same set of problems with questions). When findings are different, yet not necessarily contradictory, this may indicate that no one set of findings is complete – the different findings should be examined to see whether they complement or refute each other. Findings that are difficult to reconcile might indicate either faulty reasoning by analysts, or that interviewing has not yet yielded an adequate understanding of responses associated with a question. The former case calls for a closer look at the data, while the latter indicates a need for continued data collection.

5. Conclusions and overall assessment

Since its inception in the mid 1980s, cognitive interviewing has become a prominent method for survey questionnaire development and evaluation. It should be clear from the preceding review that cognitive interviewing is not so much one clearly defined method as a loose collection of several potential activities. These activities have a great deal in common – all involve the collection of verbal material beyond a simple survey response, which is used to evaluate whether questions are capturing information as intended. Yet there are also important differences in both philosophy and practice.

Reasonable arguments have been offered for both the data-collector and investigator paradigms of cognitive interviewing. Clearly both positions offer some advantages. Both also have shortcomings that could be addressed more thoroughly by their respective advocates. Discussions about best practices need to continue. With such continued discussions, researchers should be better equipped to use cognitive interviewing to help create questions that are clear, pose memory and recall tasks that respondents can

reasonably be expected to accomplish, and that allow respondents to express their answers accurately.

References

Beatty, P., 2002: Cognitive Interview Evaluation of the Blood Donor History Screening Questionnaire. In Final Report of the AABB Task Force to Redesign the Blood Donor Screening Questionnaire. Report submitted to the U.S. Food and Drug Administration.

Beatty, P., (In press): The Dynamics of Cognitive Interviewing. In S. Presser/J. Rothgeb/ M. Couper/J. Lessler/E. Martin/J. Martin/E. Singer (eds.), Questionnaire Evaluation and Testing Methods Development. New York: John Wiley and Sons.

Bercini, D.H., 1992: Pretesting Questionnaires in the Laboratory: An Alternative Approach. Journal of Exposure Analysis and Environmental Epidemiology, 2, 241-248.

Conrad, F./Blair, J., 1996: From Impressions to Data: Increasing the Objectivity of Cognitive Interviews. 1996 Proceedings of the Section on Survey Research Methods, Volume 1, pp. 1-9. Alexandria, VA: American Statistical Association.

Conrad, F./Blair, J./Tracy, E., 2000: Verbal Reports Are Data! A Theoretical Approach to Cognitive Interviews. In Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference. Washington, D.C.: Office of Management and Budget.

DeMaio, T.J. et al., 1983: Approaches to Developing Questionnaires. Statistical Policy Working Paper 10. Washington, D.C.: Statistical Policy Office, U.S. Office of Management and Budget.

DeMaio, T.J./Mathiowetz, N./Rothgeb, J./Beach, M.E./Durant, S., 1993: Protocol for Pretesting Demographic Surveys at the Census Bureau. Washington, D.C.: U.S. Bureau of the Census.

DeMaio, T.J./Rothgeb, J., 1996: Cognitive Interviewing Techniques: In the Lab and in the Field. In N. Schwarz/S. Sudman (eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Dippo, C.S./Norwood, J.L., 1992: A Review of Research at the Bureau of Labor Statistics. In J. Tanur (ed.), Questions about Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage.

Ericsson, K.A./Simon, H.A., 1980: Verbal Reports as Data. Psychological Review, 87, 215-251.

Ericsson, K.A./Simon, H.A., 1993: Protocol Analysis: Verbal Reports as Data. Cambridge, MA: MIT Press.

Foddy, W., 1998: An Empirical Evaluation of In-Depth Probes Used to Pretest Survey Questions. Sociological Methods and Research, 27, 103-133.

Forsyth, B.H./Lessler, J.T., 1991: Cognitive Laboratory Methods: A Taxonomy. In P. Biemer/R.M. Groves/L. Lyberg/N. Mathiowetz/S. Sudman (eds.), Measurement Error in Surveys. New York: Wiley.

Fowler, F.J./Mangione, T.W., 1990: Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, CA: Sage.

Gerber, E.R., 1999: The View from Anthropology: Ethnography and the Cognitive Interview. In M. Sirken/D. Herrmann/S. Schechter/N. Schwarz/J. Tanur/R. Tourangeau (eds.), Cognition and Survey Research. New York: Wiley.

Gerber, E.R./Wellens, T.R., 1997: Perspectives on Pretesting: 'Cognition' in the Cognitive Interview? Bulletin de Methodologie Sociologique, 11, 18-39.

Hippler, H./Schwarz, N./Sudman, S. (eds.), 1987: Social Information Processing and Survey Methodology. New York: Springer-Verlag.

Holstein, J.A./Gubrium, J.F., 1995: The Active Interview. Thousand Oaks, CA: Sage.

Hyman, H.H./Associates, 1975 [1954]: Interviewing in Social Research. Chicago: University of Chicago Press.

Jabine, T.B./Straf, M.L./Tanur, J.M./Tourangeau, R. (eds.), 1984: Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press.

Lessler, J.T./Tourangeau, R./Salter, W.: Questionnaire Design in the Cognitive Research Laboratory. Vital and Health Statistics, Series 6, No. 1. Hyattsville, MD: National Center for Health Statistics.

Loftus, E.F., 1984: Protocol Analysis of Responses to Survey Recall Questions. In T. Jabine/M. Straf/J. Tanur/R. Tourangeau (eds.), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press.

McColl, E., 2001: Protocol for Cognitive Testing of Global Health Status Rating Items. Unpublished manuscript.

Nisbett, R.E./Wilson, T.D., 1977: Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84, 231-259.

Oksenberg, L./Cannell, C.F./Kalton, G., 1991: New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349-365.

Royston, P.N./Bercini, D./Sirken, M./Mingay, D., 1986: Questionnaire Design Research Laboratory. 1986 Proceedings of the Section on Survey Research Methods, pp. 703-707. Alexandria, VA: American Statistical Association.

Russo, J.,/Johnson, E.,/Stephens, D., 1989: The Validity of Verbal Protocols. Memory and Cognition 17: 759-769.

Schechter, S./Beatty, P./Willis, G.B., 1998: Asking Survey Respondents About Health Status: Judgment and Response Issues. In N. Schwarz/D. Park/B. Knäuper/S. Sudman (eds.), Cognition, Aging, and Self-Reports. Philadelphia, PA: Psychology Press.

Sirken, M./Schechter, S., 1999: Interdisciplinary Survey Methods Research. In M. Sirken/D. Herrmann/S. Schechter/N. Schwarz/J. Tanur/R. Tourangeau (eds.), Cognition and Survey Research. New York: Wiley.

Strauss, A./Corbin, J., 1990: Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Newbury Park, CA: Sage.

Tourangeau, R., 1984: Cognitive Sciences and Survey Methods. In T. Jabine/M. Straf/J. Tanur/R. Tourangeau (eds.), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press.

Tourangeau, R./Rips, L./Rasinski, K., 2000 : The Psychology of Survey Response. Cambridge: Cambridge University Press.

Tucker, C., 1997: Methodological Issues Surrounding the Application of Cognitive Psychology in Survey Research. Bulletin de Methodologie Sociologique, 11, 67-92.

van der Veer, K./Hak, T./Jansen, H., 2002: The Three-Step Test Interview (TSTI): An Observational Instrument for Pre-testing Self-Completion Questionnaires. Paper presented at the Questionnaire Development, Evaluation, and Testing Conference in Charleston, South Carolina, November 2002.

Weiss, R.S., 1994: Learning from Strangers: The Art and Method of Qualitative Interviewing Studies. New York: The Free Press.

Willis, G., 1994: Cognitive Interviewing and Questionnaire Design: A Training Manual. Cognitive Methods Staff Working Paper Series, No. 7. Hyattsville, MD: National Center for Health Statistics.

Willis, G. (in press): Cognitive Interviewing Revisited: A Useful Technique, in Theory? In S. Presser/J. Rothgeb/M. Couper/J. Lessler/E. Martin/J. Martin/E. Singer (eds.), Questionnaire Development Evaluation and Testing Methods. New York: John Wiley and Sons.

Willis, G./DeMaio, T./Harris-Kojetin, B., 1999: Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques. In M. Sirken/D. Herrmann/S. Schechter/N. Schwarz/J. Tanur/R. Tourangeau (eds.), Cognition and Survey Research. New York: Wiley.

Willis, G.B./Royston, P.N./Bercini, D., 1991: The Use of Verbal Report Methods in the Applied Cognitive Laboratory. Applied Cognitive Psychology, 5, 251-267.

Willis, G./Schechter, S., 1997: Evaluation of Cognitive Interviewing Techniques: Do the Results Generalize to the Field? Bulletin de Methodologie Sociologique, 11, 40-66.

Wilson, T./LaFleur, S./Anderson, D., 1996: The Validity and Consequences of Verbal Reports About Attitudes. In N. Schwarz/S. Sudman (eds.), Answering Questions: Methodology for Determining Cognitive Processes in Survey Research. San Francisco, CA: Jossey-Bass.

Contact

Paul Beatty
National Center for Health Statistics
3311 Toledo Road
U.S.A. - Hyatsville, MD 20782
U.S.A.
email: pbb5@cdc.gov

COGNITIVE MODEL OF THE QUESTION-ANSWERING PROCESS AND DEVELOPMENT OF PRETESTING

ANJA AHOLA

Our testing services are mainly used by Statistics Finland's Social Statistics unit and by external social research institutes. It is, therefore, important to develop our testing activities so as to be of interest to different customer groups. Finnish social scientists are not very interested in the cognitive theory when interpreting test results, and have criticised the theory underlying the testing (cf. Ahola & Lehtinen & Godenhjelm 2002). My presentation discusses the two main criticisms as well as methods by which our results could be made more interesting for our clients.

Two critical comments

First, the cognitive model in which the question-answering process is seen as a four-phase process (comprehension–retrieval–judgment–response) should be enhanced by taking better into account the ways respondents of different subsets interpret the reasons for asking specific questions. People do not reply to any questions without thinking why they were asked, what the answers will be used for and who will be using them.

Although questionnaires must be pretested, there will always be questions that all people do not quite understand, or they will want to know why a particular question has been asked: what exactly are the researchers after. Theoretically, all this discussion about answers is disallowed because it does not take place identically in each interview, and there are personal differences between interviewers. However, in practice it is impossible to eliminate the element of human interaction that is present in the filling of a questionnaire (e.g. Alasuutari 1998).

The fact that one is responding to a survey is not a sufficient interpretive frame for answering questions. Different questions evoke different frames within which answers are given and these may also vary from one individual and population group to another.

People give their answers in terms of these frames. Likewise, because people believe that surveys may serve them or the society at large, for instance by conveying messages to decision-makers, they may respond in ways which best serve their interests.

The second criticism is that as well as for a cognitive analysis, interviewing also calls for a socio-cultural analysis, in terms of what certain conversation means in a given situation (Ronkainen 2002). The interviewee's categorisation of the situational interaction will profoundly influence what subjects will be addressed, how much information can be given, how many personal secrets should be revealed, what speech forms may be used, etc. The way respondents frame the event will significantly affect their interpretation of the questions and thus the nature of the answers.

According to the critics, some interviewees find the speech community of a survey interview (cf. Briggs: interview as a communicative event) an outstandingly good mode of talking because it allows them to say things that in a deep interviewing situation would require an entirely different vocabulary and trust relationship. The question then arises of what possibilities the survey interview situation opens for the respondent and what possibilities it excludes. There is no such speech community that would not in one way or other open and close options on what can be said and how. "If I tell a friend about my heartaches, I have certain options for doing it, but acceptance of this discoursive world means that my friend then has to respond in a certain manner, thereby encouraging me to pour out my heart. If I were to talk about the same thing at the pub, I would inevitably do it mockingly, giving details of all the partner relationship therapies we have undergone, for example. Neither account of these is inferior in terms of authenticity. For me, as the person telling them, they represent different ways of reflecting upon my experience" (Ronkainen 2002).

What is then the nature of the information produced in the survey interview situation?

Received comments and previous workshop discussion

I will next try to locate the received comments into the framework of the previous discussion at the last workshop. At the last QUEST Workshop, Elizabeth Martin (2001) described three theoretical models or approaches that she believed underlie the understanding of the question-answering process. She identified also the issues and aspects of survey questions that are implied by each theoretical perspective. Finally, she summarised whether the question evaluation methods allow us to deal with these issues.

Elizabeth Martin organised the theoretical perspectives into the question-answering process into three models or approaches: the model of the standardised survey interview,

the model of the interview as social interaction or discourse and the cognitive process model. The different approaches set different kinds of objectives for the evaluation of survey questions. According to Martin, today's pretesting methods best answer questions produced with the standardised survey interview model and the cognitive process model, although the evaluation of certain stages of the cognitive process model still deserves more theoretical consideration. Martin suggests that today's pretesting methods produce poor answers to questions generated by the model of the interview as social interaction.

According to Martin the approach considering the interview as social interaction is not a uniform theoretical perspective, but a sum of many perspectives focused on trying to understand the social context of an interview. Whereas older theories see the context as an error source in survey responses, newer theories do not analyse social interaction this way but instead see it as a resource for mutual understanding. The building of a meaning in an interview has been viewed as social discourse rather than as a cognitive process. The two comments presented above associate with the same questions that the newest theories concerning the social interaction and discourse of surveys bring to the fore.

Could test data be used from various perspectives?

We use focus groups interviews, cognitive interviews and systematic questionnaire evaluations to study the data collection tools, terminology, classifications and background concepts. Our commonest task is testing of draft questionnaires. We design the tests and interpret the results using the cognitive model and theory. Our cognitive testing still needs further development that we can bring the model to life and ask good test questions (cf. Martin 2001). The timetables for analysing responses to interviews conducted for the purpose of improving data collection tools are usually very tight, which is why we only apply the cognitive model in their interpretation.

However, besides as pretesting data, cognitive interview data can also be analysed by understanding the question-answering process as social interaction. Cognitive test interviews can also generate a lot of "superfluous" discourse as, for example, happened with interviews conducted for the purpose of developing the measurement of values. In certain projects we also try to collect such data that can be analysed from the perspective of social interaction and discourse. In these cases we attach to the test interviews additional elements allowing more "open" conversation about the studied phenomena or concepts. For example, when we tested the EU-SILC survey questions measuring subjective poverty, we added to the test questions concerning the way people talk about poverty. The study sought answers to the questions of what poverty is, how people categorise it and whether people felt themselves poor if statistics claimed they were

(Kallio 2004). The study was a good example of the fact that cognitive test data can also be used to study cultural discourses.

Last summer we collected data for an indicator project, and these are currently being analysed. The study was based on cognitive testing of two international indicators, that is, questions measuring monthly pay and the experiencing of health. These indicator questions had been embedded into the EU-SILC questionnaire in order to make their asking plausible in the context of a survey interview. We also compared the impact of two different explanations of the purpose for which the collected data would be used: 1) the data would be used for a national indicator and 2) they would be used for an EU indicator. Next I will describe this pretesting more closely.

Design of the testing of indicator questions

The underlying notion in the indicator project was that the behaviour of an individual person in each situation is dependent on that person's definition of it. Thus, answering questions also requires from the respondent inferences about the interviewer's intentions and about the reasons for asking the questions. Norenzayan and Schwarz (1999), for instance, have shown empirical results on how the institution the interviewer represents influences inferences about the meaning of a question. When respondents were asked why they thought mass murders took place, their answers stressed personality aspects or social factors depending on whether the questionnaire had been printed on the stationery of a psychological research institute or a social research institute. The answering was, therefore, influenced by the respondents' "knowledge" or mental picture of what the inquiring institution would do with the answers.

In our project, the comparative institutions were the European Union and Statistics Finland. The respondents were explained the purpose for which the data would be used in the contact letter, while making an appointment for the interview and at the beginning of the interview. Besides by cognitive testing, the interviewing protocol was steered by the following questions: How did the respondents perceive the data collection situation and the use of the data? What kinds of mental images did they have of Statistics Finland and the EU as data collectors? In addition, the perceived importance of the data collection, and trust in the survey topic and methods, and in the conductor of the interview were discussed in the interviews.

Our samples were drawn for the data collection from the areas of *Helsinki and Pohjanmaa (Finland's strongest anti-EU region)*. Efforts were made to get equal numbers of men and women interviewees aged 25 to 64 from all socio-economic groups. The

sample was split into two groups – one of which was told the data were collected for the EU and other that they were collected for national purposes. The conducted interviews numbered 43, for half of which the survey was justified by EU needs and for the other by national needs.

Comprehension of the interview situation was studied with the semi-structured theme interviewing method, understanding of the indicator questions was tested using the cognitive interviewing method and the standardised interview portion was conducted using the regular structured interviewing method. The interviews in Helsinki (23 persons) were done face-to-face, at the respondent's home or workplace or at Statistics Finland's premises as preferred by the respondents. The interviews in Pohjanmaa (20 persons) were conducted as telephone interviews.

The speech community in surveys of Statistics Finland?

The preliminary findings from comparing the face-to-face interviews was that the given justification, that is, EU use or national use, did not influence the answering. The respondents did not differentiate between the EU indicators and the national indicators. The participants of the telephone interviews paid next to no attention to the contact letter or to the justification for the purpose for which the data would be used. At the end of the telephone interview, some respondents did not even know who had commissioned the interview, and were not interested in finding out, either.

The face-to-face interviews produced interesting information about ordinary people's trust in questionnaire surveys and especially in surveys conducted by Statistics Finland. So, the factor influencing the interactive situation of an interview is the institution conducting the interview – an official, government agency, Statistics Finland – and the mental images respondents have of the way the data will be used.

The impact of the organiser of the survey is depicted well by one respondent's answer in which he stresses the social importance of Statistics Finland's data collecting and the influencing opportunity this gives to an individual person:

Interviewer (I): What made you decide to take part in this survey?

Answer (A): Well, I just generally consider it important (…) After all, Statistics Finland is an official interviewer… or maker of surveys, so that if there are things that are important in society, you can say what you think of them, I mean ordinary people have few opportunities for saying what they think, and by coincidence Statistics Finland also

recently conducted a survey at work and my workmates sort of thought it would not be worth answering but in my opinion responding to these should be taken seriously.

The answers contain several examples of this type. Participation in an interview conducted by Statistics Finland was almost regarded as a civic responsibility. It was viewed as a means of influencing decision-makers in society. Therefore they considered it **important** to respond to Statistics Finland's surveys:

I: What about when we rang you, what influenced your decision to participate then, was it the phone call or the letter you got?

A: I thought it my civic duty to do this thing, nothing more complicated, as I hadn't actually forbidden it.

In contrast, some respondents did even express reservations about other interview and survey organisations. Diverse market research surveys were especially regarded as dubious and unreliable.

However a couple of young persons preferred to participate in market research surveys instead of Statistics Finland's surveys. They thought that the topics of Statistics Finland's surveys were boring, but that they nevertheless, produces important data. However, compared to these other alternative organisations they regarded Statistics Finland as a most reliable and the safest. They had confidence in the way the data would be used.

The perceived **confidence** in the data collection situation seems to have been influenced by trust that the answers would be used anonymously, belief in the importance of the data collection (the obtained data would not used for making quick money) and the data collection method (face-to-face interview). Many respondents thought that they would probably talk about the same matters differently in a telephone interview because would not know whom they would be speaking to.

The subject of sensitivity was discussed in connection with the testing of the indicator questions. We presumed that the topics of pay and health in the indicator questions would be somewhat sensitive in the Finnish culture. Surprisingly enough, the interviewees defined Statistics Finland's face-to-face interviewing situation as confidential enough for them to talk about their income which they regard as a sensitive subject they would not discuss with their friends or acquaintances. The survey interviewing situation gave them the opportunity to talk about matters they would not discuss with friends or acquaintances. Some respondents regarded health as a more sensitive subject than pay, while others thought the opposite.

All in all, it would seem that at least as far as the face-to-face interviews are concerned it would not be appropriate to talk about a general survey speech community but about a speech community in interviews of the institution conducting the survey. It would be interesting to see this way of defining an interview situation within the framework of an international comparison.

Conclusions

Survey responses are used either as basic data for statistics or as research data for social studies. A statistical table without interpretation tempts treating a survey response as a fact representing both social reality and truth about differences between population groups. Survey responses collected with different interviewing methods become easily transformed into comparable facts in a statistical table. The conversion of mass interview responses into statistics is often a technical black box, on whose validity no information is generated, or is at least not reported or seen in interpretations. This may contain the presumption that the right measuring instruments used in the appropriate manner can produce truth that is free from errors. The aim in pretesting questionnaires is to support the above described way of using survey responses in order to develop questions that are as unambiguous as possible. It is not customary to use the information describing the question-answering process in the interpreting of results.

At the end of her article, Martin (2001) asks whether the standardising (stimulus-reaction) model on which surveys are based should be corrected or rejected altogether, because the two other theoretical perspectives have increased our understanding of survey answering as a cognitive process and as social interaction. However, changing of the current interviewing method into a conversational one in respect of questions other that so-called factual ones would seen an unlikely development direction. Instead, the trend would seem to be towards ever more efficient production of mass data. For example, the computer as an interviewing tool seems to support in many ways the stimulus-reaction based model of the question-answering process.

Cognitive pretesting of questions helps to design questions that are as unambiguous as possible. What if understanding the question-answering process as social interaction were to be used in the interpreting of survey answers? This would, of course, be more prudent for the research use than for the statistical use of survey responses.

References

Ahola, A./Lehtinen, M./Godenhjelm, P., 2002: Kysymisen taito. Surveylaboratorio lomaketutkimusten kehittämisessä. The ability of asking questions. Surveylaboratory in questionnaire development. (In Finnish) Statistics Finland. Helsinki.

Alasuutari, P., 1998: An invitation to social research. SAGE Publications. London.

Briggs, C. L., 1986: Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research. Cambridge University Press.

Kallio, M., 2004: Mitä köyhyys on? Köyhyyden kulttuurisista jäsennyksistä subjektiivisiin merkityksiin. What is poverty? From cultural analysis of poverty to its subjective meanings. (In Finnish). Statistics Finland. Helsinki. (being printed)

Norenzeyan, A./Schwarz, N., 1999: Telling what they want to know: participants tailor causal attributions to researchers' interests. European Journal of Social Psychology 29, 1011–1020.

Martin, E., 2001: Theoretical Perspectives on the Question and Answer Process: Implications for Pretesting. Proceedings of QUEST 2001, October 24-25, 2001. U.S. Census Bureau. Washington, D.C.

Ronkainen, S., 2002: Puhemaailma, kommunikaatio ja surveyhaastattelu. Speech community, communication and survey interview. (in Finnish). Evaluation presentation at the conference “The ability of asking questions”, 3.10.2002. Statistics Finland. Helsinki.

Contact

Anja Ahola
Statistics Finland
Social Statistics
FIN-00022 Statistics Finland
email: anja.ahola@stat.fi

Our Experiences from Measurement Tests in Developing Countries

Gunilla Davidsson & Birgit Henningsson

We would like to start a discussion with you about statistical assistance in developing countries. What kind of statistical assistance and/or knowledge are international organisations and national statistical offices from western countries offering? We would like to discuss this issue from our point of view: questionnaire design and cognitive measurement testing.

Too much data

The international organisations are almost always conducting their own large-scale surveys in the same way in all countries around the world. These very comprehensive surveys most often mean that the interviewer has to spend more than one day in the same household in order to collect all information wanted. This of course means that the quantity of data or information collected is much too much for the developing country to handle and take care of itself. In fact, the developing countries are only data providers to the international organisations. And at least up till now these organisations have analysed very little of the data or done very little research compared to all the data collected in countries all around the world. What kind of knowledge transference is that?

A colleague complained about all unnecessary data. One comment from another statistician was “We use only 40 % of the data”. Why do we collect so much data if we do not use it? There will be more and more questions in the questionnaire and the fieldwork will be impossible to do in a good way. The respondents do not know all the answers. Or do not remember.

In another country the administrators were really worried about the field workers. Would they really stay so long in the bush to have the data needed?

Translation

The questionnaires used are usually translated from English to the native language (-s) in question by a person with either good or rather good knowledge in the languages but not in the subject matter, or by a person with more knowledge in the subject matter than in the languages. Cultural differences in understanding the questions are very seldom taken care of. The questionnaires are literally translated. However, it doesn't matter which of the two ways the translation has been made, from our point of view the quality is mostly too poor. The questionnaires are never cognitively tested, but a large pilot is usually conducted. But these pilots are dress rehearsals of the data collection organisation and not a proper test of the questionnaire.

In one country with several languages the questionnaire was only in English. All the interviewers had to translate themselves while they were interviewing. At least they should have had a list with the most common words in the questionnaire translated. You do not know what you are measuring if every interviewer translate it in his/her own way. Teachers and participants wrote a paper about the problems. Some interesting examples showed the difficulties. Nothing happened. The problem is too big and costs a lot of money if you take it seriously.

Bad questions and low quality of interviewer training

In the countries where we have been involved with interviewer training, we have seen surveys with numbers of bad questions and low quality of interviewer training. None of the stakeholders or users seem to care about it at all. Actually they often do not have a clue about what has been collected, but still they go on financing and analysing (at least some results) from country to country. These international organisations are partly – I think mostly – financed by tax money from member countries. The question is if this really is the best way to use the money!?

"Victims of crime" is an international survey. I was asked by one country to support their interviewer training. When I got the questionnaire I told them that this would not work for the interviewer. You will end up telling them that they have done a bad job when it actually was a bad questionnaire. So to prove that I was right we had a blind mirror test. A very good interviewer soon got lost in the questionnaire and we all saw very clearly that the questionnaire did not work. But the result was that I had a tough job writing a new questionnaire with better structure, with a foldable flap attached to the form including all sorts of crimes. There were other improvements of the design as well. Preparing the training, which was my actual task, was there no time left for!

Build up surveys suitable for the country

What we would like to see are experienced statisticians with good knowledge not only in statistics but also in the culture of the country they are going to work with. The kind of assistance we think in deed could live up to be called assistance, is to help the country build up running surveys and/or registers that can provide the country with confident annual figures. Figures that are necessary and useful for the country. Today these large-scale surveys are ad hoc surveys focused on international comparability, which gives the developing countries very little knowledge and training in the ability to conduct surveys of their own in the future.

We met someone the other day that knew my colleague. He was there (in the developing country) once a year and took care of the tables in one specific survey. This year he seemed to be happy because he had more data than last year. I wondered what data? After several years still their own staff could not take care of the results.

Conclusion

So, what we are interested in is to discuss how to go on putting these problems on the international agenda. Can we create an interest in changing the policy of statistical assistance in developing countries? We would like to go from large-scale international studies to necessary national surveys with a sample size that is suitable for the specific purpose of the country. The national offices should be trained to handle the surveys themselves having in mind that they have a shortage of staff and practically non-existent own resources.

Contact

Gunilla Davidsson
SCB Statistiska centralbyrån
Statistics Sweden
Box 24 300
SE-104 51 Stockholm
SWEDEN
email: gunilla.davidsson@scb.se

Birgit Henningsson
SCB Statistiska centralbyrån
Statistics Sweden
Research and Development
SE-701 89 Örebro
SWEDEN
email: birgit.henningsson@scb.se

PRE-TESTING QUESTIONNAIRES: THE NEW ZEALAND EXPERIENCE

DENISE GREALISH

1. Introduction

This paper provides an overview of the process Statistics New Zealand (SNZ) uses to develop household, business and economic survey questionnaires. It describes the different pre-testing methods used to evaluate the questions and questionnaires, and how these methods work together and complement each other.

2. Questionnaire development process

Questionnaire development at SNZ starts with topic specifications supplied by clients. These clients can be internal to SNZ or external, from another government agency. Questionnaire developers need to be able to understand the objectives and data collection needs of a survey in order to develop questions that collect the exact information required. The topic specifications are referred to throughout the development of a questionnaire to ensure it is meeting the survey objectives. Topic specifications can be used to prioritise work, and for questionnaire review by clients and experts.

During the questionnaire development process, topic specifications are revisited when pre-testing shows up an issue with the concepts. For example, it may be necessary modify a specification, or delete one, if a robust question cannot be produced to the required standard.

When the questionnaire developers receive topic specifications questions are formed which are reviewed and tested. At SNZ, the layout of paper questionnaires is done within the questionnaire development section. This is a great advantage, as it allows for the quick implementation and retesting of recommendations that resulted from earlier pre-testing. For electronic questionnaires, flowcharts are the current method used at SNZ to specify requirements for programmers. The flowcharts detail question wording, structure, routing and edits.

A well-designed survey questionnaire should efficiently collect the data that best meets user needs, with a minimum number of errors, and the least possible burden on respondents.

Pre-testing is used at SNZ to ensure that questionnaires are easy to administer, easy to process, respondent and interviewer–friendly, meet the survey objectives, and meet SNZ's quality standards.

A variety of pre-testing methods are used, depending on the individual questionnaire being developed. The number of rounds of each type of testing, and the order of the pretesting methods used is driven by the individual questionnaire being developed, and by the time and resources available.

3. Pre-testing methods

3.1 Concept testing

During the early stages of questionnaire development, and on occasion in conjunction with the development of the topic specification by the clients, SNZ uses concept testing to help refine survey objectives. The purpose of concept testing is to ensure that a specific concept or topic required by the client is something that the survey population is able to understand and can report on. Concept testing can also be used to help define concepts that are still vague in the client's mind.

Concept testing usually takes the form of a focus group for household surveys, but in business survey development it is more likely to involve individual interviews with typical business enterprise respondents. Information gained during this early pre-testing is used to make recommendations about survey objectives so that there will be a better match between the information the survey aims to collect and the information that respondents can supply.

3.2 Expert reviews

A series of expert reviews of the questions and questionnaire takes place throughout the development. An expert draws on their previous experience, current theory and design research literature to provide a critique of the questions. The expert review provides a fresh set of eyes to critically look at the questions as the developer often gets too close to the subject matter to be able to see all the problems. As no field costs are involved, expert reviews are a relatively cheap pre-testing method. However the recommendations from the expert are based on theory rather than practice or observation.

A range of experts are used during SNZ questionnaire development. These may include:

- questionnaire developers who peer review the questions to ensure they meet SNZ's question quality and layout standards
- clients who ensure objectives are being met
- classification staff to ensure standard classifications are being used
- programmers to ensure best practise is being used
- staff from specialist areas within SNZ, for example to ensure the questionnaire meet coding and imputation requirements.

If a survey development relates to a subject matter area where SNZ has had no previous experience, an overseas expert who is familiar with that area may conduct the review.

3.3 Cognitive testing

Cognitive testing is used to discover problems experienced by the respondent and the interviewer with the 'process' of answering the questions. It helps identify problems with question comprehension, memory recall, selecting response options, interpretation of reference periods, reactions to sensitive questions, and question and response order effects.

A cognitive test is a one-on-one interview which allows the questionnaire designer to understand the process a respondent goes through to answer a question, in particular trying to detect any actions or understandings that were not intended by the designer. It can also detect usability issues with the questionnaire, such as incorrect routing or insufficient answer spaces.

Cognitive testing uses a combination of several techniques:

- observation of the respondent to see how they navigate through a questionnaire, where they have difficulties, get confused, or fail to notice instructions
- asking the respondent to 'think aloud' and verbalise their thoughts as they try to answer a question. This lets the questionnaire designer know how difficult the question was for the respondent, and if it was understood the way the designer intended
- concurrent probing (at the time the respondent answers the question) to get a better understanding of the response process
- paraphrasing (getting a respondent to repeat back a question in their own words) to check a respondent's understanding of a question
- retrospective probing, which can provide information about a group of questions or the questionnaire as a whole. This can help with information about the flow of a questionnaire, context effect or mode effects.

Recommendations from cognitive tests feed into the reviewing and rewriting of questions. The revised questions are then tested again. This process is iterative and can be repeated until the designer is satisfied with the question or for as long as time and resource constraints allow.

As a pre-testing method, cognitive testing is extremely valuable as it provides a rich source of information about the respondents' thought processes and how well questions work. It also allows a number of versions of a question to be tested in a short space of time. One of the drawbacks is that results from cognitive tests tend to focus on comprehension issues, so other types of problems may be missed, and the results can not always be generalised.

3.4 Usability testing

Usability testing is commonly used for electronic or web surveys but it can also be applied to paper surveys. Usability testing follows a similar methodology to cognitive testing but instead of looking for question comprehension issues, the test focuses on how the respondent or interviewer uses the questionnaire. That is, can they easily find their way around the questionnaire, do they read the instructions, how they enter responses, etc.

SNZ recently studied eye movement for a paper-based survey. Eye movement is tracked while a questionnaire is being completed. This allows the questionnaire designer to see which elements on the questionnaire respondents are attending to and which elements are being missed.

Business surveys also use a form of usability testing. As most of the concepts in business surveys are predefined internationally, the testing focuses on fitting the concepts into the New Zealand environment, that is, finding out how businesses store the required data and tailoring questions to fit.

3.5 Translation testing

The Longitudinal Immigration Survey – New Zealand was developed in English, the predominant language used in New Zealand, but was required to be translated into a number of Pacific Island and Asian languages. The purpose of the survey was to investigate the settlement experiences of migrants, and because those migrants with no English language skills would have a different settlement experience to those who did, it was important to interview them.

As resources are not available at SNZ to translate questionnaires, the translation for the field test was contracted out. The field test showed the translation was too formal and

non-conversational so not very usable in an interview setting. To improve the usability of the translation and the quality of the data captured, some usability testing was run with the translations.

A number of bilingual field interviewers were trained as cognitive testers, provided with the translated questionnaire, then required to run some tests. They made recommendations to the translators about the questionnaire to try and improve the usability of the translation. As the content of the survey had already been agreed on, the recommendations the field interviewers could make were limited to wording changes only. These had to be within the guidelines of ensuring the language was usable in a interview situation, the question translations were usable in context with each other, and the equivalent information to the English questions was collected.

This method of usability testing with the translation proved to be very successful in improving the usability of the translation, as well as being very time and cost effective.

3.6 Field testing

Field tests (pilot tests) are the most costly of all the pre-testing methods used to evaluate questionnaires during development. SNZ's questionnaire design standards state that a questionnaire must go through a field test before it can be signed off as being fully tested. This recognises the importance of evaluating the questionnaire in a field environment as well as in a test situation.

The main focus of a field test is to get the field interviewers' reaction to the questionnaire and how it works, as well as getting them to gauge the respondents' reaction to the questionnaire. Field interviewers fill out an evaluation form after their field test training and during the field work. It is important that interviewers write down the issues as soon as possible after they occur. Once the fieldwork is completed, interviewers are brought together as a group and debriefed by the questionnaire developers.

Questionnaire designers review the raw data from field tests and also look at the actual field test questionnaires. This is to find common mistakes that interviewers are making, non-standard answers or notes made by interviewers, and to see how answer and non-response categories are used.

As a way to increase the value of field tests for questionnaire development, SNZ also uses interview behaviour coding. This is when an interview is recorded and a questionnaire developer later codes the behaviour of the respondents and interviewers at each question, for example if the respondent asks for clarification before answering the question. While not always discovering new problems with questionnaires, SNZ has used behaviour

coding as a way to quantify problems that may already be known, which helps in the decision about which areas to focus rework resources on.

4. Summary

Pre-testing will always be a key part of questionnaire development. A variety of methods are used which work together and compliment each other. These methods are continuing to evolve and change as research into their validity and practice continues.

Contact

Denise Grealish
Statistics New Zealand
Questionnaire Design Consultancy
PO Box 2922
Wellington
New Zealand
email: denise.grealish@stats.govt.nz

MORE ON THE VALUE OF SPLIT BALLOTS

FLOYD JACKSON FOWLER, JR.

Introduction

Much attention has been given to strategies for testing how well questions are understood and answered. This kind of evaluation has great potential for improving survey measurement. Appropriate procedures for assessing how well questions are understood and the answers are becoming increasingly common, which constitutes significant progress in survey methodology.

However, the ultimate test of whether question problems matter is how they affect the data. Although our cognitive testing strategies seem to provide meaningful information about question problems, they do not tell us how much they adversely affect data and whether or not revised questions are in fact producing "better" data.

When there are two or more candidate questions to measure a particular construct, it would be best if the answers to the candidate questions could be correlated with some kind of gold standard, a measure that there was reason to think did constitute a valid measurement of the construct itself or one to which it should be related. In the absence of that, collecting distributions of the answers to candidate questions from comparable populations, and comparing the distributions, can often provide insight into whether or not the alternative wording of questions in fact affects the data (Fowler, 2004).

The contention of this paper is that it is difficult by inspection alone to know whether or not alternative wordings of similar questions will produce different estimates and, if so, which is a "better" estimate. I present six tables that were based on such "split ballot" experiments to illustrate the point.

Methods

The data presented came from two research projects that used very similar methods. In each case, a set of questions was initially subjected to cognitive testing. Some number of volunteers from the target populations were recruited, asked test questions, then asked in

various ways to explain their understanding of the questions and how they went about answering the questions. Based on those results, when problems were found, an alternative version of the questions designed to meet the same question objectives, but with better or different question design, was developed. The data in five of the tables (Tables 1-4, 6) were derived from a small national sample of adults, identified through random digit dialing, who were randomized to one or the other versions of each test question. About 75 adults answered each version of the questions. The other example, presented in Table 5, is based on a sample of health plan members who were interviewed by telephone. In this case, about 340 respondents answered each version of the question. The analysis simply compares the distributions to the two forms of the question, looking to see if the results are the same or different depending on the question wording.

Results

Table 1 presents results for questions designed to screen adults for whether or not they had ever consumed alcoholic beverages (12 drinks in any one year) before asking a series of questions about alcohol consumption.

The first alternative asks specifically whether the respondent has ever had at least 12 drinks in any one year. Alternative 2 asks if the respondent had ever had an average of more than one drink per month.

From a mathematical standpoint, the answers should be identical. However, as the table shows, many more people said that they had 12 drinks in a year than said they had an average of one drink per month. In this case, it is not clear how accurate the responses are to Alternative 1, but it is almost certain that the answers to Alternative 2 were confounded by the notion that some respondents thought they had to have at least one drink every month in order to say "yes." That is at least part of the explanation why 18% fewer said "yes" to Alternative 2 than did so to Alternative 1.

Table 2, a follow-up question, compares two different ways of estimating how often people drink alcohol. Alternative 1 asks for the number of days in the past 30 days that the respondents had any alcoholic beverage to drink. Alternative 2 asks them to summarize in the past year how many days per week, month, or year they drank any type of alcoholic beverage. As is the case for Table 1, from a mathematical perspective, we would expect the answers to be similar. Unless the "last month" was systematically unrepresentative, which we have no reason to think was the case, the numbers reported for the last month should be similar to what the reported pattern was over the past year.

In fact, as Table 2 shows, they are not similar at all. The average over the past year results in many more days of reported drinking than the question about the past month. Part of the reason may be that those people who do drink sometimes, but not often, did not report that they did not drink at all during the year but might report no days in the past month. However, if that was a main factor, the bottom two categories should add up to a similar sum, and they do not.

Table 3 is an examination of the effect of giving examples when abstract terms are used in questions. In this case, the question is about days respondents did any strenuous activities in and around their home. The original question provides a number of examples of strenuous activities, while the alternative question leaves out the examples. In all other respects, the questions are the same. As can be seen from the data, providing the examples greatly increases the number of days on which the respondents report any strenuous activity.

The data in Table 4 provide a comparison that might seem similar to Table 3. In this case, the question is about dental care. The original question provides respondents with examples of the various kinds of dentists they might have seen, such as orthodontists or oral surgeons. In contrast, the alternative assumes that respondents know what dental care means and provides no further examples. In contrast to the results in Table 3, in Table 4 there is no difference in the rates at which visits for dental care are reported. Providing the examples of kinds of dentists and dental care has essentially no effect on the answers.

Table 5 shows two series of questions designed to identify people who have chronic health conditions. In this case, for both series of questions, a chronic health condition was one that had lasted for at least three months and that either had required the taking of prescription medicines for three months or had led to seeing a doctor about the condition three or more times in the past year. The series differ in one crucial respect. The "standard" form begins with a question about whether or not the respondent has had a condition for three months, then asks if any of those conditions met the standard for either use of prescription drugs or seeking medical care. The alternative begins with whether or not respondents have had a condition for which they have taken medication for three months or seen a doctor three or more times in the past year, then asks whether the condition has lasted for at least three months. As the table shows, the series that begins with the behavioral implications of the condition, taking drugs or seeing a doctor, produce reports of many more chronic conditions than the series that started with the general question about having a condition for three or more months. In this case, we have cognitive testing results that show that people are not consistent in their understanding of what a "physical or medical condition" is. It seems highly likely that there is significant

under-reporting of conditions based on the ambiguity of the question. Hence, while we are not certain about the level of validity of the alternative, we are pretty sure, based on our testing, that the answers to the alternative are better than those to the "standard" series.

Table 6 shows two series of questions designed to identify people who had been injured in an automobile accident "because of their driving." The alternative series breaks the initial question into three parts: 1) injured in an automobile accident, 2) while you were driving, 3) because of your driving. The data show that the estimates from the two approaches are significantly different. Many fewer people end up saying "yes" when they are asked the three-question series than when they are asked the question in its initial form. Again, because of initial testing of this question, we are pretty sure that respondents did not attend to all of the issues raised when they are all presented in a single question. Based on that analysis, we are confident that the second series of questions is producing more valid data.

Discussion

The take-away point from the above is that split ballot comparisons of alternative forms of questions provide invaluable information about how question wording affects resulting data. For some of the above examples, such as Tables 5 & 6, we had a strong theory based on cognitive testing about why one of the versions might contain significant error. However, we needed the split-ballot data to prove it.

Tables 3 & 4 present similar comparisons but with different results. In Table 3, providing examples of strenuous activities greatly increased the reporting of such activities; in contrast, in Table 4, giving examples of dental care had no effect on the amount of dental care that was report. I would argue that it would be very difficult for even a skilled question design expert to have reliably predicted in advance how those two comparisons would turn out.

Tables 1 & 2, comparing alternative ways of asking about alcohol consumption, provide great examples of hard-to-predict results. Logically and mathematically, the data should be identical from those two pairs of questions. In fact, the results are very different.

Thus, cognitive testing of questions provides an extremely useful way to identify question problems, to diagnose problems with questions, and stimulate the revision of survey questions. However, in the end, we need to have data about how the resulting survey estimates will be affected by problems in our original or revised questions. The point of this paper is to encourage researchers to build in split-ballot tests of their proposed

question wording prior to fielding their full-scale surveys. Cognitive testing results in combination with split-ballot testing provides a better basis for making decisions about which questions to ask.

Acknowledgements

The collection of the data represented in this paper was funded by contract from the National Center for Health Statistics (NCHS) and by a cooperative agreement between the Agency for Healthcare Research and Quality and the Harvard Medical School (Grant # HS-09205). The author wants acknowledge to contribution of Paul Beatty at NCHS to the design of some of these experiments. In addition, it should be noted that some of these results were reported at the Questionnaire Development Evaluation and Testing (QDET) conference and will be published in the monograph derived from the conference (Fowler, 2004).

References

Fowler, F.J. (2004) "Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies" In Presser, S. et. al. (eds) *Questionnaire Development Evaluation and Testing Methods*, New York: Wiley.

Table 1:

Alternative 1

The next questions are about drinking alcoholic beverages. Included are liquor, such as whiskey or gin, beer, wine, wine coolers, and any other type of alcoholic beverage. In any one year, have you ever had at least 12 drinks of any type of alcoholic beverage?

Alternative 2

The next questions are about drinking alcoholic beverages. Included are liquor, such as whiskey or gin, beer, wine, wine coolers, and any other type of alcoholic beverage. In any one year, have you ever had an average of more than one drink per month?

	Alternative 1	Alternative 2
Yes	71%	53%
No	29%	47%
Total	100% (n=77)	100% (n=79)

$p < .02$

Table 2:

Alternative 1

In the last 30 days, on how many days did you drink any type of alcoholic beverage?

Alternative 2

In the past year, on how many days per week, month, or year did you drink any type of alcoholic beverage?

Number of Days	Alternative 1	Alternative 2
0	46%	25%
1-5	41%	45%
6-10	7%	17%
11+	6%	13%
Total	100% (n=79)	100% (n=75)
Mean Days	2.6	5.3

$p < .01$

Table 3:

Original

During the past 30 days, on how many days did you do strenuous tasks in or around your home? By strenuous tasks, we mean things such as shoveling soil in a garden, chopping wood, major carpentry projects, cleaning the garage, scrubbing floors, or moving furniture.

Alternative

During the past 30 days, on how many days did you do any strenuous tasks in or around your home?

Number of Days	Original	Alternative
0	32%	42%
1-5	34%	37%
6-10	13%	10%
11+	21%	11%
Total	100% (n=77)	100% (n=79)
Mean Days	4.66	2.72

$p < .05$

Table 4:

Original

About how many months has it been since you last saw or talked to a dentist? Include all types of dentists, such as orthodontists, oral surgeons, or all other dental specialists, as well as dental hygienists.

Alternative

About how many months has it been since you last went to a dentist office for any type of dental care?

	Original	Alternative
6 months or less	60%	57%
More than 6 months but not more than 1 year	14%	18%
More than 1 year	26%	25%
Total	100% (n=77)	100% (n=79)

NS

Table 5:

Standard

a. Do you now have any physical or medical conditions that have lasted *for at least 3 months*? (Women: DO NOT include pregnancy.)

b. In the last 12 months, have you *seen a doctor* or other health provider *more than twice* for any of these conditions?

c. Have you been taking prescription medicine for at least 3 months for any of these conditions?

Alternative

a. In the past 12 months, have you seen a doctor or other health provider 3 or more times for the same condition or problem?

b. Is this a condition that has lasted for at least 3 months? (Do *not* include pregnancy.)

c. Do you now need to take medicine prescribed by a doctor (other than birth control)?

d. Is this to treat a condition that has *lasted for at least 3 months*? (Do not include pregnancy or menopause.)

	Has Chronic Condition	
Yes	38%	56%
No	62%	44%
Total	100% (335)	100% (347)

$p < .01$

Table 6:

Original

This question is about automobile injuries, including injuries from crashes, burns, and any other kind of accidents. Have you ever had an injury because of your driving?

Alternative

a. This question is about automobile injuries, including injuries from crashes, burns, and any other kind of accidents. Have you ever had an injury while you were in a car?

b. Were you ever the driver when you were injured?

c. Were you ever injured because of your driving?

	Original	**Alternative**
Yes	8%	2%
No	92%	98%
Total	100% (n=79)	100% (n=77)

$p < .05$

Contact

Floyd Jackson Fowler, Jr.
Center for Survey Research
University of Massachusetts, Boston
100 Morrissey Blvd.
Boston, MA 02125
U.S.A.
email: floyd.fowler@umb.edu

PARAPHRASING CAN BE DANGEROUS: A LITTLE EXPERIMENT

PETER PRÜFER & MARGRIT REXROTH

Paraphrasing is a well known cognitive technique: Respondents are asked to repeat a question in their own words. Paraphrasing permits the researcher to examine whether the respondent understands the question and interprets it in the manner intended.[1]

We conducted a little experiment to find out to what extend the paraphrasing technique would be really able to produce valid information about that point.

Some time ago, we conducted a cognitive pretest for the **ALLBUS**, the German General Social Survey. The pretest sample was a quota sample with 20 persons. Among the questions we had to test, was the following one:

Here is a card with different political activities. Please tell me to what extent you personally could have an influence to reach a political goal. Please tell me for each activity, whether you think your personal influence would be very effective, rather effective, not very effective or not effective at all.
Int.: Show card 1

CARD 1

- very effective
- rather effective
- not very effective
- not effective at all

A - Express your opinion to friends and acquaintences and at work
B - Vote at elections
D - Participate in a citizens' action group
E - Voluntary work for a political party
H - Occupy houses, factories or government offices
M - Take part in an authorised demonstration
N - Not vote at elections out of protest

1 This Definition is taken from Esposito, J., 2003: A Lexicon of Questionnaire Evaluation Terminology: Concepts and Working Definitions (unpublished).

One of the most important aspects of this question is, that respondent have to consider their own political activities and not political activities in general. Therefore our main goal was to get information about this aspect:

Did the respondents consider their own political activities in the way intended?

We wanted to know, whether the paraphrasing technique would be able to give us information about this important aspect. Our expectations were not very great because we knew that paraphrasing is a technique which works more generally and is more effective when you want to find out how the whole question is interpreted. On the other hand, the aspect of "their own participation" was so important, that it seemed to be plausible, that the paraphrasing answers could give us information whether the respondents had considered this specific aspect or not.

In a first step, we used the paraphrasing technique: After the respondent had answered the whole question (all items), we asked:
"Could you please repeat the question in your own words."

In a second step, we asked a special probing question directly after the paraphrasing: *"When you answered the question, did you consider your own participation?"*

We put the answers of the respondents into two categories:

Category 1: Answers which didn't mention the aspect of "their own participation"
Category 2: Answers which mentioned the aspect of "their own participation"

Of course we sometimes had the problem of interpretation. We decided to put the answers into the categories according to their meaning.

Table 1: Distribution of the answers in the two categories

Paraphrasing	
Category 1: Answers which didn't mention the aspect of the "own participation"	**13 Cases**
Category 2: Answers which mentioned the aspect of the "ow participation	**7 Cases**

Category 1

Let‘s first have a look at Category 1: In all 13 cases, the only we can say is, that paraphrasing failed to give us useful information about our point of interest.

The following table shows three examples (out of these 13 cases) of paraphrasing answers which explicitly didn‘t mention the aspect of "their own participation”:

Table 2: Three examples of paraphrasing answers which didn‘t mention the aspect of the “own participation”

ID	Paraphrasing answers
1	*“How effective are the possibilities of political participation.”*
2	*“Whether the items listed contribute to realize a political aim.“*
19	*“Altogether how you participate in political life.”*

And now, let’s have a look at how these three respondents answered the following probing question:

“When you answered the question, did you consider your own participation?”

Table 3: Comparison between paraphrasing answers and probing answers

ID	Paraphrasing answers	Probing answers
1	*“How effective are the possibilities of political participation.”*	"Yes"
2	*“Whether the items listed contribute to realize a political aim.“*	"Yes"
19	*“Altogether how you participate in political life.”*	"Always"

In category 1 we totally counted 7 cases of this type: Respondents who didn’t mention their own participation but had considered it when answering the question.

Category 2

Let‘s now have a look at Category 2: As we noted at the beginning, 7 cases fell into this category.The following table shows three examples (out of 7 cases) of answers which mentioned the aspect of "their own participation" :

Table 4: Three examples of paraphrasing answers which mentioned the aspect of "their own participation"

ID	Paraphrasing answers
7	*“What impact my opinion has on my environment. Can I move anything at all.”*
8	*“My influence, which I can bring to bear on political decisions, e.g. with regard to acquaintances and friends, dependent on how convincing I present my opinion.”*
12	*“How effective my own personal efforts are or would be, in order to realize certain political ideas and exert an impact.”*

And now, let’s have a look at how these three respondents answered the probing question:

“When you answered the question, did you consider your own participation?”

Table 5: Comparison between paraphrasing answers and probing answers

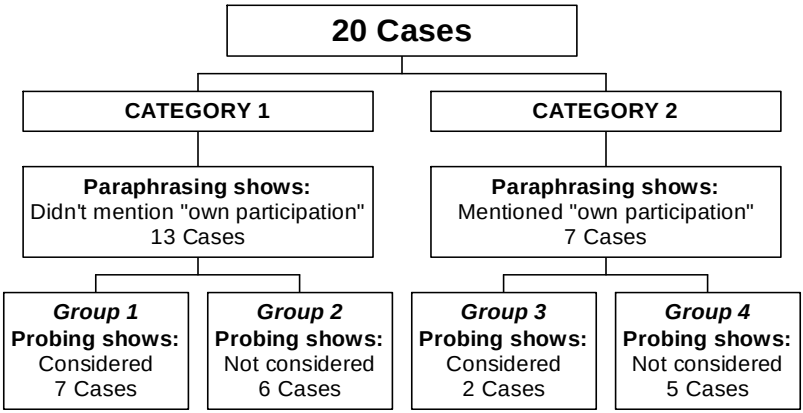
ID	Paraphrasing answers	Probing answers
7	*"What impact my opinion has on my environment. Can I move anything at all."*	*“No, I thought more general.”*
8	*"My influence, which I can bring to bear on political decisions, e.g. with regard to acquaintances and friends, dependent on how convincing I present my opinion."*	*“No, I make myself rather comfortable, when it deals with political things”*
12	*“How effective my own personal efforts are or would be, in order to realize certain political ideas and exert an impact.”*	*“I thought rather general, I didn't rather think of me. At A and B I thought general, at D personally and the rest in general.”*

In category 2 we totally counted 5 cases of this type: Respondents who mentioned their own participation but had not considered it when answering the question.

Summary Results

The following flowchart gives an overview about how the 20 cases are distributed over the two paraphrasing categories and the four probing alternatives.

Flowchart 1: Distribution of paraphrasing answers and probing answers



Let's first have a look at the two paraphrasing categories: They show that in 13 out of 20 cases the respondents had not mentioned "their own participation". The remaining 7 cases showed that the respondents had mentioned "their own participation".

Furthermore, the flowchart shows that in group 2 and group 3 the results from paraphrasing and probing are corresponding (totally 8 cases).

The results in group 1 and in group 4 however indicate clearly, that the results from paraphrasing and probing are not corresponding (totally 12 cases).

The results in group 1 show that in 7 cases paraphrasing had not provided the information that the respondents actually had considered their own participation.

The results in group 4 show that in 5 cases respondents had mentioned their own participation in their paraphrasing answers but had not or not always considered their own participation when answering the question. These results in group 4 are insofar amazing, as one should actually assume, that respondents who mention a certain aspect when they repeat a question in their own words would also have considered this aspect when answering the question.

Conclusions

The results of our experiment show:

- Even if respondents didn't mention an important aspect of a question when they tried to repeat it in their own words, they might have considered it in the manner intended when answering the question.
  That means: Paraphrasing answers may be misinterpreted because sometimes they don't show that respondents had actually considered important aspects or had actually understood the question in the way intended.
- Even if respondents have mentioned an important aspect of the question when they tried to repeat it in their own words, they might not have considered this important aspect when answering the question.
  That means: Paraphrasing answers may be misinterpreted because sometimes they include information which appears to show that respondents had considered important aspects or had understood the question in the way intended even though they actually had not considered these aspects or had not understood the question in the way intended.

Therefore our recommandations are:

- Deal carefully with paraphrasing answers.
- Don't draw any conclusion from paraphrasing answers to the actual response behaviour unless verifying it with other techniques.
- Use paraphrasing only as a starting point for additional techniques, such as probing.

Otherwise paraphrasing can be dangerous.

References

Esposito, J., 2003: A Lexicon of Questionnaire Evaluation Terminology: Concepts and Working Definitions. QUEST working paper (unpublished).

Sudman, S., Bradburn, N., Schwarz, N., 1996: Thinking about answers. The application of cognitive processes to survey methodology. Jossey-Bass, San Francisco.

Tourangeau, R., Rips, L.J., Rasinski, K., 2000: The Psychology of Survey Response. Cambridge Univerity Press.

Contact

Peter Prüfer
ZUMA
Postfach 12 21 55
D – 68072 Mannheim
email: pruefer@zuma-mannheim.de

Margrit Rexroth
ZUMA
Postfach 12 21 55
D – 68072 Mannheim
email: rexroth@zuma-mannheim.de

Appendix

Answers to paraphrasing and answers to probing

ID	Cat.	Answers to paraphrasing	Answers to probing
1	1	“How effective are the possibilities of political participation.”	"Yes"
2	1	“Whether the items listed contribute to realize a political aim.”	"Yes"
3	2	“Whether my personal activities are effective or not.”	"Yes"
4	1	“Which activities I regard as very efficient to not efficient at all with regard to politics.”	"No, rather on collectivity"
5	1	“How my stance is towards the different items, how effective I think the different items are.”	"No"
6	2	“Whether I can move something.”	"Yes"
7	2	“What impact my opinion has on my environment. Can I move anything at all.”	“No, I thought more general”
8	2	“My influence, which I can bring to bear on political decisions, e.g. with regard to acquaintances and friends, dependent on how convincing I present my opinion.”	“No, I make myself rather comfortable, when it deals with political things”
9	1	“I have not listened carefully, do not remember.”	"Yes"
10	1	“What I personally think is effective or not effective in order to reach some goal.”	"Yes"
11	1	“It is on the opinion, I have about parties and the ways to participate in different areas.”	"Yes, but personally it is not very effective."
12	2	“How effective my own personal efforts are or would be, in order to realize certain political ideas and exert an impact.”	“I thought rather general, I didn't rather think of me. At A and B I thought general, at D personally and the rest in general.”
13	1	“I am supposed to classify activities according to efficacy.”	“No, I thought of activities which anybody can carry out.”
14	1	“Do not remember”	“Not always, my answers were quite general, at A I thought of my own activities, at B in general.”
15	1	“Whether the items mentioned are effective.”	“Sure yes, but not absolutely.”
16	2	“What I personally can reach in different areas, when I utter my opinion. What impact I personally think to be able to exert.”	“Not absolutely of my own participation, but what I think about it. At H and N I didn't think personally, but about my feelings, when I read that.”
17	1	“Now I am getting confused a little bit, I do not know anything any more. Given that this is all new to me.”	“No, I can't do anything by myself.”
18	1	“I only have A, B, C in my mind, yes, friends and acquaintances, whether that was effective.”	“Yes to all”
19	1	“Altogether how you participate in political life.”	“Always”
20	2	“Whether I personally, if I personally participate, whether I can move something.”	“Only sometimes, at B and M"

EXAMINING EXPERT REVIEWS AS A PRETEST METHOD[1]

TERRY DEMAIO &ASHLEY LANDRETH

Introduction

Expert reviews are frequently used as a method of evaluating draft questionnaires. Either alone or in combination with other methods, people who have theoretical questionnaire knowledge or practical experience are asked to review draft questionnaires with an eye to identifying questionnaire problems. This can be done either by having individuals review the questionnaire alone or convening a group, also known as an "expert panel."

Presser/Blair (1994) included expert panels in their research on the effectiveness and reliability of different methods of pretesting questionnaires. But to our knowledge no work has been done to evaluate the consistency of the results produced by individual expert reviewers. We believe it is important to look at the results of individual expert reviews, because we suspect that time and resource constraints cause individual reviewers to conduct the bulk of expert reviews in the early stages of pretesting.

As part of an experiment on alternative cognitive interviewing methods, we used the results from individual expert reviews to gauge the breadth of results produced by three different teams of cognitive interviewers. Rather than having the perspective of one expert represent the potential for problems contained in the questionnaire, we chose to spread the responsibility by recruiting three experts, who worked separately to review the same questionnaire[2] using the same set of instructions.[3]

---

1 This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and are not necessarily those of the U.S. Census Bureau.

2 The questionnaire was a pre-existing CATI general population survey on recycling containing 48 items. Its objectives included determining trash removal practices among households, determining the level of participation in recycling among households, eliciting attitudes about

We expected to find a reasonable level of agreement among the experts in their evaluations of the questionnaire. But we were somewhat surprised at the outcome. In this paper, we lay out our findings and discuss what they might mean for questionnaire development and pretesting.

Methods

Three survey methodologists, from three different Federal government agencies, were enlisted to conduct the expert reviews. Each reviewer had 10 or more years of experience in questionnaire design, cognitive interview methods, and/or survey interview process research and were selected for their ability to complete the task in the time required. Each expert was asked to enumerate problems question by question in a 48-question survey on recycling. They were also asked to identify the five worst questions in the questionnaire, the five worst (i.e. most major) problems with the questionnaire, and the question numbers that reflected those problems. The experts reported their review on paper forms that were provided to them (see Attachment A for a sample of the report forms). The forms were then coded by applying a questionnaire appraisal coding scheme containing 28 problem types (see Attachment B).[4] Problem types and their locations were recorded in a database and compared across the experts.

Results

Table 1 presents the degree to which the experts agreed among themselves in identifying the number and types of problems in the questionnaire. As the top row shows, there are vast differences among the experts regarding the total number of problems each identified, with Evaluator B having identified less than one quarter (17 percent), and Evaluator C having identified less than half (41 percent), of the 158 problems identified by Evaluator A.

---

recycling, and eliciting opinions on alternative recycling strategies designed to increase the level of this behavior.

3 For information about the larger research project, see DeMaio/Landreth (in press).

4 The coding scheme is a close adaptation of that used in recent experimental research on alternative pretesting methods (Rothgeb/Willis/Forsyth, 2001), and was first created by Lessler/Forsyth (1996). The questionnaire problems documented by the expert review forms were coded by both authors, with a good level of inter-coder agreement (76.3 percent).

We examined the level of agreement among experts in identifying specific problems in particular questions, and it was extremely low (21 percent).[5] However, the possibility exists that the experts found similar types of problems (e.g. vague terms such as "recycling" and "household trash") but elected to document them at different points in the questionnaire. The data in Table 1 are consistent with this hypothesis. The percentages of problem types identified by experts at the highest level of aggregation (i.e. the categories labeled *interviewer difficulties, comprehension, retrieval, judgment,* and *response)* rank similarly across experts. The *comprehension* category ranks highest in terms of the percentage of these problem types found by each expert, between 40.8 and 60.0 percent. The *response* category ranks second highest, between 29.2 and 33.6 percent, and the *interviewer difficulties* category ranks third for at least two experts, between 13.9 and 22.2 percent. These three categories contain the majority of the problems identified by the interviewers. Agreement among experts also seems consistent for the lowest ranked categories, *retrieval* and *judgment,* which captured the fewest problem types, ranging from 0 to 7.4 percent.

Table 1 Percent of Problem Types Identified by Experts

Problem Types	Experts A (N = 158)	B (N = 27)	C (N = 65)
Interviewer Difficulties	13.9	22.2	4.6
Comprehension	50.6	40.7	60.0
Retrieval	1.3	0.0	4.6
Judgment	0.6	7.4	1.5
Response	33.6	29.6	29.2
Total %	100	99.9	99.9

However, the coder agreement among problem types does not necessarily support the notion that experts reported the same problems at different points in the questionnaire because it could also be the case that the problems they identified were of the same general type (e.g. comprehension) but focused on different terms in the questions (e.g. "recyclables," "trash," etc.). Our coding was not detailed enough to capture these differences.

5 Agreement statistic was generated by dividing the total occurrences of cases where two or more experts agreed on problem type and location (i.e. question number) by the total number of mutually exclusive problem types across all experts (N = 204).

We looked at these results in another way, focusing on the number of *questions* the experts identified as having at least one problem, rather than the number of problems themselves. The second row of Table 2 shows that there is quite a bit more similarity here, at least between two of the experts. Of course, there are external limits imposed here by the number of questions they had to evaluate. But the lowest number of problem questions identified is 41.3 percent of the highest number (19/46); in contrast, the lowest number of problems identified is only 16.4 percent of the highest number (27/165). In other words, there seemed to be a great deal more disparity across experts when comparing the number of problems each found, while the differences seem far less dramatic when comparing the number of flawed questions they identified.

Table 2 Number of Questionnaire Problems and Flawed Question by Expert

	Experts		
Problems & Flawed Questions	**A**	**B**	**C**
Number of problems found	158	27	65
Number of questions w/problems	46	19	40
Number of questions affected by major problems	38	14	10

Experts were asked to identify the five worst questions in the questionnaire. The variation in identifying problem questions was just as great as identifying individual problems. Only one question was named as the worst question by all three experts. One other question was mentioned by two experts, and all the other "worst questions" were named by only one expert. That is, there were 10 "worst questions" that none of the experts agreed on.

Experts were also asked to identify the five "most important problems"[6] they found with the questionnaire, starting with the most broad ones (i.e. those broad problems that could potentially affect more than one question or aspect of the questionnaire). For each problem, they were instructed to list the question numbers of the items that were most likely to be affected by it. Again, we found great variability in the evaluations of the experts. The magnitude of the problems varied from large things like "the survey wants household level information but the questions ask for person level estimates" to fairly

6 Experts were not provided any criteria for identifying the most important problems. They relied on their own interpretations of this concept.

small things like “the ordering of the scale items.” In terms of agreement across the experts, there was no major problem that was mentioned by all three experts. Three of the major problems had agreement by two of the experts, and six problems were only mentioned by one of the experts. In one case, an expert listed two major problems that were related and both fit under one problem listed by another expert. So, experts agreed less often on these general overarching problems than we would have expected.

Even when the same problems were identified, however, there were large differences in the number of questions that were reported as being affected by the problem. Overall, as the bottom line of Table 2 shows, many more questions were identified by Expert A than Experts B or C as being affected by the major problems they reported. Specifically Expert A reported an average of 7 questions affected by each major problem, while Experts B and C each reported an average of 2.2 problems.

Discussion

An analysis of the output of the experts suggests different review styles are operating. Expert A clearly has a very detailed focus, finding more than twice as many problems as the nearest other expert. This included problems with almost all of the questions (46 out of 48), with each question having an average of 3.4 problems. In addition, the five worst problems enumerated by this expert were reported to affect almost 80 percent of the questions (38 out of 48). A very careful and critical review was necessary to elicit the level of evaluative information contained in this report.

Expert B, in contrast, can be thought of as having a minimalist focus. Relatively few problems were identified by this expert compared to either of the other two. Less than half the questions (19) were seen as having problems, with each question having, on average 1.4 problems. The five worst questionnaire problems identified by this evaluator were fairly narrow in the number of questions they applied to (14). On the basis of this information, one would have a hard time believing that experts A and B were reviewing the same questionnaire.

Expert C has a more middle-of-the-road focus. The number of questionnaire problems identified was in the middle of the other two experts. The number of questions identified as problematic was similar to Expert A, while the number of questions affected by the five worst problems was similar to Expert B. While more questions were found to be problematic, the number of problems identified per question is fairly low (1.6 on average). In fact, there is a lower ratio of problem questions that are affected by a broad

problem (10/40) for Expert C than for either of the others – 38/46 for Expert A and 14/19 for Expert B.

There are several explanations for the disparity of these results. One is the amount of time the experts were able to devote to the task. A priori, one could argue that the more time is allotted, the more comprehensive the review. (However, this information was not collected for this project.) Second is differential expectations about the level of detail required in the assignment. Once an expert identified a problem in question 3, for example, he/she may not have felt it necessary to report it again in the item-by-item portion of the task for every other question that suffered from the same problem. Third is a difference in the perceptions of the experts as to what constitutes a good or bad question. Fourth, although the experts all had 10 or more years of experience in questionnaire design, cognitive interview methods, and/or survey interview process research, their particular experience and expertise may have left them better or worse at evaluating questionnaires. Finally, some experts may be used to working in review panels rather than individually, and feel hampered by the non-collaborative style here.

The low level of agreement among the experts in our research is enough to cause concern about the generalizability of expert review results. While expert reviews are typically considered as a quick, low-cost method of obtaining input about questionnaire problems, some thought should be given to specific aspects of the review procedures. Who does the expert reviews and how they are done may have important implications for the quality of the review. We thought we were providing specific guidelines to our reviewers. But although the reporting format was standardized, the process of problem discovery was not. Some experts may have used a question appraisal scheme to guide their review, while others may have taken a less structured approach. Without more controlled research on this topic, we would suggest that the results of a single expert may not be sufficient, either by itself as a pretest method or as a preliminary step for cognitive interviews. It seems to us that expert review panels (even small ones) would, by their collaborative nature, yield more consistent results. And in addition, some structured procedures such as using a question appraisal scheme to guide their review should be presented to experts. Perhaps further research in this area could determine what the best method of approaching the expert review task would be.

References

DeMaio, T.J./Landreth, A.D., Cognitive Interviews: Do Different Methodologies Produce Different Results?, in S. Presser/J. Rothgeb/M. Couper/J. Lessler/E. Martin/J. Martin/E. Singer (eds.), Questionnaire Development Evaluation and Testing Methods, New York: Wiley Interscience (forthcoming).

Lessler, J./Forsyth, B., 1996: A Coding System for Appraising Questionnaires, in N. Schwarz/S. Sudman (eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Presser, S./Blair. J., 1994: Survey Pretesting: Do Different Methods Produce Different Results? In P.V. Marsden (ed.), Sociological Methodology: Volume 24, Beverly Hills, CA: Sage, pp. 73-104.

Rothgeb, J./Willis, G./Forsyth, B., 2001: Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? Paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, May 2001.

Contact

Terry DeMaio
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: Theresa.J.DeMaio@census.gov

Ashley Landreth
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: ashley.denele.landreth@census.gov

Attachment A:
Selected Pages from Expert Review Report Forms

INSTRUCTIONS

Independent Evaluator Record Sheets

Feel free to use the attached forms to record your analyses. If you prefer to submit typed feedback, please adhere to the general format outlined in the following pages.

PART I: Question-by-Question Problem Identification

1. Question wordings, instructions, and response categories are considered in-scope for this evaluation. For each survey question, identify and briefly explain each specific problem you find. Each problem should be recorded separately and given a number. The attached form only allows for seven (7) problems, but if you find this is insufficient, feel free to continue the numbering scheme to add to the problem list. If additional problems are identified, remember to record all the relevant data associated with these problems – as outlined in items 2 and 3 below.
2. For each problem you identify, mark one box – labeled "H" or "L" – to classify it as a high or low priority problem according to the following definitions:

   High priority: A problem that <u>should</u> be addressed before the instrument is fielded, because it <u>will likely</u> adversely affect the response process in unacceptable ways.

   Low priority: A problem that <u>could</u> be addressed before instrument is fielded, but <u>may not</u> adversely affect the response process in unacceptable ways.
3. For each specific problem identified, mark one box – labeled "A" or "R" or "B" – to record whether it will be a problem with administering the question (A = administration), a response problem (R = response), or both (B = both).

PART II: Five (5) Most Important Problems

1. Briefly state the five (5) most important problems you found with this questionnaire. Please list any broad/general problems first (i.e. those that apply to more than one question or aspect of the questionnaire).
2. For each of the problems you identify, please list the question numbers that are likely to be affected.

PART III: Five (5) Worst Questions

1. Identify the five (5) worst questions. For each, please include a short (i.e. 1-2 sentences) explanation for its selection.

RECORD SHEET EXAMPLES

PART I: Question-by-Question Problem Identification

Q10

	Priority:	Problem for:
Problem 1:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**

This question is double-barreled; it asks respondents to enumerate the number of years since they bought the horse AND moved to Montana.

PART II: Five (5) Most Important Problems

Briefly state the problem and list affected question numbers:

Problem 1:

Awkwardly worded questions will be difficult for respondents to comprehend the first time the question is read.

Affected question numbers: Q9, Q42, Q75, and Q99

PART III: Five (5) Worst Questions

Identify question number and provide brief explanation for why it was selected (1-2 sentences):

Problem 1. Q #: Q76 Explanation:

This question's response set is not mutually exclusive, and it will be impossible for respondents to select only one option – as the question's instruction suggests.

PART I: Question-by-Question Problem Identification

Q1

	Priority:	Problem for:
Problem 1:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 2:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 3:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 4:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 5:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 6:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**
Problem 7:	☐ **H** ☐ **L**	☐ **A** ☐ **R** ☐ **B**

PART II: Five (5) Most Important Problems

Briefly state the problem and list affected question numbers:

Problem 1:

Affected question numbers: _____

Problem 2:

Affected question numbers: _____

Problem 3:

Affected question numbers: _____

Problem 4:

Affected question numbers: _____

Problem 5:

Affected question numbers: _____

PART III: Five (5) Worst Questions

Identify question number and provide brief explanation for why it was selected (1-2 sentences):

Problem 1. Q #: _____ Explanation:

Problem 2. Q #: _____ Explanation:

Problem 3. Q #: _____ Explanation:

Problem 4. Q #: _____ Explanation:

Problem 5. Q #: _____ Explanation:

Attachment B:
Questionnaire Appraisal Coding Scheme

Interviewer Difficulties	**Comprehension**	**Retrieval**	**Judgment**	**Response Selection**
IR Difficulties	Question Content	Retrieval from Memory	Judgment & Evaluation	Response Terminology
1 Inaccurate instruction 2 Complicated instruction 3 Difficult for interviewer to administer	4 Vague/unclear Q 5 Complex topic 6 Topic carried over from earlier Q 7 Undefined/vague term	17 Shortage of memory cues 18 High detail required or info unavailable 19 Long recall or reference period	20 Complex estimation, difficult mental calculation required 21 Potentially sensitive or desirability bias	22 Undefined term 23 Vague term
	Question Structure			Response Units
	8 Transition needed 9 Unclear respondent instruction 10 Question too long 11 Complex/awkward syntax 12 Erroneous assumption 13 Several questions			24 Responses use wrong or mismatching units 27 Unclear to respondent what response options are 28 Multi-dimensional response set
	Reference Period			Response Structure
	14 Period carried over from earlier Q 15 Undefined period 16 Unanchored/rolling period			25 Overlapping categories 26 Missing response categories

EVALUATION PLAN FOR THE DUTCH STRUCTURAL BUSINESS STATISTICS QUESTIONNAIRES: USING OUTPUT TO GUIDE INPUT IMPROVEMENTS

DEIRDRE GIESEN[1]

1. Introduction

In establishment surveys issues of usability and respondent-friendliness are often neglected. Dillman (2000, p 345) tellingly describes the implicit model for government business surveys as "a Cost Compensation Model". In this model the goal to minimize monetary cost determines the questionnaire design and implementation practices. To compensate for the resulting flaws in the design of the instruments the data collection agency relies on the fact that the participation in most government establishment surveys is mandatory. However, recently there has been more and more interest in the improvement of data collection for establishment surveys to reduce response burden and increase data quality (e.g. Goldenberg et al. 2002, Jones 2003, Hak & Willimack 2003).

A similar trend can be seen at Statistics Netherlands (SN). In a time of decreasing resources, the efficiency of the production of statistical information is of utmost concern. To work more efficiently Statistics Netherlands has redesigned the statistical process for the Structural Business Statistics (SBS). In 1998 a project was started with the aim to integrate and standardize the data collection, the editing, and the publication of the SBS. This project is now known as IMPECT (IMPlementation of the EConomical Transformation process). In the first years of IMPECT emphasis was put on creating the logistics of the system. Recently, the attention has moved to the evaluation and improvement of the content of the questionnaires. For 2004 an evaluation and revision of

1 The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

the SBS questionnaires is planned. This paper describes the strategy we developed to evaluate this set of almost 200 questionnaires which vary depending on branch and size of the establishments surveyed.

2. Goal of the evaluation

The goal of the evaluation is to improve the phrasing of the questions and the lay out of the questionnaires in order to increase the quality of the data and reduce the response burden[2]. These goals may prove to be conflicting in many situations, as better measurement at the micro level often means more detailed questioning.

3. The Structural Business Statistics Questionnaires

The SBS questionnaires, also known as the 'Production Surveys', measure a number of indicators of the activity and performance of enterprises in manufacturing, construction, trade, transport, commercial services, energy and water. Variables collected include detailed information on turnover and expenditure of the past year. Important sources for the asked information are the business balance sheets and profit-and-loss accounts. The specification of items such as personnel costs and housing or stock value, however, calls for a consultation of other administrative records.

The data is collected by mail-out, mail-back forms. Each questionnaire is sent out with a so called 'remark sheet' which respondents should use to make comments about the questionnaires, make changes in the name or address of the firm or to ask for delay of the deadline for sending back the questionnaire. Responding to the SBS questionnaires is mandatory. Of all SN establishment surveys, the SBS rank second with respect to response burden, measured as the time needed to fill out the questionnaire. Questionnaires of more than 15 pages are typical. All size classes are covered, but for smaller firms sampling is used. In 2002 almost 80,000 questionnaires were sent out, with a response rate of 68%.

One of the goals of IMPECT was to uniformize the SBS questionnaires in order to gain more efficiency in the data collection process. The process of data collection has indeed been standardized completely. All questionnaires are automatically generated from a system called LogiQuest. The questionnaires are uniformized to a large extent at the level of the *variables* measured. However, similar variables can (and must) be measured with

2 Other projects at SN explore different ways to reduce response burden, for example the possibilities of automatic tapping of records and reducing the number of questions.

different questions or response categories, since it would be rather difficult to specify the turnover of a shoe store and a construction company with the same items. Thus, all SBS questionnaires have a uniform part that is the same for all branches, and a part with branch specific questions. Within each branch there is also a short and a longer form, depending on the size of the businesses (or more correctly: the level of detail needed to construct the statistics). The combination of size and branch specific *questions* results in 183 different questionnaires. Within these 183 groups of questionnaires forms are not always identical, as they may contain product lists that are uniquely compiled for a specific establishment, according to information already available from that firm.

4. Strategy for Testing

We have chosen a strategy that mixes both quantitative and qualitative methods and where the cheaper methods are used to prioritize the more expensive qualitative methods. Given the large amount of different SBS questionnaires and the heterogeneity of respondents it is simply not feasible to test all different questionnaires qualitatively. Fortunately, as the questionnaires have already been in the field, there is quite some process and survey data available that we can use in our evaluation. This triangulation is comparable to the formative evaluation described by the UK Office of National Statistics (ONS) as part of their framework of reviewing data collection instruments in business surveys (Jones, 2003).

4.1 Office based analysis of the questionnaires

Our first step in the evaluation is an office based analysis of the questionnaires. The SBS questionnaires have been in the field for three years now. This means that there are survey data and process information available that can be used to make inferences about the questionnaire. We will also use qualitative data from different sources available at SN, such as the data editors who work with the questionnaire. The goal is to reach an empirically grounded overall analysis of the SBS questionnaires. After this round we should know which questionnaires, questions and types of respondents are most problematic with respect to data quality and response burden.

Survey data and process data

Differences in *unit response* may be an indicator of problematic questionnaires. We must therefore take into consideration how branch and size class characteristics relate to unit response for the SBS questionnaires. If there are groups with particularly low or late response rates, it will be useful to further investigate whether these differences can be attributed to characteristics of the questionnaires. For example, one could test the relation

between the number of specific questions in a questionnaire and the likelihood of a timely response. Patterns in *data quality* may be another important indicator of the quality of the questionnaire. We will examine three ways to operationalize the data quality at the level of questions: 1) items non-response, 2) plausibility of the data as calculated in the editing process 3) percentage of changes made in the data during the editing process. If we manage to develop useful quality measures at the item level, these data present excellent material to investigate the effect of questionnaire and respondent characteristics on data quality.

Content analysis of respondents' remarks and filled out questionnaires

All remarks made by respondents about the questionnaire are documented in *LogiQuest*. This system contains both the information provided on the so called 'remark sheet' (see paragraph 3) as well as remarks about the questionnaire that are made to the call center staff. So far we have not systematically looked into that data base and it will be interesting to see if a content analysis of these remarks will provide useful information for the evaluation of questionnaires.

At ONS samples of questionnaire images are analyzed as part of the process of questionnaire evaluation (Jones, Williams & Thomas, 2003). A first look at some filled out SBS questionnaires shows that crossed out questions, accolades written in the margin to group specified items and comments about the questions give interesting insight in parts of the response process. A very attractive feature of this analysis is that it can be done systematically.

Interviews with SN staff

In their work with respondents, questionnaires and the collected data, employees from different departments of SN have gained insights into possible strength and weaknesses of the questionnaires. We will organize a round of focus groups and open interviews to make these ideas and information available to our evaluation. Four types of informants can be distinguished:

Interviews with our *field officers* - who visit non-responding firms and sometimes help firms to fill out questionnaires - have proved very useful in previous projects (Snijkers, 2000; Giesen, 2003). Rowlands, Eldrigde and Williams (2002) found that *data editors* also provide important and new insights to questionnaire problems. A third group of relevant informants are the *call center staff* who make the non-response follow up calls and answer the first helpdesk request by respondents. A last, but not least important group of SN employees to talk to are the *users of the data*, the people working on the analyses and publication of the data. They may know of patterns in the data that indicate

problematic questions. Experiences with household surveys show that the data users can provide important points of interest for testing and evaluation. Also, interviews with data users on their ideas of possible flaws in the questionnaires present a great opportunity to involve them in the evaluation of 'their' questionnaires. This will hopefully help create commitment among this group for any changes in the questionnaire made later on.

Expert Review

Next to the analyses of the existing information about how the questionnaires work in the field, we will give a small sample of typical SBS questionnaires to experts in the field of questionnaire design. If possible, we will present the questionnaires together with the results of the review described above and a first concept of the field-tests planned. The experts will be asked to comment on the questionnaires, our conclusions and plans so far, as well as to come up with possible solutions for problems already discovered.

4.2 Diagnoses of questionnaire problems in the field

The round of office based analyses should provide us with a good overview of the most problematic questions and questionnaires and the groups of respondents where these problems occur most. Some of these problems may be straightforward and it will be easy to decide if and how they can be solved. In other cases we will need information from respondents to analyze why questions do not work for them and how we can improve these. We will use the results of the office based analyses to decide where we will focus our fieldwork. When prioritizing our limited capacity for field testing, we will also consider practicalities such as the importance of problematic questions or groups of respondents for the output of the survey and whether or not a question can be changed.

The goal of this second step is to diagnose the problems found and look for possible remedies. Recent experiences at SN have given us a good idea on how we can collect useful information on problems with establishment questionnaires by studying the response process in the field. A pilot at SN by Hak and van Sebille (2002) has shown that focused on-site interviews yield useful insights in problems of the SBS questionnaires. For this pilot four constructing companies that were known as good respondents were visited. The goal of the focused interviews was to approach an observation of the actual process of filling out the questionnaire as close as possible. With the already completed and returned questionnaire at hand, the researcher and field officer reconstructed the response process with the respondent. This meant that item for item it was assessed whether and how respondents had come to an answer. Respondents were able and willing to explain if and how they had estimated or calculated the numbers given. This detailed

information revealed misinterpretations of questions and definitions and satisficing behavior.

One of the conclusions of this pilot was that on-site observation of the actual response process might very well be possible. This was successfully tried in the field in the context of the evaluation of the Transportation Survey (Giesen, 2003). Here a methodologist, a field officer and a camera man visited three respondents. The respondents were observed and filmed while filling out an electronic questionnaire.

During these visits we had three main goals: observing what respondents do, understanding why they do it and collecting good survey data. Firstly, we wanted to observe how respondents go about when they work with the questionnaire. For this purpose we encouraged the respondents to start with the questionnaire as they would if we had not been present. During this phase we tried to restrict the interaction with respondent to questions that were necessary to clarify what the respondent was doing at the time ("What are you looking for now?" or "What are those records?"). Secondly, we needed insight in *why* respondents filled out the questions the way they did and how they evaluated the instrument. For this purpose, after the completion of the questionnaire, we asked the respondents how they had understood and answered crucial questions and how they felt about the user-friendliness of the instrument. Thirdly, as real data were collected in these sessions and respondents were likely to have to complete similar questionnaires in the future, we wanted to correct errors respondents had made and to explain how they should have done it.

It proved rather difficult to strictly distinguish the three phases and goals of the visit. Especially when respondents got stuck in the questionnaire it was sometimes impossible not to intervene eventually. Without a doubt, our mere presence and our interventions will have influenced the motivation of the respondents and the ease of filling out the questionnaire. However, even with this bias, we gained insight in where and how respondents made errors in the questionnaire and what aspects of the questionnaire were particularly burdensome. We believe some of the problems found, especially with respect to navigational issues, could only have been obtained by some form of on-site behavioral observation.

Our experiences so far indicate that on-site observation or – if actual observation is not possible – retrospective focused interviews on-site, yield rich and useful data to evaluate and improve questionnaires. It is needless to say that these rich data come at a high cost. It may take several days to organize and actually do an on-site observation.

4.3 Development and testing of improved questionnaire

A round of qualitative fieldwork with on-site observation should result in recommendations for the improvement of the SBS questionnaires. These changes in the questionnaire have to be tested with respondents, to make sure that the changes are in fact improvements for the problems found and have not created new problems. The scope and methods for this test round will be developed when we have the results from the first two steps. Than we will know how many changes have been made to the questionnaires and we can assess the risks of these changes.

5. Future plans

The implementation and further development of the testing strategy will undoubtedly give us a lot of information about the usefulness of our evaluation methods. Besides the practical goal to improve the SBS questionnaires, we also have a research goal to increase our insights in the response process of establishments and the best ways to study these processes. In this research project questions will be addressed such as: Can we indeed distinguish problematic questions by desk research? Which of the sources of information available about questionnaires in the field are more useful to evaluate questionnaires and which are less useful? Can and should we incorporate these kinds of evaluation in a process of systematic review?

Literature

Dillman, D. A., 2000: Procedures for conducting government-sponsored establishment surveys: comparisons of the total design method (TDM), a traditional cost-compensations model, and tailored design. In ICES II, Proceedings of the Second International Conference on Establishment Surveys, Alexandria, VA: American Statistical Association, pp. 343-352.

Giesen, D., 2003: Het gebruiksgemak van de elektronische vragenlijst Verkeer en Vervoer 2003. [The user-friendliness of the electronic transportation questionnaire 2003] Heerlen: Statistics Netherlands.

Goldenberg, K. L./Anderson, A. E./Willimack, D.K./Freedman, S.R./Rutchnick, R.H./Moy, L.M., 2002: Experiences Implementing Establishment Survey Questionnaire Development and Testing at Selected U.S. Government Agencies. Paper presented at QDET conference, June 2002, Charleston, SC.

Hak, T./van Sebille, M., 2002: Het respons proces bij bedrijfsenquêtes. Verslag van een pilot studie. [The response process in establishment surveys. Report of a pilot study] Rotterdam/Voorburg: Erasmus Research Institute of Management/Statistics Netherlands.

Jones, J., 2003: A framework for evaluating and redesigning data collection instruments. Survey Methodology Bulletin. Office for Nationals Statistics, 61:4-9

Jones, J./Williams, S./Thomas, M., 2003: The Use of Administrative Sources in the Evaluation & Design of Data Collection Instruments. Paper presented at the QUEST Workshop, October 21-23 2003, Mannheim, Germany.

Jong, A. de, 2002: Uni-edit: Standardized processing of structural business statistics in the Netherlands. Paper presented at Conference of European Statisticians, UNECE Work Session on Statistical Data Editing, May 2002, Helsinki, Finland.

Rowlands, O./Eldridge, J/Williams, S., 2002: Expert review followed by interviews with editing staff – effective first steps in the testing process for business surveys. Paper presented at QDET conference, June 2002, Charleston, SC.

Snijkers, G., 2000: Eindverslag focusgroepen met nieuw concept omslagvel PS (IMPECT). [Final report focus groups on new concept cover Production Survey IMPECT] Heerlen: Statistics Netherlands.

Contact

Deirdre Giesen
Statistics Netherlands
Kloosterweg 1
NL-6412 CN Heerlen
The Netherlands
email: igin@cbs.nl

IMPROVING BUSINESS SURVEY DATA COLLECTION INSTRUMENTS: GOVERNANCE AND METHODOLOGIES

JACQUI JONES

Introduction

Over the past two years Data Collection Methodology, Methodology Group, in the UK Office for National Statistics (ONS) has researched, developed and implemented methodologies for improving business survey data collection instruments[1]. As part of this, in April 2003 ONS agreed a six year programme for reviewing all ONS statutory business survey paper self-completion data collection instruments (Jones & Scott, 2003). The reviews have an agreed governance structure and are undertaken by conducting the processes in the framework for reviewing data collection instruments in business surveys (Jones, 2003).

The objective of this paper is to provide an overview of this work by outlining:

- the characteristics of business survey data collection instruments in the ONS
- how improvements have been made to the development of business survey data collection instruments
- the framework for reviewing business survey data collection instruments

An overview of business surveys in the Office for National Statistics

The Office for National Statistics has approximately 90 statutory business surveys with 913 different data collection instruments (excluding Telephone Data Entry (TDE) data collection). The primary mode of data collection is paper self-completion. Every year 1.5 million paper instruments are sent to businesses. For some surveys that collect less than

1 For the purpose of the paper the term data collection instrument(s) refers to the questionnaire(s). This includes questions and the guidance notes.

nine data items TDE is also offered as an alternative mode of data collection. There are also a very small number of surveys that use TDE as the only mode of data collection. For example, the Vacancies Survey.

ONS's business survey paper self-completion data collection instruments are typically designed as administrative forms rather than dialogues with respondents. Even the terminology used characterises this. For example the use of the terms forms, inquiries and contributors rather than questionnaires, surveys and respondents. On the positive side a one data collection instrument fits all approach is not implemented for all business surveys. There are many instances of data collection instruments being designed for specific industries or groups of industries.

Prior to 2001 there was no data collection methodology provision for business surveys. No standard method for reviewing and developing business survey data collection instruments. Little interaction with respondents in either a pre-field or field environment. Instead internal results and validation persons were responsible for developing and designing data collection instruments. Development generally occurred in a piecemeal fashion as respondents questioned facets of the data collection instrument and/or data requirements changed.

The design facets of existing business survey paper self-completion data collection instruments

Prior to 2003 business survey paper self-completion data collection instruments were predominately designed from an administrative form and scanning perspective rather than a respondent perspective. Combined with this there were no standards for facets such as question construction, and instrument layout & instructions.

Most of the existing business survey data collection instruments demonstrate the following design facets:

Layout and formatting:

- printed on white paper
- use black and red ink
- data capture boxes in red ink outline
- no standards for formatting of font emphasis. For example the use of bold, italics and underlining
- very little spacing between questions

Instrument instructions:

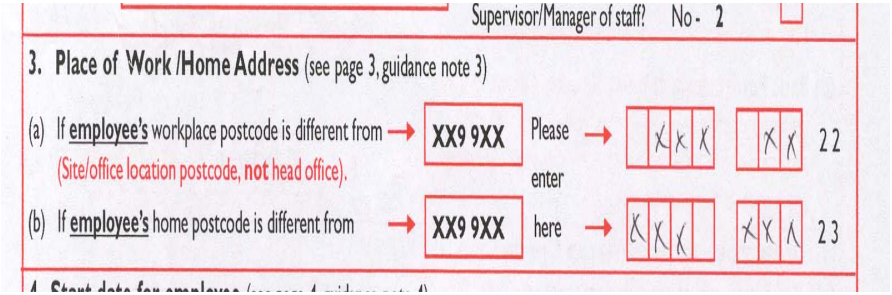
- no standard instructions for instrument completion
- no routing around not applicable questions
- respondents asked to complete the instrument in black ink
- the use of **X** rather than tick to indicate a chosen response

Several of the data collection instruments also demonstrate one or more of the following design facets:

Questions:

- questions with two questions in one

For example, the New Earnings Survey postcode question:



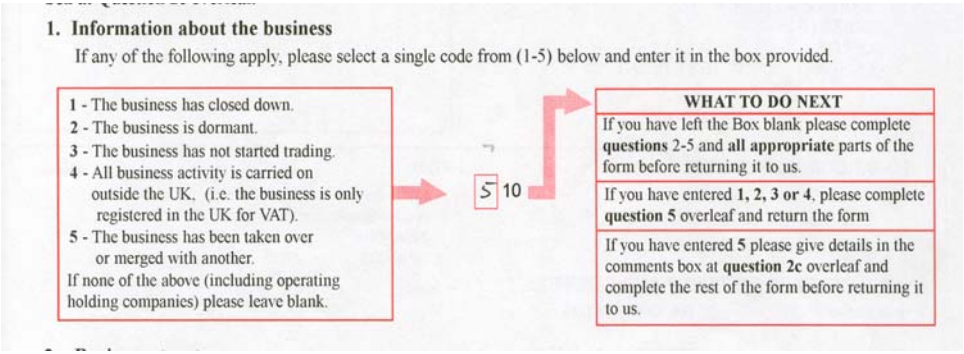
- no question just a heading with response boxes

For example, the Long-Term Insurance questionnaires:

6. UK Corporate Securities
6.1 Other than shares
6.1.1 UK corporate sterling bonds
6.1.1.1 Issued by banks.....
6.1.1.2 Issued by building societies.....
6.1.1.3 Other UK corporates.....

- questions where respondents are asked to select a response category, then the response code and then enter the response code

For example, the Annual Register Inquiry:



1. Information about the business

If any of the following apply, please select a single code from (1-5) below and enter it in the box provided.

1 - The business has closed down.
2 - The business is dormant.
3 - The business has not started trading.
4 - All business activity is carried on outside the UK, (i.e. the business is only registered in the UK for VAT).
5 - The business has been taken over or merged with another.
If none of the above (including operating holding companies) please leave blank.

5 10

WHAT TO DO NEXT
If you have left the Box blank please complete **questions** 2-5 and **all appropriate** parts of the form before returning it to us.
If you have entered **1, 2, 3 or 4**, please complete **question 5** overleaf and return the form
If you have entered **5** please give details in the comments box at **question 2c** overleaf and complete the rest of the form before returning it to us.

Possible impact from the design facets of existing business survey paper self-completion data collection instruments

Much has been written about the importance of data collection instrument design and the detrimental effects that poor design can have on the final survey results (Dillman, 2000; Biemer & Fecso, 1995). Data collection instrument design and impact are both multi-faceted. For example, instrument design consists of facets such as question construction, appropriate response categories, position of guidance notes, layout of questionnaire and survey introduction. Design impact can affect facets such as measurement error, non-response error, processing error and respondent burden. It is important to address each facet of instrument design and impact separately. Each facet needs to be researched and evaluated to ensure that instrument design minimises the negative aspects of the facets of impact. Lohr (1999) highlights this point by stating that:

> *"The quality of survey data is largely determined at the design stage. Fisher's (1938) words about experiments apply equally well to the design of sample surveys: "To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of". Any survey budget needs to allocate sufficient resources for survey design and for nonresponse follow-up. Do not scrimp on the survey design; every hour spent on design may save weeks of remorse later" (page 262).*

As described in the previous section, the characteristics of existing ONS business survey paper self-completion data collection instruments often demonstrate facets of instrument design that could have a negative result to one or more of the facets of impact. Understanding these facets is part of the on-going research to improve business survey data collection instruments.

Improving business survey data collection instruments

In 2001 a data collection methodology branch for business surveys was established in the Methodology Group of ONS. Since 2002 this branch has been working with business survey areas to improve the process of designing and developing business survey data collection instruments for all modes. A major part of this work has been:

- agreeing and establishing governance for business survey data collection instruments
- researching and developing standards for the design of business survey data collection instruments. For example, standard instructions, layout and formatting
- researching, developing and implementing a standard framework of methodological processes for reviewing business survey data collection instruments

The governance of business survey data collection instruments is now the responsibility of the Business Questionnaire Steering Group that is supported by survey specific project boards and working groups as each survey is reviewed. Membership of the Steering Group, project boards and working groups generally consists of persons from the data validation unit; results, analysis and publication; information management; the forms processing centre; communication division; and methodology group.

Research and development of standards for the design of business survey data collection instruments are now at the stage of implementation for instruments currently being reviewed. The standards are now being drafted for use within the office.

The framework for reviewing data collection instruments in business surveys was first used in 2002. In March 2003 it was reviewed and slight adjustments made to it. To date the framework has been used for reviewing the E-Commerce Survey, New Earnings Survey, Annual Register Inquiry and Long-term Insurance questionnaires & General Insurance questionnaires. The E-Commerce Survey is the furthest progressed having completed all stages of the framework. The evidence from this work shows that the stages and processes carried out provide a valuable evidence based approach to the design of data collection instruments. The results of the E-Commerce Survey stage 5 post implementation evaluation shows that reductions have been made to the measurement error, non-response error, processing error and respondent burden.

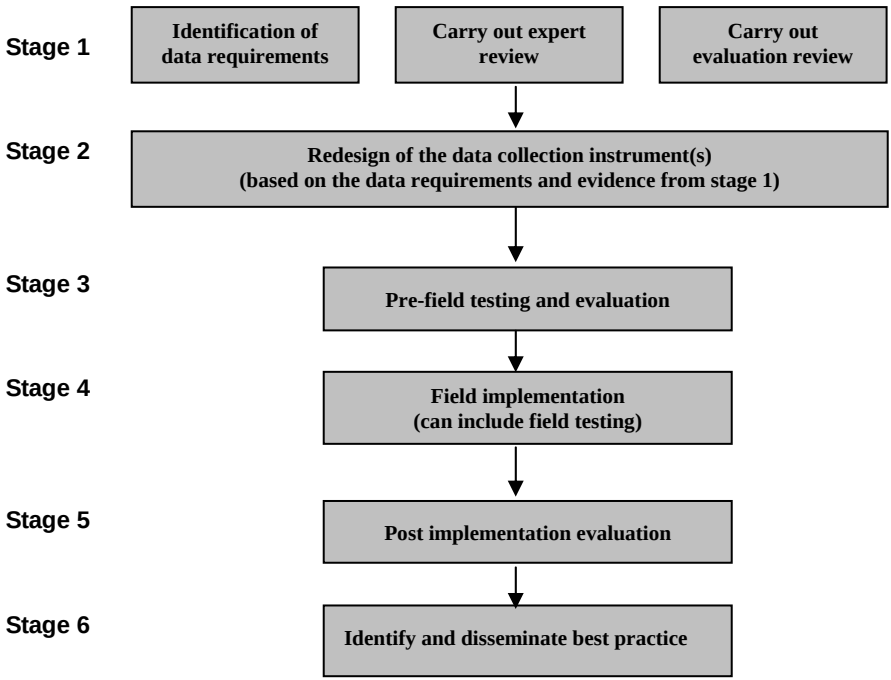
It is envisaged that the framework will continue to be reviewed on an annual basis to ensure that the best approach is maintained.

The framework for reviewing data collection instruments in business surveys

The objectives of the framework (figure 1) are to use a system of stages and processes, which:

- provides evidence to inform decision making in the design of data collection instrument(s)
- provides a consistent, comprehensive, timely and reliable approach to reviews
- pre-field tests the data collection instrument(s)

Figure 1: A framework for reviewing data collection instruments in business surveys



The stages of the framework

Stage 1: identifying data requirements, expert review and evaluation review

This stage of the framework involves three simultaneous processes. In process 1 (the identification of data requirements) the relevant survey output manager is responsible for carrying out a consultation exercise to identify and agree the objectives and data requirements of the survey. For the New Earnings Survey this involved contacting users and asking them to respond to the following questions:

(1) What data do you require?
(2) Why are the data required?
(3) How often are the data required?
(4) What alternative sources of data are available?
(5) Do alternative sources of data meet your data requirements (if not, why not)?

The output manager then assesses the collected information and further user consultations maybe undertaken to agree the objectives and data requirements for the survey. These requirements need to be signed off before stage 2 of the framework can commence. If this is not done then it is impossible to redesign the data collection instrument(s) in stage 2 of the framework.

In process 2 (carry out expert review) at least one expert review is carried out. At a minimum a data collection expert in Methodology Group will carry this out. Where possible an independent data collection expert and a subject-matter specialist will also be asked to carry out an expert review. An expert review is a desktop review of the design of the existing data collection instrument(s). The review focuses on all the design facets of the instrument(s)[2].

Process 3 (carry out evaluation review), is a formative evaluation. It triangulates data from a variety of qualitative and quantitative sources. For example, unit non-response, item non-response and respondent feedback.

Stage 2: redesign of the data collection instrument(s)

Stage 2 of the framework looks at the agreed objectives and data requirements of the survey and based on the evidence produced from analysis of the expert review and evaluation review identifies facets of the data collection instrument design that require further development. During this stage the data collection instrument(s) are redesigned.

---

2 For example, the wording of questions, response categories, instructions, definitions and the layout and formatting of the instrument.

The redesigned data collection instrument(s) are then presented to the relevant output manager to clarify that the data requirements for the survey are being met. Where appropriate other experts are consulted. For example, the ONS accountancy advisor and National Accounts. Prior to moving into stage 3 of the framework the data collection instrument(s) must be signed off by the survey output manager and Methodology Group.

Stage 3: pre-field testing and evaluation

Stage 3 involves pre-field testing and evaluation. To date cognitive interviews using either concurrent or retrospective probing have been the pre-field testing methods used. However, in November 2003 the first focus group with business respondents was successfully conducted. Both cognitive interviewing and focus groups are new methodologies for ONS business surveys. When conducting cognitive interviewing people in the business survey areas are also involved in carry out the cognitive interviews. Methodology Group is responsible for providing training, conducting at least one cognitive interview per iteration, quality assuring the cognitive interviews and analysing the qualitative data.

The cognitive interviewing training is divided into two sessions each lasting approximately two and a half-hours. The first training session outlines:

- what cognitive interviews are
- why we conduct cognitive interviews
- when to conduct cognitive interviews
- how to conduct cognitive interviews

The session also gives participants the opportunity to practice cognitive interviews using a variety of probes[3] and receive feedback on their cognitive interviewing skills.

The second training session takes place two days after the first training session. This time interval has been found to be the most effective in allowing participants time to consider and practice their probing techniques. It gives participants further opportunities to practice cognitive interviewing and receive feedback. It is also used to standardise the cognitive interviewing by developing standard probes. This ensures that all interviewers systematically probe specific design facets of the data collection instrument(s).

In the agreed timetable the pre-field cognitive interviews are carried out in at least four iterations. Each iteration incorporates at least four face-to-face cognitive interviews, an interviewer debriefing session, analysis of the data and further development of the data

3 Probes such as comprehension, paraphrasing, confidence judgement and recall.

collection instrument(s). Interviewing in carried out by two interviews and recorded using a mini disc player. Methodology Group quality assure this process by carrying out at least one interview per iteration, leading the debriefing sessions, analysing the qualitative data and redeveloping the data collection instrument(s) based on analysis of the data. If there are still design issues at the end of the fourth iteration further face-to-face cognitive interviewing or telephone interviews are carried out. At the end of stage 3 the data collection instrument(s) is signed-off for implementation. The survey specific working group & project board and the Business Questionnaire Steering Group undertake sign-off.

Stage 4: field implementation

Stage 4 implements the data collection instrument(s) into the field. This stage involves the printing and despatch of the instrument(s), the capture and validation of the returned data, dealing with respondent queries, response chasing and the analysis of the data. Throughout this stage pre-defined data is collected for use in stage 5 (post implementation evaluation). For example, respondent queries on a question by question basis, problems in data capture and problems in data validation.

To date implementation has not included any field-testing. However, in 2004 ONS will be piloting field testing in stage 4 (field implementation) of the reviewed data collection instruments for the Business Register Survey (previously named the Annual Register Inquiry) and the Annual Survey of Hours and Earnings (previously named the New Earnings Survey).

Stage 5: post implementation evaluation

Summative evaluation is carried out in this stage. The data collected during stage 4 field implementation plus data such as unit and item non-response are collated and analysed to identify any aspects of the design of the data collection instrument(s) that show:

- measurement error
- non-response error
- processing error
- unnecessary respondent burden

Follow-up telephone interviews with respondents to the survey are also carried out. If the analysis provides evidence of any type of error or respondent burden with the data collection instrument design then the procedure is to move back to stage 2 of the framework and work through the processes again.

Stage 6: identify and disseminate best practice

This stage involves both the identification of best practice for the processes involved in the framework and the design of data collection instruments in business surveys. Best practice is disseminated through ONS standards and guidance. For the processes involved in the framework this will provide evidence to inform the annual review of the framework. For facets of the design of data collection instruments this should ensure that stage 1 expert reviews and stage 2 redesign of the data collection instruments are based on best practice evidence.

Next Steps

The governance structure and framework for reviewing data collection instruments in business surveys will continue to be used. In 2004 to 2005 the reviewed and redesigned data collection instruments for the Business Register Survey and Annual Survey of Hours and Earnings will be field tested. Post-implementation evaluation will proceed the field tests. The next round of reviews are commencing. The Business Questionnaire Steering Group has agreed the following reviews for 2004:

- Quarterly Profits Inquiry
- Research & Development
- International Trade in Services
- Monthly Production Inquiry
- Retail Sales Index
- Monthly Inquiry into Distribution and Services Sector
- The 12 monthly review of the framework will also take place.

References

Babbie, E., 1995: The Practice of Social Research Wadsworth Publishing Company.

Biemer, P./Fecso, R., 1995: Evaluating and Controlling Measurement Error in Business Surveys in: B. Cox et al. (eds), Business Survey Methods, John Wiley & Sons, Inc. pp. 257-282.

Bryman, A., 1988: Quantity and Quality in Social Research Unwin Hyman.

Converse, J./Presser, S., 1986: Survey Questions: Handcrafting the Standardized Questionnaire Sage University Paper Series on Quantitative Applications in the Social Sciences.

DeMaio, T. J. (ed), 1983: Approaches to Developing Questionnaires Statistical Policy Working Paper 10, United States Office of Management and Budget, Washington D.C.

Dillman, D., 2000: Mail and Internet Surveys John Wiley & Son.

Dippo, C. S. et al., 1995: Designing the Data Collection Process in: B. Cox et al (eds), Business Survey Methods, John Wiley and Sons, Inc. pp. 283 – 301.

Eldridge, J./Martin, J./White, A., 2000: The use of Cognitive Methods to Improve Establishment Surveys, in: Britain The Second International Conference on Establishment Surveys (ICES11) Buffalo, New York, USA.

Fowler, F. J., 1995: Improving Survey Questions: Design and Evaluation Applied Social Research Methods Series 38, Sage Publications.

Gower, A. R., 1994: Questionnaire Design for Business Surveys Survey Methodology 20, pp. 125 - 136.

Jones, J./Scott, L., 2003: The Review Programme for Business Survey Data Collection Instruments Survey Methodology Bulletin, No. 52, pp. 1 - 3, Office for National Statistics, UK/

Jones, J., 2003: A Framework for Reviewing Data Collection Instruments in Business Surveys Survey Methodology Bulletin, No. 52, pp. 4 - 9, Office for National Statistics, UK.

Lohr, S., 1999: Sampling: Design and Analysis Duxbury Press.

Patton, M., 1990: Qualitative Evaluation and Research Methods Sage.

Statistics Canada, 1994: Policy on the Development, Testing and Evaluation of Questionnaires.

US Bureau, 1998: Pretesting Policy and Options: Demographic Surveys at the Census Bureau U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census.

Contact

Jacqui Jones
Office for National Statistics
Room D136
Government Buildings
Cardiff Road
Newport
Wales
United Kingdom
email: Jacqui.Jones@ons.gsi.gov.uk

A VALUABLE VEHICLE FOR QUESTION TESTING IN A FIELD ENVIRONMENT: THE U.S. CENSUS BUREAU'S QUESTIONNAIRE DESIGN EXPERIMENTAL RESEARCH SURVEY[1]

JENNIFER M. ROTHGEB

1. Introduction

Survey methodologists within the U.S. Census Bureau conduct questionnaire design research, including questionnaire pretesting and evaluation. Typically, the pretesting research is conducted using cognitive interview methods. Frequently, however, we want to expand that research by conducting "split-sample" field experiments to compare different questionnaire designs, such as different question wording, question sequencing, etc.

In the past, the only available option has been to piggyback onto one of the demographic surveys like the Current Population Survey (CPS) or the Survey of Income and Program Participation (SIPP) which typically presents many constraints (time, procedural, managerial). Usually the lead time for production surveys is too long, researchers are not allowed much control when experiments are piggybacked onto production surveys, and the fact is, that production surveys don't want experiments attached to them. Because of the lack of available field surveys in which to do question testing, in 1998 researchers in the Statistical Research Division (SRD) proposed to establish an independent omnibus demographic household survey intended solely for research purposes. We proposed to call

1 This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

the survey the Questionnaire Design Experimental Research Survey (QDERS). To contain costs, we proposed that the QDERS use a nationally representative Random-Digit-Dialing (RDD) sample, be conducted from one of the Census Bureau's centralized telephone centers, be no longer than 15 minutes in duration, and produce approximately 800-1000 completed interviews. The proposal stated that involved researchers would be responsible for questionnaire development and specifications, developing and administering interviewer training, monitoring the survey's progress, developing the SAS data file and conducting analysis. When presented to senior management within the Census Bureau's research area, the proposal was well received, viewed as an exceptional opportunity for researchers, and provided generous funding.

As well as getting internal funding, in order for QDERS to be implemented, it was necessary to get "buy-in" from the Telephone Center Coordination Staff at Headquarters and the Hagerstown Telephone Center (HTC), which is one of the Census Bureau's centralized calling centers. Once our proposal was posed to these groups, they enthusiastically supported the project.

At the Census Bureau, survey instruments administered through computer-assisted interviewing (CAI) are authored in a group outside the survey methods research area. Researchers do not possess knowledge of the authoring language to program the QDERS instruments. Authoring resources are costly and scarce. In addition, much lead time is usually required for CAI instrument authoring. Given that the production surveys sometimes experience a scarcity of authoring resources, we knew allocation of any available resources would obviously be given to the production surveys rather than a research survey. Since the QDERS project had minimal funding available for instrument authoring (and a short lead time until field implementation) we did not think it was feasible to use a CAI instrument for QDERS.

2. Fielding QDERS

As previously stated, the purpose of QDERS is to serve as a vehicle for SRD researchers to have an opportunity to independently conduct questionnaire design field experiments in a timely manner and within a flexible environment. The flexibility of the HTC staff was instrumental in the success of QDERS. The HTC staff has been very generous in allowing involved researchers free rein over the project. QDERS researchers identify staffing requirements for different interviewer groups for various questionnaire treatments, balancing groups by interviewer experience, skill level, gender, tenure and experience with RDD surveys. The HTC staff permits the researchers to develop the interviewer training package and administer training without getting input or feedback

from the HTC staff. QDERS researchers also request that interviewers in different assigned groups not discuss the questionnaire treatment on which they are working at a specific time during QDERS. To avoid interviewer effects, interviewers are rotated among all questionnaire treatments so all interviewers are exposed to all experimental treatments for the same duration of time during the fielding of QDERS.

The QDERS researchers are also permitted to determine the case management parameters such as how many call attempts will be made to an unreached number, whether refusal conversions calls will be made, etc. In addition, the HTC staff permits the QDERS researchers to develop an interviewer debriefing protocol and conduct the debriefing without HTC involvement and without supervisory staff present during the debriefing.

Provided below is some basic information about the first two implementations of QDERS in 1999 and 2000.

A. QDERS 1999

QDERS 1999 included two treatment groups and a four-week data collection. Twenty-two interviewers participated in QDERS and the interviewers were “flipped” midway through, so all interviewers administered both versions of the questionnaire, but were not exposed to both versions at the same time. Paper and pencil questionnaires were used because resources were not available for authoring of a computer-assisted-telephone-interview (CATI) survey instrument.

QDERS 1999 experiments included examination of the following issues:

- The effects of person-level versus household level questionnaire design strategies on survey estimates and data quality;
- Methodological issues in measuring the uninsured; and,
- Using alternative question strategies to reduce income nonresponse.

QDERS 1999 received additional funding from another division that allowed us to double the planned sample size and also afforded a reinterview to be conducted to provide a reliability measure. Our target was to have 1800 completed interviews (900 of each treatment) and to have 900 completed reinterviews (450 of each treatment).

We used an RDD sample of 5400 sample telephone numbers to produce 1291 completed household interviews. (We had been advised that we needed three sample telephone numbers for each completed interview desired.) The 1999 response rate ranged from 36 to 46 percent (using AAPOR RDD response rate standards), depending on whether cases with unknown eligibility were included in the denominator. In 1999 we did not attempt any refusal conversions. Our targeted number of completed interviews fell short of our

goal. We suspected this was largely due to the lack of refusal conversion attempts. However, because our analyses focused on relative differences between treatment groups, the low response rate was not as large of a concern as it would be had this been a production survey. In addition to the 1291 interviews, we reinterviewed over 900 households.

B. QDERS 2000

QDERS 2000 had four treatment groups, an eight-week data collection period (divided into four ten-day interview periods) and 24 interviewers (split into four groups). In addition to rotating the interviewers among questionnaire versions, we also decided to use sample replicates so that new sample could be released at the beginning of each new interview period. This allowed each group of interviewers to begin work on a different questionnaire treatment using a fresh sample and without any sample cases remaining from the earlier interview periods.

Because of the increased complexity of the QDERS questionnaire design and the number of experimental treatments, we used CATI instruments in QDERS 2000. Also, some of the production surveys for which the experiments were relevant are automated surveys and it was more methodologically sound to conduct experiments using the same mode of interview as that used in the production environment. (We did not have financial resources available for instrument authoring but we were able to acquire the resources through bartering of our research services in exchange for authoring.)

Experiments in QDERS 2000 included examining:

- Question ordering issues related to health insurance;
- Question wording experiments to facilitate pretesting evaluation research;
- Topic-based income reporting versus person-based reporting; and,
- An interviewer training experiment (refusal aversion training).

As with QDERS 1999, we received additional funds in order to double the sample size for QDERS 2000. Our targeted number of completed interviews was 2000 interviews (500 for each treatment). Due to the low response rate in 1999, we decided to make two revisions to how we approached QDERS 2000. First, since the suggested ratio of 3:1 sample cases to completed interviews had proved inadequate, we decided to increase the sample telephone numbers to 8000 in the hopes of reaching our goal of 2000 completed interviews (a 4:1 ratio). Second, based on our experience in 1999, we decided to devote resources to refusal conversion attempts in an effort to boost response rates.

It is worth noting that within one year, QDERS expanded in terms of the types of experiments included. It grew from including only questionnaire design experiments in 1999 to including experiments focused on interviewer training and another on the evaluation of pretesting techniques. In such a short time, researchers were realizing the multiple utility that QDERS permitted.

With QDERS 2000, the sample of 8000 telephone numbers produced 1862 completed household interviews. The response rate ranged from 42 to 52 percent (using AAPOR standards), depending on whether cases with unknown eligibility were included in the denominator. Refusal conversion attempts were made for all households for which an initial refusal was obtained. As part of one of the QDERS experiments, refusal aversion training was provided to some interviewers. No doubt this contributed to the higher response rate. We were disappointed that our goal of 2000 completed interviews was not reached, but we were encouraged that we came much closer to meeting that goal than we had a year earlier.

3. Benefits of QDERS

QDERS has proved to be a valuable tool by which survey researchers can conduct methodological research. Many more controlled split-sample experiments are conducted at the Census Bureau now than ever before. Researchers realize that when they need to follow up some laboratory research with field testing, they now have an available vehicle by which to continue their research. The availability of QDERS has served to stimulate researchers to further develop their research ideas. QDERS has also resulted in more collaboration between survey methodologists and content experts and between researchers inside and outside of the Census Bureau. The availability of QDERS as an independent research vehicle has prompted organizations external to the Census Bureau to provide funding to increase QDERS sample size to provide enough power for certain experiments. Census Bureau researchers have produced more journal articles, book chapters, and conference papers about question design and survey methodology than would have been possible without QDERS.

Some of the QDERS 1999 and 2000 experiments have resulted in the introduction of revised question design and new approaches to interviewer training in some production surveys. One of the experiments within QDERS that was used to evaluate pretesting techniques demonstrated that pretesting does appear to reduce measurement error.

QDERS was conducted again in 2002 and 2003. Preparations are underway for QDERS 2004.

4. QDERS Publications and Conference Papers

The details and results of the QDERS 1999 and 2000 experiments would consume too much space for this paper. For readers interested in specific experiments, I refer you to the publications below which are products of the QDERS project.

Forsyth, B./Rothgeb, J./Willis, G., 2004: Does Question Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. Forthcoming in S. Presser/J. Rothgeb/M. Couper/J. Lessler/E. Martin/J. Martin/E. Singer (Eds.), Methods for Testing and Evaluating Survey Questions. New York: Wiley Interscience.

Hess, J./Moore, J./Pascale, J./Rothgeb, J./Keeley, C., 2002: The Effects of Person-level vs. Household-level Questionnaire Design on Survey Estimates and Data Quality. Public Opinion Quarterly, Winter 2001, University of Chicago Press.

Hess, J./Rothgeb, J./Moore, J./Pascale, J./Keeley, C., 2001: Measures of Functional Limitations: The Effects of Person-level vs. Household-level Questionnaire Design, in: S.Barnartt/B.Altman (Eds.), Exploring Theories and Expanding Methodologies: Research in Social Science and Disability, Volume 2, Oxford, England: Elsevier Science Ltd.

Landreth, A./O'Brien, E., 2001: Respondents' Understanding of the Vague Economic Concept "Cash": A Comparative Study. Presented at the annual meeting of the American Association of Public Opinion Research. Montreal, Canada.

Mayer, T./O'Brien, E., 2001: Interviewer Refusal Aversion Training to Increase Survey Participation, in: Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA.

Moore, J./Loomis, L., 2001: Reducing Income Nonresponse in a Topic-Based Interview. Paper presented at the annual meeting of the American Association of Public Opinion Research, Portland, OR.

Pascale, J., 2002: A Quantitative and Qualitative Assessment of the Data Quality of Health Insurance Measurement Methodologies. Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark.

Pascale, J., 2001: Measuring Private and Public Health Coverage: Results from a Split-Ballot Experiment on Order Effects. Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA.

Pascale, J., 2000: Alternative Questionnaire Design Strategies for Measuring Medicaid Participation. Paper presented at the 128th Annual Meeting of the American Public Health Association.

Pascale, J., 1999: Methodological Issues in Measuring the Uninsured, in Proceedings of the Seventh Health Survey Research Methods Conference (PHS01-1013.) Department of Health and Human Services. Washington, D.C.

Pascale, J., 1999: Effects of Questionnaire Format and Reference Period on Measuring the Uninsured. Paper presented at the 127th Annual Meeting of the American Public Health Association.

Contact

Jennifer M. Rothgeb
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: jennifer.m.rothgeb@census.gov

INTERACTIVE CODING IN THE FIELD: A TEST

RACHEL VIS[1]

1. Introduction

At Statistics Netherlands all coding of answers to open questions from household surveys is currently done by specially trained coders at the office. The interviewers in the field are instructed to probe the respondent in order to get as much information as possible. Still, the interviewers never know for sure whether they have gathered sufficient information. The respondent's answers have to be typed down verbatim. All this information in the form of text strings is sent to the office. The coders have these text strings at their disposal to try and link a specific code to the answer. This can partially be done automatically, by means of a data base in which frequently used text strings are already linked to codes. All text strings that are not present in the data base have to be coded by hand. Here the coders rely on experience and on vast manuscripts with standard classifications.

Interviewers at Statistics Netherlands are not trained to be coders. Of course, a part of their task is to code respondents' answers, but this task is limited to assigning a given answer to one of the response categories of a question. For several reasons they are not asked to code open answers. Coding to the level of at least four digits is a very specific skill, which needs a lot of training and experience. Since there is a significant turnover of interviewers they have no time to gain experience and it would be too costly to give them all a thorough training.

In 2002 a project was started to make the process of coding answers to open questions for respondents' education, occupation and company more efficient. The division of Social and Spatial Statistics and the division of Technology and Facilities joined forces to

1 The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

develop alternative ways of coding. The result was a coding programme which is integrated in the questions. In short this means that after the answer of a respondent is typed in the laptop by the interviewer the coding programme will try to find a corresponding code in a data base on the laptop. The idea behind coding in the field is that the codes are possibly more accurate, because the interviewer can verify the coded answer with the respondent. The interviewer, then, is not an actual coder. Another advantage of this coding programme is that interviewers get feedback whether enough information has been gathered, because the coding programme generates the appropriate follow-up questions until a sufficiently detailed code has been reached.

During the development process the new questions have been tested several times. The coding programme has been thoroughly tested whether it links the right code to the right answer. In August 2002 the new questions were tested by themselves, i.e. not integrated in a survey. This was the first test in which it was tested whether interviewers could work with the coding programme. Some conclusions from this test were: (1) that the instruction of a few hours was too short to familiarise the interviewers with the coding programme, (2) that the coding programme took much time to search through the data bases which caused awkward silences in the dialogues, (3) the interviewers reacted often surprised and out loud to the actions of the computer. After this first test several improvements to the questions and the coding programme were made. These improvements were tested in the follow-up test in April 2003 (Vis/Beukenhorst, 2003), which will be discussed in this paper. In the present test the impact of the coding programme on an interview was investigated as well, i.e. what happens before and after the coding programme has done its work. The interaction between the interviewer and the respondent and between the interviewer and the new questions were the main interest of the test. The aim of the questionnaire lab test was to find out:

1. how long the new questions take,
2. whether the respondents can answer the new questions,
3. how do respondents react to the questions,
4. whether the interviewers can handle the new questions, and
5. whether the new questions disturb the interview.

First we will discuss the questions and some aspects of the coding programme that were tested. After that we will describe how we conducted the test. The methods that were used to analyse the test data will be discussed next. Finally the results from the test will be presented.

2. The questions to be tested

In most household surveys some questions about a respondent's education, occupation and company are asked. The questions are not dealt with elaborately in every survey, but in the Labour Force Survey (LFS) they are most prominently present. The questions are not literally the same in each survey, for this test we took the questions from the LFS. The original open questions in the LFS are:

- *What kind of company or institute do you work for?*
- *What is your occupation or function at this moment? What are your main (executive) activities at work?*
- *Which course or training have you followed? What kind of course or training was it? Was there a specific subject or discipline?*

In this paragraph a basic explanation of the working of the new questions will be given (adapted from Michiels, 2003), illustrated by quotations from several interviews.

2.1 Respondent's company

The first part of the new question-unit is about the respondent's company. If a respondent has indicated that the company he works for has more than 100 employees, the questionnaire will ask for the company's name. The interviewer will enter the answer of the respondent in the questionnaire. The answer is then looked up in a data base of company names. The names in this data base are all linked to a certain SIC-code (Standard Industrial Classification). The coding programme suggests up to six possible options to the interviewer. And the interviewer can consult with the respondent which option suits best. When a company name is selected, the programme stores the corresponding code.

Interviewer: *What is the name of the company or the institute you work for?*
Respondent: *ABN AMRO.*

The coding programme now suggests several possible options to the interviewer, who has to present them to the respondent.

I: *Do you work for the bank itself, or for the insurance office, or for the exchange office, or for the building fund, or for the lease holding?*
R: *No no, for the bank itself.*

When there is no match to the name of the company, or the respondent has indicated that the company has less than 100 employees, the questionnaire asks for a description of the company type. The respondent's answer to this is looked up in a data base with company descriptions, which are also all linked to a classification code.

If there is no match to the company's name or the type of company, then there is one last option to get at least a less detailed code of the type of company. This last option was still in an experimental stage and was first tried in this test. The respondent is asked whether he or she can indicate what sector the company belongs to. The answer has to be looked up in a tree diagram (see below). The idea is that you start high up in the diagram and by probing the respondent you work your way down. Each step down means an extra digit and extra information to the code. For some sectors there is only one step down, but at most there are five steps. Because the diagram is very extensive only the first sector is unfolded:

1. Agriculture and Fishing
    - 1.1. Agriculture
        - 1.1.1. Agriculture and horticulture
        - 1.1.2. Breeding and keeping of animals
        - 1.1.3. Mixed farm
        - 1.1.4. Agricultural service
            - 1.1.4.1. Veterinary
            - 1.1.4.2. Non veterinary
        - 1.1.5. Other
    - 1.2. Fishing
2. Industry and Building
3. Trade, Commercial services and Letting
4. Transport, Storage and Communication
5. Government, Education and Health care
6. Other

2.2 Respondent's occupation

The next part of the new question-unit is about the respondent's occupation. The answers are looked up in a data base of occupation descriptions.

I: *What is your occupation or function at this moment?*
R: *Carpenter.*
I: *The computer is searching… Is it construction carpenter or universal carpenter?*
R: *Construction carpenter.*

If there is no direct hit, the main activities at work are asked. The coding machine will then try to find an occupation that fits the given main activities.

I: *What is your occupation or function at this moment?*
R: *Post office employee.*
I: *What are your main activities?*
R: *Processing post.*
I: *Let me see. Post sorter?*
R: *Yes.*

The programme also combines the answers to occupation and company. This means that not always all questions about occupation have to be asked, since some companies imply just one occupation (or a limited number of occupations). For example, if a respondent says that he works in a barbershop, the coding machine immediately suggests the occupation "hairdresser".

2.3. Respondent's education

The last part of the new question-unit is about the respondent's education. In this part, questions are asked about the respondent's whole educational history from secondary school on. A special feature in this module is that the answer to the first question determines the following question(s). The coding programme generates the appropriate follow-up questions until a code is reached. It is not to be said in advance how many questions are needed to reach a code. It is possible that with one question the exact code corresponding to the educational level is reached:

I: *Which course or training did you follow first?*
R: *Havo. [Specific kind of grammar school.]*
I: *Okay.*

But it can also take longer:

I: *Which course or training did you follow then?*
R: *Law.*
I: *At university I suppose?*
R: *Yes.*
I: *Was there a specific subject? Constitutional law, or…?*
R: *Labour law.*
I: *Labour law, let me see. Yes. Did you do your master's degree or doctoral degree?*
R: *Master's degree.*

There is no fixed order in which to go through the questions. The interviewer can enter the first answer the respondent gives. In the previous example the respondent answered with the subject of his studies, but it is also possible that a respondent first says to what type of school he went before saying which subject he followed. The coding programme is an intelligent programme. In order to reach a certain code it presents the relevant questions to the interviewer. In other words, the programme guides the interviewer through the questions, each time following another route.

There is one drawback to the programme. It only suggests a maximum of six possible hits for each entry. If there are more than six possible hits the programme indicates "too many hits" and either the interviewer has to gather more information from the respondent to get

a more specific answer, or no code is selected and the programme skips to the next question. The answers that have not been coded by the programme will be sent back to the office to be coded by hand.

3. The test

Usually a questionnaire lab test is conducted in the laboratory, yet for this test we wanted to reconstruct a field situation as close as possible, so the test was conducted in the call centre of Statistics Netherlands. This meant that the researchers, the developers and the interviewers' supervisors could be present during the test. And any problem with the questionnaire or any other problem could be solved on the spot. Eventually the new questions will be integrated in both CAPI-questionnaires (Computer Assisted Personal Interview) and CATI-questionnaires (Computer Assisted Telephone Interview).

Six interviewers participated in the test. In the previous lab test the instruction was too short, so for this test the interviewers received a whole day of instructions. At the end of the instruction day each interviewer had had at least three test interviews with test respondents. These test respondents are specifically recruited for participating in tests, and for this they receive a participation fee. Since these respondents knew it was a test the interviewers could familiarise themselves with the questions in a non-stressful way.

The next day the interviews were held with real respondents. The respondents were not aware that they participated in a test, they were asked to participate in the Consumer Sentiments Survey. The new questions about education, occupation and company were integrated in this survey. In order not to surprise the respondents with the elaborate new questions, they received a standard advance letter with an additional announcement: "*The interviewer will also ask you some questions about your education, occupation and job.*" The questions have to be asked for all household members older than 14, in many cases this will be by proxy.

For the test three interviewers made their calls in the afternoon and three interviewers worked in the evening. Between 1 and 9 o'clock in total 42 interviews were completed. Of all interviews a recording was made.

Afterwards there was an evaluation with the interviewers.

4. Methods used

The interviewers evaluated the test during two focus groups. The aim of these focus groups was to find out how the interviewers had experienced the interviews, and what they thought of the new questions.

To analyse the interviews conversation analysis was used. In conversation analysis an interview is taped and afterwards transcribed and analysed by the researcher (Visschers, 2002). For this test the sections with the new questions about education, occupation and company have been literally transcribed. The dialogues contain information about the interview and the interactions. Does the respondent understand the question? Does the interviewer understand the respondent's answer? Does the interviewer read out the question as worded? For more information about conversation analysis see Houtkoop-Steenstra (2000) or Psathas (1995).

For this test it was also possible to reconstruct the actions of the interviewer in the questionnaire by following the dialogue. The results from conversation analyses can be compared to the remarks of the interviewers, e.g. do the interviewers see the mistakes that were made and do they remark on them?

5. Analysis

5.1. Length of the interview

Whenever a new question is tested the length is a point of concern. The length of the interviews can be analysed by comparing the old interview time to the new. During the test it appeared that in some interviews the new questions about education, occupation and company took longer than the original questions. In some interviews the section with the new questions lasted up to 15 minutes. An explanation for this could be that, because it was a test, the interviewers wanted to be more thorough than usual. Still the dialogues should be analysed to see whether there is another explanation.

The questions about education for elderly respondents, and the tree diagram question in general caused the longest dialogues. This was because these questions triggered a discussion between the interviewer and the respondent. The following dialogue lasted about 6 and a half minutes, I will cite only a fragment:

…
I: Did you follow that course at university?
R: No, not at university.
I: Then what was it?

R: Well, I cannot remember.
I: Was it the LOI or something? [Distance learning]
R: No.
I: You did go to school?
R: Yes.
I: Was it your intention to become a teacher?
…

The interviewer has to get a usable answer, in order to get through the questionnaire. Yet the respondent has problems recalling this detailed information. He is talking about a short course he did almost forty years ago, which was not even part of his main education.

The dialogues for the tree diagram take long for several reasons. The respondent does not always understand what "company sector" means, and sometimes the interviewer has to read out a lot of options. Especially "Industry and Building" has a lot of sub-items, and when a respondent doesn't immediately say to which sub-sector the company belongs, the interviewer has to read out all possible options. Very often these dialogues do not even lead to a code, and even if there is a code at the end of the sequence, it is doubtfully the right one.

An encouraging finding for the future was that the length of the interviews decreased over time. The length of the interview apparently depends upon the interviewer's experience with the new questions.

Even if the length of the interview decreases, it is not said that these discussions will not occur any more. We concluded that the questionnaire should be improved so that there is an elegant escape route to get out of these endless discussions, for both the education questions and the tree diagram.

5.2. Can the respondents answer the new questions?

In the dialogues there were several instances in which the respondent was not able to answer the question. Most frequently this happened for the tree diagram question. Respondents do not often ask for clarification, but the misconception shows from the given answers. The answers almost never fitted one of the answer categories or sectors (see paragraph 2.1). Some examples of answers to the question: "*Can you say to which sector the company you work for belongs*?"

"Conductor of an orchestra."
"Seafood sector."
"Building company."
"Cocoa factory."

The answer categories in the tree diagram are literally taken from the classifications that are used at the office. A respondent, nor the interviewer, thinks in terms of such classifications. Copying the literal terms from the classifications into the answer categories caused other misinterpretations as well:

I: Can you say to which sector the company you work for belongs?
R: Chicken farm.
I: So that is agriculture.
R: Yes.
I: Is it breeding and keeping of animals?
R: No.
I: No?
R: I only keep chickens, I don't breed them.

A similar situation occurred when the respondent and the interviewer interpreted the category "wholesale business, trade mediation" as two activities belonging together. Since the respondent worked for a wholesaler the category was not chosen and a lengthy search for the right code commenced. It is not clear to the interviewer and the respondent that the comma in the answer category means "and/or".

Our recommendation is that the answer categories should be made more respondent and interviewer friendly and they should be closer to a respondent's train of thought.

5.3. Reactions of the respondents

Before the test one of the risks we anticipated was that the respondents would wonder "how do they know all that?" The respondent is not aware of the fact that the interviewer has a lot of information available in the coding programme. The hypothesis was that a respondent could feel it as a violation of his privacy if, for example, the computer would know that a respondent works in a bakery, after he had given only some minor information about the company's name and size. The fear was idle, this situation never occurred. In fact, several respondents were pleasantly surprised that the interviewer seemed to know what they were talking about. In the focus group interviewers made similar remarks on this. In these situations the coding programme made them feel more secure, and they could show more interest in the respondent.

5.4. Can the interviewers handle the new questions?

Even with the longer instruction, the interviewers learned most by simply doing the interviews. Clearly the interviewers got more confident with the questions during the day. This resulted in shorter interviews over time. The coding programme was completely new to the interviewers. At first they had to search for the right commands to give to the

computer, later this became a routine. The interviewers were surprised by every move the computer made and they were confused because the computer did not follow one specific route through the questions. Later this became a bit easier, but it was difficult for the interviewers to rely on the programme to guide them. One interviewer got in a bit of trouble, because she assumed that there was a fixed routing in the questions:

I: *What kind of company or institute do you work for?*
R: *For an architectural firm.*
I: *Let me see, I now have two options. Architect for civil, commercial and industrial building design, or architect for other technical design.*
R: *The second.*

The interviewer does not wait for the computer and asks:

I: *What is your occupation or function at this moment?*
R: *I'm the director.*

Here the interviewer sees that the computer, without presenting the question first, suggests "architect" as occupation. This is very different from what the respondent just said, so she has to start a discussion to set things right.

I: *Er… let me see. You do not work as an architect yourself then?*
R: *Yes, I do.*
I: *What would you say is your main occupation, director or architect?*
R: *Architect.*
I: *Okay.*

It became clear that the interviewers still thought in terms of the old questions in which they had to probe and probe. Pronouncing the questions verbatim was secondary. Getting an answer (any answer) was the most important thing. In the following example a mother is asked to answer questions about her son, but there are some things she really does not know.

I: *What is his occupation or function at this moment?*
R: *Administrative worker.*

For most people this occupation would seem perfectly clear, but linked to this term are many different codes depending on main activity.

I: *Is it an executive function?*
R: *No.*
I: *And what are his main activities?*
R: *Well, I don't know. He has been working there only for a few days. I haven't yet heard much about it.*

The interviewer does not settle for "don't know", so she suggests a possible activity. In fact she directs the respondent in a certain direction.

I: Is he concerned with the bookkeeping?
R: Yes.
I: At least that then. Let me see, let's fill that in.

The interviewer now gets several options, among other things "junior accountant" and "certificated accountant".

I: Is he certified? (In Dutch this could also mean, "does he have a diploma?")
R: Yes.
I: Okay.

In this example a boy, who has just started working as an administrative worker, has been coded as a certificated accountant. The interviewer's urge to get a direct hit has caused a false code. This is a dilemma, because we want the interviewer to gather as much information as possible, but we do not want them to direct the answers. Schober/Conrad (2002) say: "Since interviewers always influence responses, this raises the question of which kinds of influence are benign and which are not. We argue that the criterion should be how interviewer behaviors affect response accuracy – that is, how well responses correspond with the definitions the survey author had in mind." This means that the interviewers should be trained well, and that they should know the aims of the researchers. If it is not possible to explain the precise aims of the researcher the interviewer should be instructed to accept not reaching a hit in some occasions.

5.5. Does the coding programme disturb the interview?

It is very unusual in a telephone interview that there are silences, this is logical since the only contact with the other person is by means of speech. Neither party knows what happens at the other end of the line. By talking or by making noises an interlocutor shows that he is still there and that he is still participating in the conversation. There are two different kinds of silences, the first occurs between the end of a question and the beginning of an answer and the second occurs between the end of an answer and the beginning of the next question. During the first the respondent is thinking which answer to give. During the latter the interviewer is most probably typing the answer, and the clicking of the keys can be heard. What usually happens during an interview is that a respondent makes thinking noises while searching for the right answer, for example sighing out loud, "er", or clicking with the tongue. The interviewer also tries to keep the silences as short as possible, by quickly repeating the question, or by adding some

information for example. While the clicking of the keys can be heard the silence is often not broken, since the reason for the silence is clear.

In the test the coding programme still wasn't fast enough to avoid awkward silences. Also there occurred another kind of silence, i.e. the silence while the interviewer is waiting for the computer. These silences triggered slightly different reactions from the interviewers. In many cases the interviewer started explaining what was happening. Remarks that are often heard are:

"The computer is searching."
"This is a new programme."
"I'm waiting for the computer."

In some cases the interviewers already beforehand said to the respondents that they are working with a new questionnaire and that things might go a little slow.

The respondents also react to the silences since they cannot hear what is happening on the other end of the line. They sometimes interpret the silence as a sign that there is something wrong with the given answer. So they start to help the interviewer:

I: *Can you say to which sector the company you work for belongs?*
R: *Er, food, right. Er, cocoa, cocoa butter…*
I: *Is it for manufacturing or trading or -…*
R: *Yes yes it's for de the products we make, for the chocolate-industry. You should look in the cake-sector.*
I: *[silence] That's not right in any case. I have to look where I'm going to put it. I will try Trade. Let me see what he does. Oh, no I can't find it there, that's not right. I will have to find it in Industry. I'll put it there. [silence]*
R: *The address is Xxx-street in Haarlem.*
I: *Well he won't know the address any way. No I will put it under "Other".*

With the questions about education the respondents sometimes interpret the silence as a sign that they can go on. It happens that the respondent already mentions the next training while the interviewer or the computer is still processing the first training.

The coding programme, or the computer as a whole, becomes rather important in the conversation. It not only guides the questions and the conversation, but it also becomes a conversation partner. In the dialogues you hear very often the interviewers' surprised reaction to actions of the computer. The interviewers do not just accept the actions of the computer, but they respond to it and they draw the computer into the conversation.

"I'm waiting for the computer."
"The computer says…"
"Oh no the computer says something else."
"The computer is searching."

"Let me see whether the computer can find it."

One explanation for this phenomenon could be that in other surveys the questionnaire follows a fixed route so the interviewer is never surprised by the computers actions. With the new questions and the coding programme, the interviewer not always knows for sure what the next step in the questionnaire will be. It is expected that the interviewers eventually will get more used to the new coding programme, but at the moment the programme intrudes into the conversation.

Drawing the computer's view into the conversation can be positive, because it can force the interviewer and the respondent to clarify themselves or to discuss their point further. It can also be negative, for example when the computer cannot find a direct hit and says "Too many hits". This can push the respondent in the defence, because what he perceives as the truth is not acknowledged by the computer. Or as a respondent said: "*But my company really exists!*"

6. In closing

In this test, conversation analysis turned out to be a rather fruitful tool to find several shortcomings in the new questions. The change in length of the interviews can be shown very easily, but conversation analysis additionally shows the cause of the lengthening. For instance, the phrasing of the questions and the lack of an escape route caused the long dialogues for the questions about education for an elderly respondent.

The coding programme has generated a list of coded answers. This shows that it functions well, that is, it is able to generate output. Yet with conversation analysis you see that in some cases the code that has been selected is not or doubtfully the right one, e.g. the administrative worker that became a certified accountant.

During the evaluation the interviewers said that they were unaware of drawing the computer into the conversation, nor were they aware that it disturbed the interview. However by analysing the dialogues the impact of the programme on the interviews became clear.

In the focus groups the interviewers were rather positive about the interviews. There were some that mentioned the length of some interviews, but overall they were rather positive about how everything went. Still with conversation analysis you see that not everything went well during the interviews.

Besides this, conversation analysis has some drawbacks as well. For instance, it is a rather time consuming method, especially when all interviews have to be transcribed completely.

After this test, the questions and the coding programme were modified according to our recommendations. In September 2003 a field test was conducted and finally in January 2004 the questions will be integrated in all surveys of Statistics Netherlands.

References

Houtkoop-Steenstra, H., 2000: Interaction and the standardized survey interview: the living questionnaire Cambridge University Press

Michiels, J., 2003: Handleiding PRAT voor CBS-ers (Concept). [PRAT Guide (concept)] Heerlen: Statistics Netherlands.

Psathas, G., 1995: Conversation Analysis: the study of talk-in-interaction Sage Publications Inc. Thousand Oaks

Schober, M.F./Conrad, F.G., 2002: A Collaborative View of Standardized Survey Interviews. In D.W. Maynard et al. (eds.), Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. (pp. 67-94) New York: John Wiley & Son, Inc.

Vis, R./Beukenhorst, D., 2003: Vragenlabtest PRAT april 2003. [Report Questionnaire lab-test PRAT April 2003.] Heerlen: Statistics Netherlands.

Visschers, R., 2002: Conversatie Analyse en het testen van survey-vragen. [Conversation Analysis and testing survey questions.] Heerlen: Statistics Netherlands.

Contact

Rachel Vis
Statistics Netherlands
Kloosterweg 1
NL-6412 CZ Heerlen
The Netherlands
email: RVCS@cbs.nl

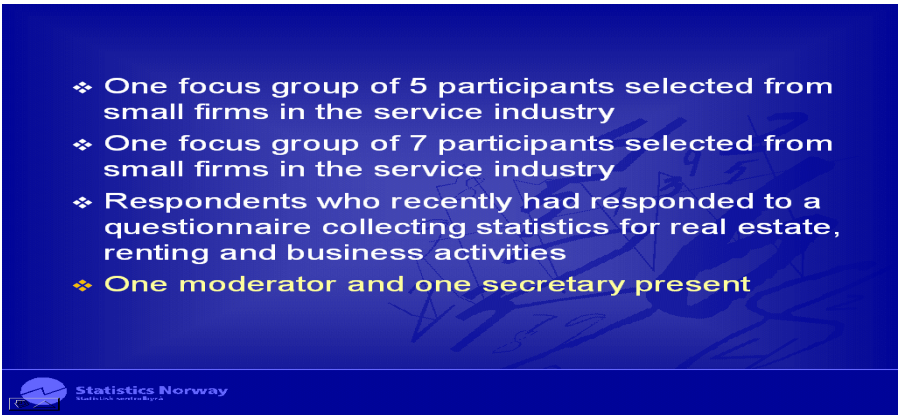
SEARCHING FOR RESPONSE BURDENS IN FOCUS GROUPS WITH BUSINESS RESPONDENTS

GUSTAV HARALDSEN

This presentation is based on experiences that have been made by Statistics Norway in the Eurostat-founded project *Developing Methods for Assessing Perceived Response Burden (PRB).* The project aims to develop methods for assessing perceived response burden in business surveys. It is led by Statistics Sweden. In addition the Office of National Statistics in UK is taking part. The strategy followed in the project is first to identify what the business survey respondents perceive as burdensome, and on this basis to develop relevant questions, which tap the perceived response burden. In this paper I will present the design we used in a couple of focus groups that we used to learn more about what was considered to be burdensome tasks in relation to business questionnaires. Our intention is of course to inspire others to run similar focus groups and share their results about perceived response burdens with us.

In the trials reported from in this paper, two focus groups were carried out with respondents from businesses in the service industry. The participants where selected among those who had recently responded to a questionnaire collecting statistics for real estate, renting and business activities. All the participants came from firms in the Oslo region. The first group gathered seven participants from rather big firms; most of them with more than 100 employees. The second group had five participants from smaller firms; some of them with only one or a few employees.

Figure 1: Focus Groups in first trial

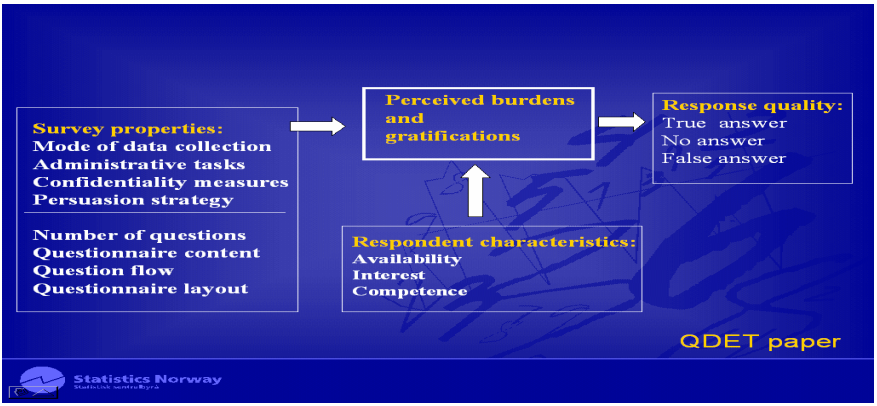


The focus group guide was based on a revised version of the conceptual model of perceived response burden presented in the paper *Identifying and Reducing the Response Burden in Internet Business Surveys* (Haraldsen 2002). In this version, eight aspects of the survey design which may affect the response burden are named:

1. Mode of survey communication
2. Administrative tasks
3. Confidentiality concerns
4. Persuasion strategy
5. Number of questions
6. Question content
7. Question flow
8. Questionnaire layout

The first four of these properties has to do with the data collection procedure, while the four last are different aspects of the questionnaire used. The perceived response burden is seen as a result of these survey properties in combination with the availability and initial motivation and competence of the business respondent.

Figure 2: Conceptual Response Burden Model



The model is described in more detail in Haraldsen (2002). What is important here is that we wanted to initiate a discussion that covered both the survey design aspects and respondent characteristics indicated in the model during our focus group discussions.

The focus group guide consisted of a mix of topics for discussion and practical exercises. Most of the exercises and visual tools that were used were printed in an eight page booklet given to each participant.

The Focus Group Guide

The focus group agenda was as follows:

1. **Introduction** covering a presentation of the topic of the discussion, of the moderator and the secretary and of how a focus group is set up and run.

2. The participants **introduce themselves** and the firm they are representing. The main purpose of this presentation was to reveal what the participants had in common, and in this way to create a sense of commonness. The important common denominator of the focus groups was that they belonged either to a group of small firms or a group of bigger firms.

3. An **open and general discussion** about the response burden of questionnaires received from Statistics Norway. This discussion was wrapped up by asking each participant to indicate on a scale running from 0 to 6 how easy or burdensome they found the questionnaires sent to them by Statistics Norway.

4. **Prepared discussion** based on a four-page questionnaire used to collect statistics for real estate, renting and business activities. All the participants had completed this questionnaire just a few weeks ago. First the envelope with a letter of introduction, the questionnaire and a separate description of difficult terms were given to the participant. They were asked to open the envelope and read the material the way they would have done if they had received it in their office. Next the moderator focused on five topics. The moderator was free to present these topics in the order he found most natural. The topics were….

4.1. **Readability**. The focus of this discussion was on the layout and the length of the questionnaire, and on the order of questions.

As a point of departure one of the participants was asked to read aloud and comment on question 2, which consisted of 12 sub questions about the international relationships of the firm. The question had a rather long introduction. It consisted of a mixture of questions with fixed response categories and open questions where the respondents should fill in an amount or a percentage. The sub questions were presented over two columns, while the other questions in the questionnaire were presented in a full A4 format. The reading of the question was followed up with a discussion of how easy or difficult it was to understand and find one's way through the different sub questions.

The participants were shown a list of the order of sub questions in question 2 and asked if they found this to be a natural order of questions about international relationships.

The participants were also showed the same kind of list of all the topics covered in the questionnaire and asked to comment on the order of the questions.

Finally the participants were asked…

- if they found this to be a short or long questionnaire
- if they based this impression on the number of pages, the number of questions or on other characteristics of the questionnaire
- how long time they would normally use to complete this questionnaire
- if they considered this to be a short or long time
- if the deadline for completion was considered to be long or short
- if they would prefer to receive the questionnaire at an earlier or later point of time

4.2 **Question problems.** The focus of this discussion was on the definition of question terms, the tasks that the respondents should perform and the response formats and level of detail asked for in the questions. For each of these three aspects we had picked a question from the questionnaire to illustrate the problem.

The discussion about problems with terms and definition of terms used a question about investments (in tools, means of transport and buildings) as a point of departure. In this question it was referred to budget estimates already reported to the tax authorities. The respondents were told to add together some of these, to add investments that were not covered in posts referred to or to exclude some investments that were embedded in the sums given in the previous form. On the separate help sheet the same terms were explained with the help of a more formal definition.

The respondents were asked what kind of explanation they preferred and if these two ways of explaining the terms would give the same result. They were also asked if they found it easy or hard to draw the borderline between those expenses that should be included and excluded.

Four of the questions were explained in more detail on the help sheet, while two remaining questions did not have a separate explanation. The respondents were asked if they found some of the explanation unnecessary or if they missed some explanations. They were also asked if they preferred to have difficult terms explained to them on a separate sheet or in the questionnaire itself.

As an example of **questions that may cause calculation problems** we used a question that asked for the average number of owners working in the firm, the average number of employees and the total man-labor carried out in the firm. The focus group participants were asked how difficult it was to answer these questions and how they estimated the figures asked for. They were also asked if it would be easier or more difficult to give monthly figures instead of estimating an annual average.

In the third part of the discussion a question that asked for total and activated expenses used on computer hardware and software was used to discuss **how easy or difficult it is to give detailed figures**. In this question the expenses should be given in 1000 kroner (= 125 €).

The discussion about question problems was wrapped up by asking the respondents to indicate in the exercise booklet whether it was difficult terms, difficult tasks or difficult response formats that caused most frustration for the respondents in business surveys. This evaluation exercise was simply presented like this:

Which aspects of the questions cause most and least troubles in statistical questionnaires? Write in the words "most" and "least".

Definition of terms _____

Calculation of answers _____

The level of detail in the answers _____

Similar cards were also used in later evaluations.

4.3 **Administrative tasks.** The participants were asked to write down which tasks that took place before, during and after the completion of the questionnaire. The results were presented around the table and discussed.

The participants were asked to indicate in the booklet which of the steps from preparation to posting the answers that was felt less and most burdensome. They were also asked, if the had the opportunity, which of these activities they rather would be spared for.

4.4 **Attitudes towards the task.** The focus in this part of the discussion was on the interest for and attitudes towards statistical information and confidentiality concerns.

In the introduction letter enclosed with the questionnaire it says that statistics about real estate, renting and business activities are used as planning and management tools, both by politicians and in the business world. Our first question to the participants was if the could suggest in what way this statistics was useful for policy makers and the industry.

A table and press release from statistics produced from the test questionnaire was shown to the participants. They were asked if they had ever sought information from any of the publications or web pages of Statistics Norway. They were also asked if they found the press release and the table presented to them interesting and useful. If not, it was discussed how the statistics could be presented to make it more interesting.

The establishments are obliged by law to answer the questionnaire used in this test. There is a reference to the relevant paragraphs in the introduction letter. The participants were asked if the knew what these regulations said about the duty to answer and about what may happen if one refused. They were also asked what they though would be the consequence if they were free to decide if they wanted to fill in the questionnaire or not.

4.5 **The burden of specific questionnaires vs the total burden of all questionnaires one has to fill in.**

- Questions posed in this part of the discussion were…
- How many statistical questionnaires do you complete during a year?
- How much time do you use on this kind of work?
- Are you the only person who fills in all questionnaires or is this job given to more than one person in the firm?
- Do you feel that these questionnaires represent a high or low workload?

After the prepared discussion, the participants where given a 10 minutes break.

5. Summary with the help of Concept Mapping

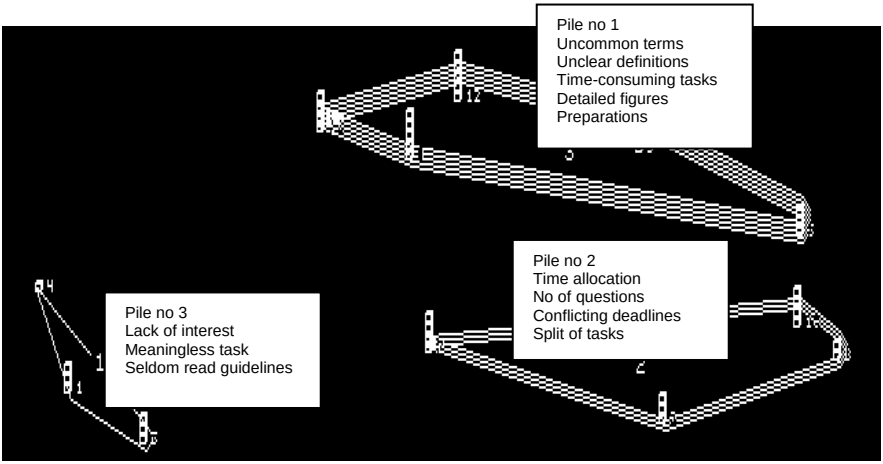
During the break, the moderator and secretary wrote down what they considered to be the main conclusions about what caused response burdens in business surveys. After the break the participants were given the opportunity to subtract or add new points to the list.

After a list was agreed upon, we used it as a basis for a concept analysis that was carried out in the following way:

- Each participant was asked to write down each statement on a small card
- Afterwards they were asked to indicate by a number from 0 to 6 how easy or burdensome they found the aspect described by the statements on the cards were-
- Finally they were asked to put cards that they felt described similar statements together and put a paper clip on each pile of cards.

The cards and the exercise booklet were left in a blank envelope as meeting broke up. The cards were later analyzed with the help of The Concept System, which is a program for concept mapping developed by William M. K. Trochim. The program offers a visual presentation of what statements the focus group participants have grouped together and the weight they have given the cards in each group.

As an illustration of the method, we have copied the concept map developed from the first focus group (with bigger firms) In this example the statements have been split in three piles. Pile no 1 contains those statements that were considered to burden the respondents most, while the statements listed in pile 2 and 3 were generally considered to be less important.



What did the Focus Groups tell us about Perceived Response Burden?

The main purpose of this paper is not to go into detail about substantial results from the focus groups, but rather to present the procedure used, and to discuss if this kind of focus groups can help us to operationalize the concept "Perceived Response Burden". The first step towards this aim, however, must be to get a clearer picture of what the business survey respondents perceive as burdensome. Only then do we know what concepts we should seek to measure. Therefore we will sum up some main conclusions from the focus group discussions.

Put in a slogan form, the results can be summarized in the following statement: ***The burdens of answering the questions seemed to be lower than the burdens that the respondents recognized in the questionnaire.***

Both groups recognized many problems in the questionnaires. But these problems did not cause as many problems for the respondents as we would expect. There are two reasons for this. The first one is rather encouraging, while the second one is more depressing.

1

The good news is that it seems to be a happy correlation between response competence and response burden. What we found in these two focus groups was that companies that should report a lot of complicated figures also had the most competent respondents; while firms with less competent respondents also had an easier task.

Respondents in bigger firms are more professional than respondents in small firms. In our focus groups the participants from the bigger firms generally were economist that had a

controller function in the firm. Their job was to perform different kinds of quality controls and they had established a documented practice for how the statistical questionnaires should be answered. Because of this they did not find the normal reports to Statistics Norway especially burdensome. In smaller firms the respondents had different educational backgrounds and held no formal controller function. On the other hand the smaller firms also had less information to report and had an easier task when the questions asked for calculations. Consequently even they found the respondent's job to be rather easy. For instance, when the questionnaire ask for the average number of employees last year, this needs to be calculated in bigger firms but normally has a straightforward answer in firms with just a few employees.

This observation is a very good illustration of the point that the perceived response burden is the result of the combination of survey design and respondent characteristics.

2

The not so good news is that when there initially is a response burden problem, the respondents seem to be rather clever to find short cuts that lift off some of the burdens. When we looked at the different aspects of questions, it was obvious that "unclear terms" and "terms that did not fit with the business records" were the most important problems in the questionnaires. One participant phrased it this way:

"Even if it is information gathering and the calculations that takes most time, it is the descriptions of what we shall report that leads to most frustration".

Even in the small sample of questions used in these focus groups and the small sample of business respondents gathered for the discussions, we revealed several examples of terms that were interpreted in different ways. And the most common complaint about the questions was that the terms did not fit with the records available. But we were also told that the standard solution to this kind of problems was to do a qualified guess based on existing records rather than bothering with complicated definitions and extra calculations. In other words, the respondents very often seem to have solved potential response burden problems with the help of simplified response strategies. The cost of this way of reducing the response burden is of course that the data quality may suffer.

3

In addition to these two observations, there is a third one that we think is important in the overall picture of how business respondents react to statistical questionnaires. This is a unison ignorance and skepticism towards the value of the statistics produced from the information the respondents provide. Only a few of the focus group participants had ever logged into Statistic Norway's homepage or ever read a statistical publication. The only

statistics they knew about was the price index, which some of them had used to revise their own prices. When they were showed a press release and a copy of one of the tables produced from the questionnaire in focus, none of them had seen this kind of tables earlier and none of them could see any immediate use of the results.

There seem to be two reasons for this lack of interest. The first is that the tables statistical agencies generally publish are not tailored to statistical needs in the companies. Several of the focus group participants told that they collected figures from competitors in order to compare their own resources, investments and results with those of other firms running the same kind of business. But none of them were aware of that Statistics Norway perhaps could produce the same kind of tables.

The other reason for the minimal interest in statistics seems to be a rather negative evaluation of one's own contributions. Because many of the respondents choose short cuts when they are faced with difficult question, and that short cut seem to be accepted, this has a negative effect on the credibility of statistical products. One participant told us

"The first time I responded to the questionnaire I spent a lot of time on it, but still expected that Statistics Norway would call me up because something was wrong. But I never heard from them. As a result I do not take the task so serious anymore".

In other words, the absence both of useful statistical products and of quality controls, seem to lead to a laissez faire attitude toward the tasks we ask the respondents to perform. For the project this means that it may be difficult to measure the perceived response burden for other respondents than those who fill in the questionnaire for the first time. This based on the assumption that, as a result of the initial exposure, the potential response burdens of the questionnaires are avoided at later encounters. Instead the quality- and credibility problems appear to be more serious than a heavy response burden.

How to Improve the Focus Group Guide and Procedure?

We have four ideas for improvements of the methods used to map the perceived response burdens in establishments:

1. More effort should be made in order to identify companies and respondents that have a high response burden. In coming focus groups we think it is important to look for questionnaires where both big and small firms need to answer all questions and look for small firms that have a lot of questionnaires to answer.
2. We also think that it is important to ask the participants more about their personal characteristics than what we did in the first focus group session. Also when we later operationalize a new response burden concept we believe that we should pay more attention to the characteristics of the respondent in addition to company characteristics.

3. The concept mapping that we performed in the summary section of the focus group wrapped up the focus group discussion very well. Therefore, more time should be set off for this exercise. The list of statements should more clearly state response burden problems than what was the case in these two test groups. In a focus group that run for two hours, the concept mapping can only be run half ways. Ideally, the respondents should be confronted with the results of the mapping and be asked to comment on the results. But there is not time enough to process and present the data so quickly that this can be done. An alternative procedure could be to send the results to the participants after they have been analysed, and ask them to respond with their comments. A positive side effect of such a procedure might be that, in contrast to earlier experiences as respondents, the participants this time receive some interesting feedback.
4. In addition to focus groups, individual interviews should be run. Statistics Sweden is presently following up this advice.

The two focus group test referred to in this paper address the perceived response burden, while the question how the response burden should be measured remains unresolved. In later focus groups we should also try to discuss and test different measurement methods. We think that the distinctions we have made between recognized problems and actual response burdens, is important. The response burden seems to be those burdens that the respondents both recognize and are willing to bear. A measurement that is not able to distinguish between these two aspects may not be very useful. Thus, it might be that the original term "perceived response burden" need to be split up into the perceptions of survey problems and the perceptions of respondent problems; and that both these terms need to be operationalized.

References

Haraldsen, Gustav, 2002: Identifying and reducing the Response Burden in Internet Business Surveys. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET). Charleston, South Carolina, November 14-17, 2002.

Contact

Gustav Haraldsen
Statistics Norway
P.P. box 8131 Dep
0033 Oslo
Norway
email: GHa@ssb.no

DEVELOPING TAILORED COGNITIVE PROTOCOLS: CAN COGNITIVE INTERVIEWS BE CONDUCTED OVER THE TELEPHONE?

CAROL COSENZA

What's the best way to do cognitive testing? This question, both explicitly and implicitly asked by researchers world-wide, may not have an easy answer. Although intuitively it makes sense that there may be not be only one single correct way to cognitively evaluate a survey, research is just beginning in this field. Those interested in question evaluation are experimenting with different methods and modes trying to determine what works best under what conditions. Tests are being conducted looking at interviewer effects, effects of type of probes used, and effects of the mode of administration. This paper will focus on an experiment looking at what differences, if any, are found comparing phone to in-person administration of a cognitive interview protocol.

Survey researchers know that when they conduct a study, there is no monolithic "correct" way to collect the best data. There are variables, such as mode of administration, length of interview, and types of questions asked, that can influence the cost, response rates, and even the quality of the data. Just as survey researchers juggle these different variables when designing a study, we are beginning to see that we have many of those same options when planning cognitive testing. *Who* does the interview influences not only how the interview is conducted, but may also be a factor in what we learn from the interview. *When* probes get asked in relation to the test questions can also vary. Cognitive information can be obtained retrospectively (either immediately after a question, following a series of questions, or at the end of the entire interview), concurrently (using the think-aloud model), or prospectively (when, before asking the test question, the interviewer elicits some background information about the respondent that is not based or tied directly to any one specific question). The *structure* of the protocol and cognitive probes can range from a completely unstructured interview, where interviewers have no

pre-scripted probes to a totally structured interview, with a pre-scripted protocol. There are also many ways for *researchers to learn or "get results"* from cognitive interviews, including interviewer debriefings, individual summaries written by the interviewer, reviewing audio or video tapes, and coding the interview.

At the Center for Survey Research (CSR), when deciding which cognitive testing method to use, we consider both content and operational issues. For example, when we have pre-identified concerns about item phrasing or vocabulary, we have found that we often learn more by asking more standardized probes that focus specifically on areas of interest. When doing surveys with a particular sub-group or population that we may not have had a lot of experience with, our interest may be more centered on topics and questions actually make sense to respondents in their particular situations. In these cases, we feel it's important to understand the respondent's situation and often start the cognitive interview by asking some prospective questions. Issues of usability are very important in self-administered surveys. Different cognitive methods can be used to find out about the cognitive process a respondent goes through when filling out a survey as well as how the individual questions are cognitively understood.

In addition to content specific issues, as in any research, there are operational issues that must be considered. How much money and time you have to complete the cognitive interviews, who is part of the sample (and where are they located), and what mode of administration you are using all have some impact on the study design.

Our experiences using different methods of cognitive interviewing have led us to tailor our cognitive protocols based on our needs. One recent interest has been the effect of doing cognitive interviews over the telephone. CSR has used telephone administration in several studies in the past - a decision most often driven by the population we are studying. For example, we've done telephone cognitive interviews with doctors and residents about end-of-life care and with CEOs of building supply companies about their inventory. Often these populations are very busy and don't have the time or inclination to come in to be interviewed. By conducting the interviews over the phone, we have greater access to more people. Doing cognitive interviews over the phone also breaks down physical barriers to participating. Everyone knows that the best test of a survey is done with the population that will be doing the survey, not necessarily a general population cognitive pool of respondents. However, this is often difficult to accomplish if the sampled population is not in your area (for example, studying workers with injuries in another state) or they are unable to come into a lab setting because of health issues (respondents who had recently been released from the hospital). We were able to accomplish both of these using telephone cognitive interviewing. We've also used the

telephone cognitive interview as part of a methodological test - comparing closed-ended/ structured interviewing to open ended/unscripted interviewing.

There are both advantages and disadvantages to doing cognitive interviewing over the phone. We've found that the main advantages are sample accessibility and possible monetary savings. Being able to talk to cognitive respondents over the phone allows a greater range in whom we can interview. We are not limited by geography. Respondents can live nearby or across the country. It allows us to get a better representative sample of potential respondents than if we were forced to use only those near our center. It also allows people with limited physical mobility the opportunity to participate. There are also potential savings in interviewer time, and travel costs.

There are also significant drawbacks that center on not being able to see the respondent. Visual cues from the respondent, such as nodding, looks of confusion, or attention loss, are all things that could lead an interviewer to do additional probing. These cues are unavailable to the interviewer when doing phone interviews. If testing usability of the instrument is a concern, phone interviews are not helpful since the interview cannot be video taped and the interviewer does not have the ability to watch the respondent fill out the forms. Also, any of the other visual components of an cognitive interview, such as using show cards or doing card sorts are impossible in this mode.

The Experiment

While we understood there would be trade-offs when doing cognitive testing over the phone, we didn't know how, or even if, these trade-offs would influence the cognitive information we learned. Our original plan was to conduct cognitive interviews for the same study, both in person and on the phone, and see if there were differences in what we learned. The plan was built in to a large study of job skills and training. Unfortunately, the study was delayed and was not ready to be tested before QUEST. A last-minute alternative plan, a mini-experiment, was developed in the hopes that it would provide leads about the protocol for future research

The mini-experiment was conducted as part of a study of doctors asking about their views on aggressive treatment and testing for certain health issues. Six cognitive interviews were done by 2 senior staff members. Three were done in-person by one of the interviewers and the other 3 were done by the other interviewer over the phone. We found no differences in the length of the interviews or in what we learned from the respondents (the same questions were found to be problematic by both those that did the interview over the phone and in-person). The one possible difference we saw was that doing it in-

person might have helped establish rapport with the respondent more quickly. In-person respondents may have been a little quicker to slip into their role as cognitive respondent and see themselves as part of the "team" evaluating the questionnaire.

Obviously there were major limitations with this mini-study. First, was the scale of the study. The sample size was extremely small and there were only 2 interviewers. There was also a lack of clear, objective criteria for identifying differences. We found it very hard to tease out interviewer differences from mode differences.

The Next Step: A Bigger Study

The original study on job training will be happening and will include cognitive interviewing on a larger scale. Four trained cognitive interviewers will do 3 interviews each - 2 doing them on phone and 2 in-person. All interviewers will all be briefed together so that everyone receives the same information about the study and the same goals for each question. The phone and in-person interviewers will be debriefed separately in the hopes that we will be able to focus on what was found (or not found) using each mode. We will try to distinguish whether the types of cognitive difficulties respondents had differ by mode and whether the number of problems found in the instrument differ.

At CSR, we use a model of cognitive interviewing that involves creating cognitive goals for each test question so that the interviewer understands what the researcher is trying to find out. The use of these cognitive goals may decrease the interviewer effects since both groups of interviewers will be using the same instrument.

Questions we hope to be answer with this research include:

1. What differences, if any, will be found between what we learn from the cognitive interviews done on phone interviews and those done in-person?
2. How will phone cognitive interviews work for a non-elite population?
3. Will using different modes mean we have to alter our cognitive interviewing protocol structure?

Next Steps: Defining Outcome Measures: It's different but is it better?

Outcome measures of validity are always hard to define for cognitive interviews (and for most other question evaluation techniques as well). What criteria should be used to figure

out whether one method is better than another? Possible criteria could be whether the number or types of cognitive problems differ between modes. For example, do some modes find more comprehension or retrieval problems than other modes. Still, the question remains about whether finding more problems necessarily equates with a better mode. We could also look at interviewer or respondent behavior. However, this may be more influenced by specific characteristics of the respondent rather than a mode effect We could try to create some kind of interviewer assessments, asking how the interviewers felt the interactions went and whether they thought the respondent had any cognitive difficulty at specific questions.

Future research in cognitive testing will need to focus not only on comparing different methods and modes but, perhaps more importantly, on how to interpret the differences that are found.

Contact

Carol Cosenza
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
U.S.A.
email: Carol.Cosenza@umb.edu

COMPUTER ASSISTED PRETESTING OF CATI QUESTIONNAIRES (CAPTIQ)

FRANK FAULBAUM

1. Introduction

Observational or standard pretesting of CATI-Questionnaires is not easily performed since in the strict sense this would mean that the recording of observed respondent behavior has to be done *during* the interview process. In this case, the coding system has to be designed in such a way that its handling does not influence the interviewer-respondent interaction. Otherwise the pretest would no longer constitute a pure field pretest but rather a pretest under specific conditions. In contrary to laboratory pretest methods like cognitive procedures (*think aloud*, *paraphrasing*, *probing*, etc.), pure observational pretesting exclusively relies on passive observation of respondents' behavior (for an overview of pretest methods see Exposito/Rothgeb 1997; Presser/Blair 1994; Prüfer/Rexroth 1996). Below, we present a method for Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ) which allows

- a *behavior coding* of the question-answer episodes in *real-time* under *field conditions (standard pretest)*, i.e. during the interview at that time where the episode really takes place without interrupting the natural flow of the interview;
- the *reliable identification of certain types of problems* occurring during the interview
- the assessment of *respondent and interviewer specific influences* on data quality on the basis of pretest data;
- the immediate transfer of codes to a data file while the interview process is going on;
- the using of *big random samples* in order to *reduce the sampling error* of pretest results and to do *more complex statistical analyses* already in the pretest stage of questionnaire development

(see Kleudgen/Faulbaum/Deutschmann 2001; Deutschmann/Faulbaum/Kleudgen 2003; Faulbaum/Deutschmann/Kleudgen 2003). The approach is considered to be a first attempt to integrate coding of response behavior into a normal CATI interview. Associated with the pretest procedure is a specific graphical presentation of pretest results which is called IPG (Interview Process Graph). The IPG like an electrocardiogram reveals the problem zones occurring in the complete interview. By this method of presentation, it is possible to identify problems with response scales as well as possible learning processes initialized by the respondents while going through item batteries. Problems of understanding and weaknesses in question wording manifest themselves in oscillations of the IPG.

Behavior coding of respondent behavior, which basically constitutes a variant of standard pretesting methodology, in its traditional form tries to classify response behavior along the dimension adequate vs. inadequate. The coding is done with respect to each question in the questionnaire. In principle, this could either be done by categorizing the responses *after the interview* or *during the interview*. The first variant has the disadvantage of requiring automatic recording of the whole interview which, in turn, at least in Germany, requires the consent of the respondents. Since this might disturb the pure field character of pretesting and might introduce a bias into response behavior, the decision was to use the second variant, i.e. coding the response behavior during the interview. While behavior coding of tape-recorded responses after the interview has the apparent advantage that it could be done *by the researcher* himself, coding during the interview requires that the coding is done *by trained interviewers*. This, however, is not easy to deal with because of the higher time pressure in case of telephone interviews. The interviewer has to do coding and interviewing at the same time without interrupting or delaying the interaction between interviewer and respondent which might constitute a heavy burden on the interviewer. This kind of multi-tasking demanded on the interviewer could be circumvented by letting the coding be done not by the interviewers but by specifically trained personnel equipped with separate computers and headsets who does the coding in parallel with the interview. This strategy, however, would also require the agreement of the respondents. Furthermore, for big sample sizes it requires a costly equipment.

Observation and categorization of response behavior during the interview process require a quite simple coding system which could easily be managed by the interviewers. Nonetheless, the simultaneous task of interviewing and coding puts some burden on the interviewers who have to be trained extensively. Only the most competent and experienced interviewers should be selected for the pretest phase.

2. Coding system and coding procedure

The coding principles used are derived from behavior coding systems described elsewhere (see Fowler/Cannell 1996; Morton-Williams 1979; Oksenberg/Cannell/Kalton 1991; Prüfer/ Rexroth 1985, 1996) and adapted to the properties of the telephone mode. In contrary to PAPI, to which most coding procedures originally refer computer assistance allows the integration of the coding system into the CATI software (and, in principle also the CAPI software) by reserving certain keys for particular types of respondent behavior.

The basic idea of coding respondent behavior can be illustrated by what Zouwen/Dijkstra/ Ongena (2000) called a "paradigmatic question-answer sequence". In a paradigmatic, ideal and unproblematic sequence, the interviewer poses each question correctly and the respondent gives an answer which the interviewer is able to assign to one of the response categories. This, in fact, means that the respondent only gives adequate responses. Thus, the central aim of behavior coding and its underlying coding system is to classify for each question occurring in the interview the adequacy or inadequacy of the respondents' answers and to identify certain types of inadequacy. Since no coding of the interviewer-behavior is done, i.e. no real interaction coding is involved, we cannot decide whether an inadequate behavior of the respondent has been caused by inadequate *interviewer* behavior. The latter possibility can only be ruled out by an extensive interviewer training. Moreover, if a sufficiently high number of respondents is pretested and many interviewers are involved, the problem is not so serious since systematic interviewer influences can be accounted for in the statistical analysis.

The coding system is described systematically in figure 1. The basic types of behavior categories upon which the coding system is based are:

- **Spontaneous answer to the question:** The respondent in his first reaction tries to give a direct answer to the question or refuses the question.
- **Non-spontaneous answer to the question:** The respondent in his first reaction wants a further clarification by the interviewer before she/he gives an answer, refuses or says „don't know". Thus, this class of responses collects all those which cannot be counted as direct attempts to select a response category.

To each of these classes there corresponds a number of behavior subcategories leading to a specific code. The codes are entered into the computer by the use of function keys in order to allow for a rapid input**.**

The behavior subcategories belonging to the above basic category types are:

Subcategories for "Spontaneous answer to the question":

- *Answer corresponds correctly to the response categories (response scale)* and can be assigned to the response categories including the categories „refuse" or „don't know" without any problem (Interviewer presses function key F1 in order to indicate that the answer was assignable without problems)
- *Answer does not exactly meet the response categories*, but the response can be assigned to the response categories without further probes by the interviewer (press function key F2)
- *Answer is assignable after further probes:* Respondent answers directly but must be asked, to which response category his answer should be assigned (press F3)
- *Anticipated answer:* Respondent answers already while the question is read by the interviewer (press F4)

Subcategories for "Non-spontaneous answers to the question":

- *Question understanding/acoustics/ language:* Respondent does not clearly understand the question because of acoustic reasons or he knows the language not well enough or the phone connection is bad and there is noise in the phone line (press F5).
- *Concept meaning:* The meaning of a concept is not understood, the respondent doesn't know the concept or the word (press F6)
- *Question comprehension:* Respondent doesn't understand the meaning (sense) of the question. He doesn't understand why the question was posed (press F7)
- *Response categories:* Respondent forgot the response categories, response scale too complicated (press F8**)**

Of course, various subcategories can be rearranged according to certain properties and collected in new specific response classes like "adequate" or "inadequate".

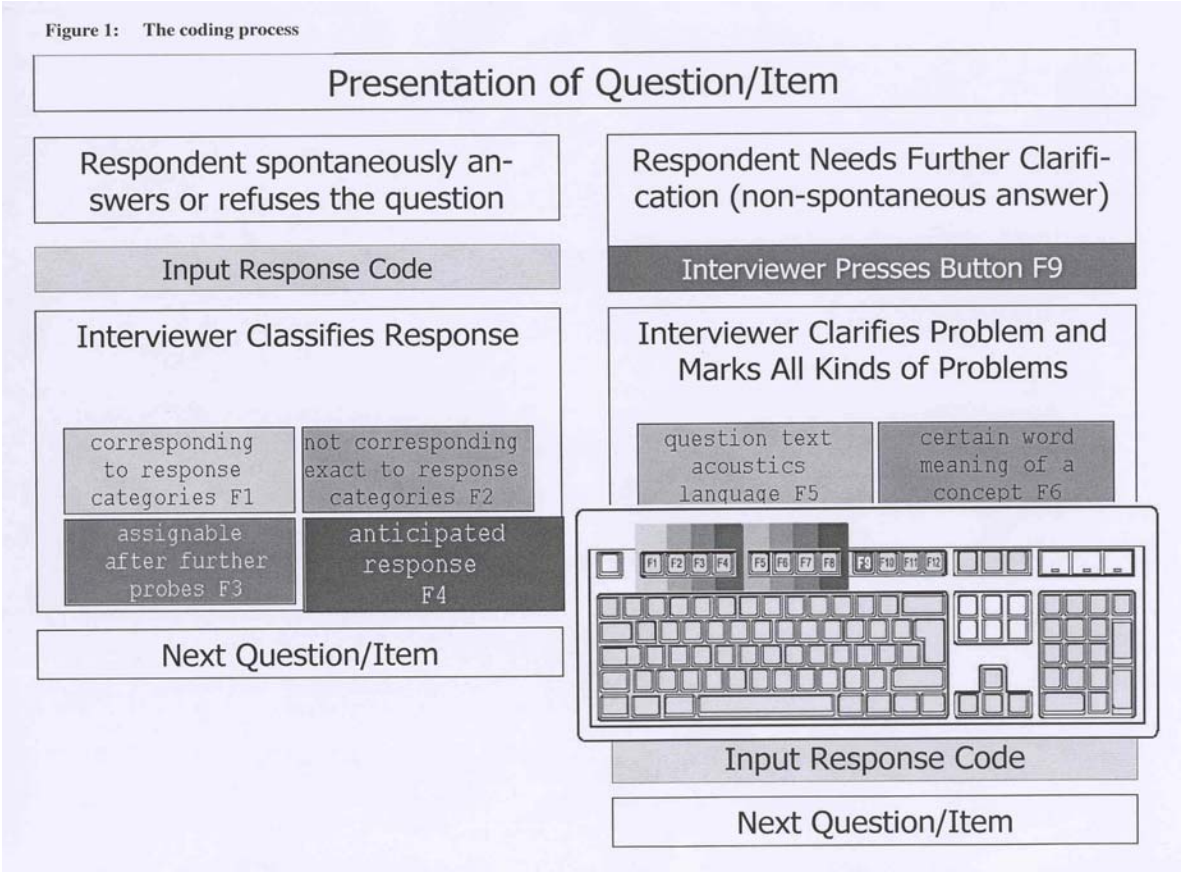
3. Analysis of pretest results

3.1 Structure of pretest data file and types of analyses

The pretest data file for each case contains the following information:

- Characteristics of respondent's interviewer (demographic variables, etc.
- For each question the response category including refusal information
- For each question and each coding category the classification code
- Further information about the interview (interview length, interviewer's impression about the respondent's behavior like cooperative attitude, etc.)

Figure 1: The coding process



These data admit different types of analyses: the analysis of interviewer differences in code statistics (frequencies, percentages) across questions and respondents, the analysis of differences in code parameters between types of respondents (male/female etc.) across interviewers and questions and the analysis of differences between questions or questions types in code parameters across interviewers and respondents. Examples of these types of analyses are given below. Of course, specific analyses for one interviewer, respondent or question can be done. A prerequisite for these analyses is a sufficient number of respondents and also questions. With a sufficient number of respondents also more complex statistical analyses like factor- and regression modeling or cluster analysis could be done.

3.2 Visualization of pretest results: The Interview Process Graph (IPG)

For each question of the questionnaire statistics of the different types of statistical coding results like frequencies, percentages, etc. of refusals and/or don't knows, of inadequate spontaneous responses, of comprehension problems, etc. can be plotted in various types of graphs we call *interview process graphs (IPGs)*. The horizontal axis of an IPG consists of the question numbers appearing in the same order as in the interview. The vertical axis refers to the statistics of certain types of coding. Thus, we can e.g. consider an IPG for the percentage of inadequate spontaneous responses, an IPG for total numbers of inadequate responses, an IPG for the percentages of meaning problems, etc.

IPGs allow for the identification of possible problem zones occurring during an interview and for the analysis of question/item problems in the context of neighbor questions/items which is especially important in case of big item batteries. They also permit the visualization of learning and adaptation processes occurring during the interview. One could e.g. visualize how fast the respondents learn to handle a certain type of response scale.

Figure 2 shows an example of an IPG. It is based on a CAPTIQ-pretest in a Health & Media Survey which dealt with media use and medical information seeking behavior. The sample size was 2.000. The questionnaire consisted of 124 questions of different types: simple yes/no questions about diseases and health problems, questions using various kinds of response scales for assessing the time dimension of health related behavior, item batteries for the identification of attitudes concerning different health topics using agreement scales as well as questions about knowledge of different diseases and the extent of media use in seeking medical information.

The size of the pretest sample was 100. The IPG in Figure 2 integrates different types of pretest information for all questions/items of the questionnaire: percentages of spontaneously given adequate and nearly adequate responses, percentages of spontaneously given inadequate responses and percentages of non-spontaneous answer

due to a problem. The codes defining theses response classes are indicated in the figure. The items indicated by a double star have been presented in a randomized fashion. We see that for some questions the percentages of adequate or nearly adequate responses were nearly 100 percent. An example are the thirteen questions named FR5_1 to FR5_13. The high percentages reflect the simplicity of the questions. The respondents were asked whether they already suffered from certain diseases. They had only to answer yes or no.

However, other items tell a completely different story. The item battery FR37_1 – FR37_10 introduced by the phrase „How do you feel personally informed about...“ followed by a list of different diseases like cancer/tumor, venereal diseases/Aids, heart condition, diabetes, etc. apparently seems to be more problematic. Respondents had to give a judgment on a verbal scale with respect to each disease. The scale values were (in English translation) „very well informed“, „well informed“, „somewhat informed“, „barely informed“, „not informed at all“ In 14% of all cases the interviewer could elicit an adequate answer only after further probes (spontaneous inadequate answer: Code F3).

A further example for weaknesses in an item battery is given by the six items named FR18_1 to FR18_6. The initial question was:

In the following I tell you some statements people sometimes make with respect to their health. Please tell me if you totally agree, almost agree, almost disagree or totally disagree.

Examples of items were:

- *My health is principally a matter of constitution and luck.*
- *My health is at first dependent of what I personally do.*
- *My health is determined by the physicians.*
- *Etc.*

On average, in 39% of the cases the respondents had to modify their spontaneous answers after probing by the interviewers in order to admit an assignment of the answer to an admissible response category. In addition, in 7% of the cases respondents apparently had problems and asked for clarification which may be seen as an indication of the larger complexity of task and a higher potential for response errors.

There is still another interesting finding which can well be illustrated by this item battery but which also occurs in other batteries. Items occurring earlier in the item list showed worse response behavior than items occurring later in the item list. This may either indicate the effect of further clarification in that respondents are becoming better in coping with the task in the sense of a learning process or that they return to constant response tendencies.

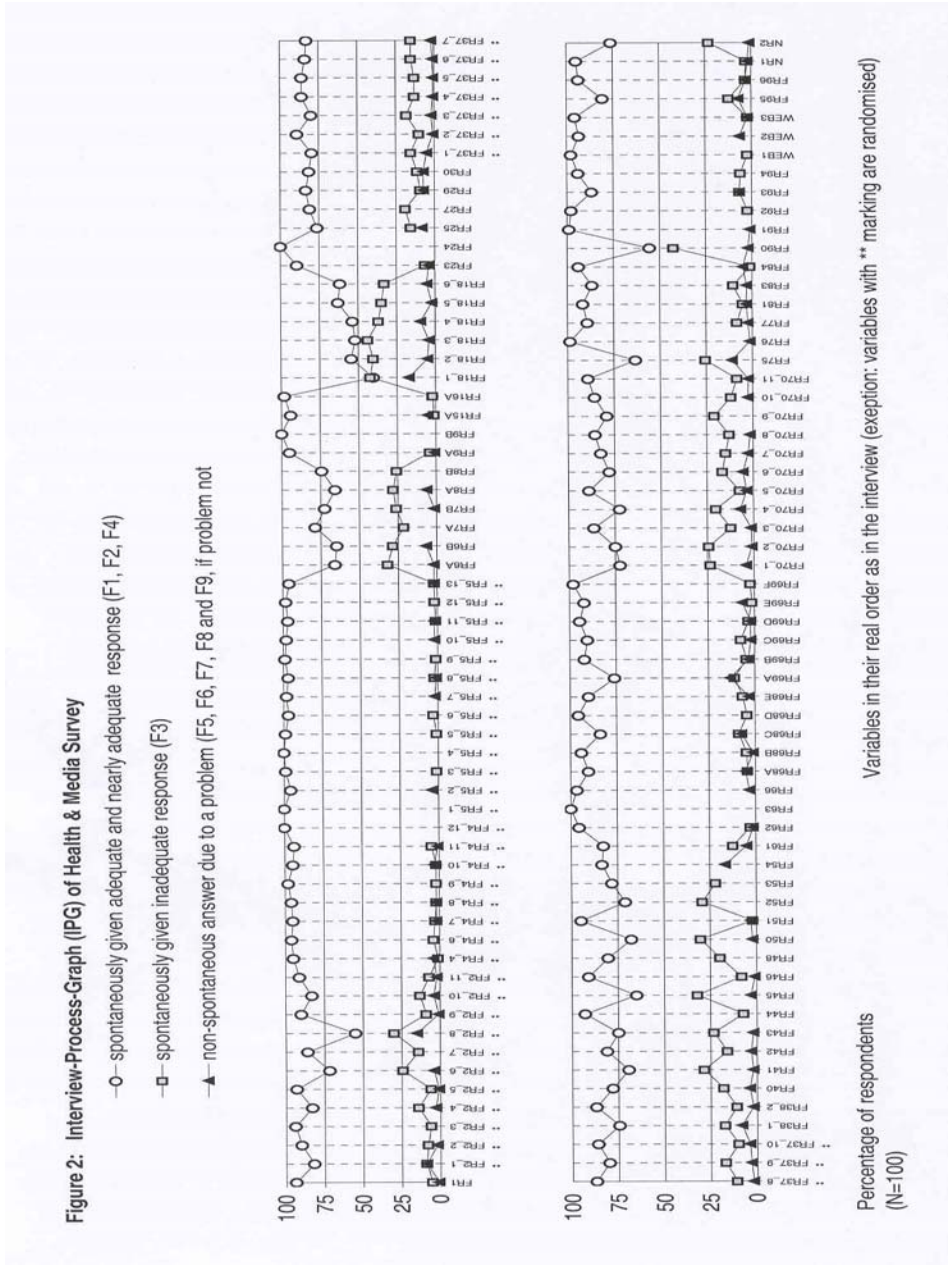


Figure 2: Interview-Process-Graph (IPG) of Health & Media Survey
–o– spontaneously given adequate and nearly adequate response (F1, F2, F4)
–□– spontaneously given inadequate response (F3)
–▲– non-spontaneous answer due to a problem (F5, F6, F7, F8 and F9, if problem not
Percentage of respondents (N=100)
Variables in their real order as in the interview (exeption: variables with ** marking are randomised)

The presentation of the first item FR18_1 causes problems for 17% of the respondents. The problems mainly concern the item or question understanding (7%) and problems with respect to the response categories (6%). In 4% of the cases the items only needed to be repeated by the interviewers. At the same time we observed an increase in the proportion of spontaneous adequate or nearly adequate answers (from 40.4% to 61,7%). The successive items were causing significantly less problems. The relevant percentages of the IPG are summarized once more in table 1.

Table 1: Proportions of adequate and inadequate answers

	spontaneous adequate or nearly adequate answer (F1, F2, F4)	spontaneous inadequate answer (F3)	non-spontaneous answer due to a problem (F5, F6, F7, F8)
FR18_1	40.4	42.4	17.2
FR18_2	54.5	40.4	5.1
FR18_3	52.1	43.8	4.2
FR18_4	53.7	36.8	9.5
FR18_5	63.0	34.8	2.2
FR18_6	61.7	33.0	5.3

n=100

3.3 Respondent- and interviewer-specific analyses

3.3.1 Respondent-specific analysis

The preceding section concentrated on item-specific analyses of pretest data, i.e. on the quality of the instrument. The advantage of the CAPTIQ method is that it can handle larger sample sizes which also admit respondent- and interviewer-specific analyses. Thus, questions like “Are there specific subgroups of respondents having more problems with respect to certain types of questions than other subgroups” or “Which properties of respondents have the biggest influence on response behavior?” can, in principle be investigated.

As an example, let us consider the relationship between the demographic respondent variables “Gender”, “Age” and “Education” and the response behavior. Table 2 gives an overview of the proportions of various types of adequate and inadequate answers. The proportions are based on a summation of codes over items and interviewers. The table

shows significant differences between males and females. Females apparently give more spontaneous inadequate answers and more non-spontaneous answers due to a problem than males. The proportion of spontaneous inadequate answers also increases with age and decreases with education.

Table 2: Respondent-specific analyses: Demographic variables and response adequacy

		spontaneous adequate or nearly adequate answer (F1, F2, F4)	spontaneous inadequate answer (F3)	non-spontaneous answer due to a problem (F5, F6, F7, F8)
Gender	male	86.8	9.9	3.3
	female	84.2	13.0	2.8
Age	16 - 29	89.8	7.3	2.9
	30 - 44	86.9	9.7	3.3
	45 - 59	83.3	14.1	2.5
	60 years and more	78.6	17.9	3.6
Education	low	80.8	15.9	3.3
	high	88.2	9.1	2.7
Total		**85.3**	**11.6**	**3.0**

n=100

Though these results are far from surprising they underline the plausibility of the method. Similar results have been obtained by Prüfer/Rexroth (1985) in their work on interaction coding and by Reuband (1998).

3.3.2 Interviewer-specific analysis

Under the condition of big pretest sample sizes already simple statistical description may reveal interviewer differences with respect to the classification of behavior types. In the pretest example from the Health and Media Survey the respondents were randomly selected for the pretest sample and randomly assigned to the interviewers so that differences in proportions are not considered to be confounded with other background

variables. Table 3 shows for each interviewer the proportions of respondents who gave spontaneous adequate or inadequate answers non-spontaneous answers due to a problem.

It can easily be recognized that there are important differences between the interviewers. While, e.g., interviewer „BM“ coded non-spontaneous answers due to a problem in 6% of the cases, interviewers „GA“ and „ZI“ assigned these codes only in 1,6%. of the cases. Interviewer „ZI“ had the highest proportion of the behavior category „spontaneous inadequate answer. The results indicate that the intense interviewer training did not lead to a full standardization of coding behavior.

Table 3: Example of interviewer-specific analysis: Comparison of interviewers

	Number of complete interviews	Spontaneous, adequate or nearly adequate answer (F1, F2, F4)*	Spontaneous, inadequate answer (F3)*	non-spontaneous answer due to a problem (F5, F6, F7, F8)*
Interviewer: AE	12	84.3	11.5	4.2
Interviewer: BM	18	81.7	12.3	6.0
Interviewer: GA	20	92.2	6.2	1.6
Interviewer: KA	11	86.2	11.1	2.7
Interviewer: SC	12	86.1	10.5	3.3
Interviewer: ZI	27	82.9	15.5	1.6

n=100
* percentages

4. Conclusions

The CAPTIQ-Method was specifically designed for evaluating CATI-Instruments with comparatively large pretest samples. The device is far from ideal. In fact, it has to rely on rather robust and rough coding principles. However, this does not mean that further refinements and modifications could not be done. In this respect the work presented here only represents a first step. What is needed in any case, are studies of intercoder reliability.

It is just the roughness of the method which guarantees its applicability to large pretest sample sizes which, in turn, allows for the application of more sophisticated statistical methods in the analysis of pretest data. Above, only the results of elementary inspections of the IPGs have been reported. More sophisticated analyses could involve factor

analyses and clustering of inadequate responses for the identification of problem types, methods of serial statistical analysis, subgroup analyses taking into account age, gender and other socioeconomic variables, etc.

The use of CAPTIQ is not limited to classical pretest applications which mainly concentrate on question quality. In addition, the method may also be used for the identification of interviewer-related as well as respondent-related causes of quality. Thus, response behavior is conceived to be decomposable into a respondent part, an interviewer part and a question wording part.

As a kind of observational pretest method CAPTIQ ideally should constitute the last member in a chain of pretesting stages all dealing with the improvement of the same instrument. It is clear that, at first, the standard rules for designing good questions should be followed (see Fowler 2001; Fowler/Mangione 1990) though in most research this is not the case. Also appraisal systems for questionnaires could be used (see e.g. Willis/Lessler 1999) at the first stage. The number of inadequate responses is expected to be substantially reduced if cognitive pretests are done before. In any case, the procedure serves diagnostic purposes. Though it is not able in every case to put into concrete terms what exactly has to be changed in the questions the procedure can give hints where to look for. It can also indicate problems not due to the question wording but rather to respondent- or interviewer-related properties.

CAPTIQ may also be useful if no extensive pretesting can be done. In most surveys which are not devoted to academic or governmental research but are done by commercial companies usually no extensive pretesting is taking place because of costs. Questionnaires are designed and then immediately submitted to the field. In these cases the method presented here could offer a quite cheap and routinely applicable method for the identification of severe questionnaire problems by inspecting the Interview Process Graph.

Contact

Prof. Dr. Frank Faulbaum
Lehrstuhl für Sozialwissenschaftliche Methoden/Empirische Sozialforschung
Sozialwissenschaftliches Umfragezentrum
Universität Duisburg-Essen, Standort Duisburg
Lotharstraße 65
D – 47048 Duisburg
email: faulbaum@uni-duisburg.de

References

Deutschmann, M./Faulbaum, F./Kleudgen, M., 2003: Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ). Proceedings of the American Statistical Association, Survey Research Section, New York: ASA.

Exposito, J. L./Rothgeb, J. M., 1997: Evaluating survey data: Making the transition from pretesting to quality assessment. In: Lyberg, L. et al (eds.) *Survey measurement and process quality.* New York: Wiley

Faulbaum, F./Deutschmann, M./Kleudgen, M., 2003: Computerunterstütztes Pretesting von CATI-Fragebögen. ZUMA-Nachrichten 52, S.20-34.

Fowler, F. J., 2001: Why it is easy to write bad questions. ZUMA-Nachrichten 48, S.49-66

Fowler, F. J./Mangione, Th. W., 1990: Standardized survey interviewing: Minimizing interviewer-related error. Newbury Park

Kleudgen, M./ Faulbaum, F./ Deutschmann, M., 2001: Computer assisted observational pretesting of CATI-questionnaires. Paper presented at the International Conference on Methodology and Statistics, Ljubljana.

Morton-Williams, J., 1979: The use of "Verbal Interaction Coding" Evaluating a questionaire. Quality and Quantity 13, 1979: S.59 – 75.

Oksenberg, L./Cannell, Ch./Kalton, G., 1991: New Strategies for pretesting survey questions. Journal of Official Statistics 7, S.349 – 365.

Porst, R., 1998: Im Vorfeld der Befragung: Planung, Fragebogenentwicklung, Pretesting. ZUMA-Arbeitsbericht, 98/02.

Presser, S./Blair, J., 1994: Survey pretesting: Do different methods produce different results? Sociological Methodology, S.73 – 104.

Prüfer, P./Rexroth, M., 1985: Zur Anwendung der Interaction-Coding-Technik. ZUMA-Nachrichten 17, S.2 – 49.

Prüfer, P./Rexroth, M., 1996: Verfahren zur Evaluation von Survey-Fragen: Ein Überblick. ZUMA-Nachrichten 39, S.95 – 115.

Reuband, K.H., 1998: Der Interviewer in der Interaktion mit dem Befragten – Reaktionen der Befragten und Anforderungen an den Interviewer. In: Statistisches Bundesamt (Hrsg.): Interviewereinsatz und –qualifikation. Band 11 der Schriftenreihe Spektrum Bundesstatistik, S.138-155.

Van der Zouwen, J./ Dijkstra, W./Ongena, Y., 2000: What Characteristics of Questions in Survey-Interviews make the Interaction between interviewer and respondent 'problematic' or even 'inadequate'? Department of Social Research Methodology, Vrije Universiteit, Amsterdam. Paper presented on the Fifth International Conference on Logic and Methodology, Köln, October 2000.

Willis, G. B./Lessler, J. T. (1999): Question Appraisal System-1999, Research Triangle Institute.

TANGIBLE EVIDENCE: USING MODERN TECHNOLOGY FOR RECORDING AND ANALYSIS OF INTERVIEWS

DIRKJAN BEUKENHORST[1]

1. Introduction

Two broadly defined methods exist to pre-test or evaluate a questionnaire. On the one hand we can try to detect problems by observing the interaction between interviewer and respondent during this 'conversation with a purpose'. On the other hand we can try to reconstruct the cognitive processes taking place inside the respondent in order to discover how respondents answer the questions. Both methods have their advantages and their drawbacks, and both uncover certain types of problems that the other can not detect. These methods can both be adapted to evaluating self-completion questionnaires.

Both benefit from collecting much detail in a systematic way because the cues pointing at problems often are rather subtle and can be easily overlooked. Problems not only risk being overlooked, sometimes they are 'found' without real evidence. This happens, for example, when respondents are overly stimulated to mention problematic aspects. In these cases it is as important to have detailed material to check this possibility.

Both methods, too, have a common problem: it is very difficult to strike a balance between the richness of detail giving insight in the existence and nature of a problem and on the other hand the ease and efficiency of analysis in qualitative or (semi-) quantitative ways. In this paper it will be argued that deployment of modern recording techniques in

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

combination with the use of dedicated software can contribute enormously to the quality and ease of analysis of pre-tests based on cognitive interviews and observational data.

Problems in a questionnaire can arise in the interaction process between the interviewer and the respondent. An interview is a conversation and by consequence should meet (sub-) cultural expectations concerning a conversation (Maynard a.o. 2002). It is, for example, rather unusual if someone asks "what does that mean?" to answer "whatever it means to you". This is, however, prescribed practice for interviewers. Just reading the question a second time if someone does not take his turn, is another rule interviewers follow that can annoy or confuse respondents. An interview thus violates many cultural expectations and presupposes furthermore knowledge about 'rules' peculiar to interviews (Houtkoop 2000, Stanley 1996). Not all respondents will have that knowledge: does everyone know what "tick all that apply" means or know how to state opinions on a five-point scale or how to express quantities in that way? (Miller, Integrating Socio-Cultural Factors into the Question Response Model, this volume). More trivial problems that can arise during an interview based on a questionnaire are difficulties an interviewer can have reading the question, or formulations that encourage premature answers (putting a question mark in the middle of the question followed by some illustration, for example). Such interaction problems are most of the time studied by some kind of observation of the interview process.

Many observational methods exist to test or evaluate these interact ional aspects of an interview. A simple and cheap method is having CATI-interviewers code the behaviour of respondents. Up till five different codes can be assigned in such a 'respondent behaviour coding' without distracting the interviewer too much from her main task (Burgess and Paton 1993). Codes are usually only given to a limited number of questions because of the extra burden for interviewers. Prüfer and Rexroth (1996) asked field-interviewers to indicate on the paper form if a question is problematic or not ('problem scoring'). Obvious limitation of both these methods is that the behaviour of the interviewer remains out of sight. More elaborate behaviour coding, including the behaviour of interviewers, is done by supervisors or researchers. This is applied mainly in CATI-settings where conditions to listen in are relatively easy to realise. Frequently used codes are the following: coded is whether interviewers read the question exactly as written, whether they make a slight change or whether they make a major one. For respondents is coded whether they ask for repetition of the question, whether they ask for clarification, whether they interrupt the interviewer, whether they give a qualified answer or whether they give an inadequate answer. Presser and Blair (1994) conclude this method of behaviour coding gives good results compared to cognitive interviews. The Australian Bureau of Statistics (2001) uses this method for uncovering interviewer errors and respondent fatigue.

Behaviour coding is often done 'real-time', that is during the interview itself, because of efficiency considerations or technical difficulties recording the interviews. This calls for simple codes, especially if the interviewer himself has to do the coding. Even with simple codes one cannot have very much confidence in the validity of the assigned codes (for an example of pre-tests with well-trained cognitive interviewers see Schechter a.o.1996). What is even worse, without a recording one lacks the possibility to check this validity.

Another reason for recording the interview, even if coding is done real-time, is that the necessarily simple codes do not give much information about the causes of problems. If we want to know more about the nature of a problem or want to remedy it, we will have to go back to the original interview sequences. In the second paragraph we will describe what level of detail we would like to record and what level is possible.

An interview is a very complex interaction process between an interviewer and a respondent. The interviewer reads the questions, the respondent answers, asks for some explanation, gives spontaneous comments etcetera. The interviewer reacts (adequately or not) by repeating the question or explaining a concept, the respondent takes his/her turn again and so on. During this response process the respondent has to accomplish different cognitive tasks which can lead to cognitive problems that can be the cause of invalid answers. The most common classification of different types of cognitive problems arising during the response process are problems with the comprehension of the question, those that arise during the retrieval of the information from memory, and problems concerning the formulation of the answer or choosing the right answer category. Because we cannot look directly inside the head of the respondent to study the ways in which he accomplishes the cognitive tasks, we have to rely on information the respondent can give about his cognitive activities. We do this by using techniques borrowed from cognitive psychology. The respondent often is asked to 'think aloud', that is to tell us what he is thinking while looking for an answer, or we ask him to describe the ways in which he searches his memory to find the required information, or we ask to repeat the question in his own words ('paraphrasing'). These tasks are for many respondents far more trying than answering the questions of the questionnaire themselves. Because these tasks are so difficult measurement errors in cognitive interviews may occur more frequently than in normal interviews.

Although cognitive interviews are called 'cognitive', this does not mean that we can rely solely on 'verbal information from the inside' elicited by cognitive techniques to diagnose questionnaire problems arising in the 'black box'. These verbal reports need to be supported by other evidence to establish the validity of problems. The complexity and subtlety of the content of a *pre-test* interview (cognitive, in-depth) is so great that based

on an audio recording conversation analysis as a technique to analyze all text, turn taking and para-linguistic behaviour of the (pre-test) interview is not sufficient because visual data are not recorded or analyzed. This omission is unfortunate because non-verbal behaviour, facial expressions and the like can tell a lot about comprehension problems, affective reluctance to answer and other response problems. For a thorough and rigorous analysis all relevant data have to be recorded in as much detail as possible. This means that cognitive techniques rely as well on observational data as the so-called observational studies proper.

Errors may arise in the reporting of cognitive process information by the respondent but also in the interpretation of this information by the researcher. The relative freedom of the interviewer during the test-interview to probe in case of suspected problems or to let pass unusual incidents as negligible, reinforces the subjectivity and complicates the analysis afterwards (Conrad and Blair 1996). This is another reason to record all data, in order to create the opportunity to invoke the judgement of other researchers. Another very important and not to be underestimated reason to record everything in detail is the fact that in a cognitive interview the normal course of the interview is often disrupted and the interview may therefore produce artificial results. If an interviewer is too persistent in probing for problems respondents may become inclined to mention more difficulties than they really encounter (Garas, Blair and Conrad 2003). In self-completion interviews interruptions by the researcher may distract the respondent so much that this in fact becomes the cause of problems. To diminish this risk of finding artificial problems careful analysis afterwards of interview sequences is necessary. This can only be done if a complete recording is at hand.

In conclusion, we need detailed observations not only for pre-tests using different types of behavioural coding but also for pre-tests using primarily cognitive or in-depth interviewing techniques. In the next section we will describe what kind of data we ideally would like to collect, depending on the data collection mode.

2. What evidence to record

In the case of computer assisted self-completion questionnaires (CASI) a pre-test can be executed by an interviewer or the pre-test can be done within the self-completion context: the respondent has to answer meta-questions or write comments on the questionnaire itself while answering the questions themselves (Beukenhorst, Giesen & de Vree 2002). In both cases we have to record everything that happens on the computer screen, the movements of the cursor, the entering of answers, the correction of answers etcetera. This can be done with a screen capture device that records all screens and what happens on

those screens. While recording it is possible, for example, to highlight the cursor or make mouse clicks audible. With the help of some hardware the recording can be done on another computer. This is convenient if test interviews are done on location where the respondent uses his own computer and does not want software installed on his pc. Another possibility is to make with a fixed camera a video of the computer screen. The video can be digitized and saved as a computer file. So-called audit trails make a file of all or a selection of keystrokes, including very precise timekeeping. This way one can reconstruct what the respondent did before and while entering answers. If an interviewer or researcher asks questions or helps the respondent with the questionnaire we should have a video and audio recording of all those events. If the respondent is asked to think aloud, an audio recording of all speech is of course indispensable. An extra camera focused on the face of the respondent can be very useful when interpreting the audio recording. If the test is especially aimed at the lay-out of the electronic questionnaire an eye-tracking device which records all eye-movements can be handy (Redline and Lankford 2001). A test of a self-completion paper questionnaire can be recorded more or less in the same way except that a camera has to be directed at the paper questionnaire and on the hands of the respondent in stead of making a screen capture and an audit trail.

Face-to-face interviews offer still more challenges for the technical equipment because the interaction between interviewer and respondent has to be recorded in detail, not only the spoken dialogue but also the facial expressions of both persons because they can point at reluctance to answer, feelings of uncertainty and other emotions which can hinder the flow of the interview. Sometimes it will be necessary to bring along a cameraman. Apart from a detailed video and audio recording an audit trail is helpful if it concerns a CAPI-interview.

In a CATI environment the only relevant data are the speech and some paralinguistic utterances, the interviewer and respondent can only hear each other after all. However, what the interviewer enters in the computerized questionnaire is important too and should be easily accessible during analysis. An audit trail of the interview should therefore also be made. A video of the faces of both respondent and interviewer can of course be useful too, although they do not influence the interaction. Such video's can only be made in a laboratory context.

A first, rather obvious, objection one could raise against such elaborate recording is the risk that the natural setting is so much changed that the interview becomes an unrealistic representation of a real life interview. This is mainly the case if one works on location and has to install all kinds of cameras and an extra pc. However, most respondents get quickly used to being observed and filmed and modern equipment is not so voluminous as in the

past. Pre-testing on location may in many cases even with cameraman and equipment be more realistic than a lab interview. Besides, as mentioned before, especially cognitive interviewing has in itself already a rather disturbing effect because the natural character and flow of the interview are much affected by the extra questions. So the cognitive interviewing research method is by itself not an unobtrusive method. Several mentioned recording techniques are indeed unobtrusive: the making of an audit trail and a screen capture remain unnoticed by respondent and interviewer alike. In practice one has to balance pragmatically on the one hand the (un)realistic situation of a test which can threaten the validity of the data and on the other hand the validity and rigor of the analysis which are highly dependent on detailed recording. Depending on personality characteristics of the test person and on the kind of interview and interviewing mode one can choose a not too disturbing technique that still records all relevant information for a valid analysis.

A second objection against such level of detail in the (recorded) data is that analysis of this rich material becomes far too complicated and time consuming for normal pre-testing practice. In the third section it will be explained that modern software makes such analysis and coding of data maybe not simple but at least feasible.

3. Coding and analysis

If we would describe in words all data gathered with the above mentioned techniques during ten interviews, we would end up with an at least one hundred pages long ‘thick description’ (Geertz 2000) that would constitute a really qualitative analysis. Doing that would take too much time and no client would read it. So we have to find ways to analyse the data in another way. We will have to reduce the amount of data without however loosing during this reduction process the possibility to easily access the complete dataset again. Sometimes data reduction is completed by writing out only the pieces of a test interview that at first hearing seem ‘relevant’. Often some other pieces can appear highly relevant during the ongoing analysis. This is common in a qualitative analysis method where description of the data and analysis intermingle. If one cannot easily retrace in the audio or video record the part of the interview that is found afterwards relevant, the analysis gets very tedious. So we need a method to mark all major parts of an interview in order to retrace them if one wants to study them in a later phase of the analysis.

Preparation for analysis and data reduction are executed by different forms of coding. Before we can code the data, however, we need to combine all data existing in separate

files into an easily accessible format. Software to do exactly that is offered on the market[2]. This software can be used for the coding and analysis of the behaviour as well. All audio and video recordings can be digitized and linked to audit trails. Synchronizing those different records facilitates finding tracks or parts of the interview. In this way one can analyze all audio, video and other records made of one test person. To compare results between different test persons marks have to be placed in the different records at relevant places in the questionnaires. Most of the time these marks will be placed at certain questions.

After having organized the material in this way, coding can start. This can be executed in two fundamental different ways. On the one hand, coding can be done like simple 'problem scoring'. If any kind of problem arises in one of the records one can check with the other data concerning that respondent if there really is a problem. If so, one enters a 'problem' in the behaviour coding analysis software. In the same way the records of other respondents are analyzed. After analyzing all records one makes a tabulation of how many problems with how many respondents arose at which points in the questionnaire. Besides this simple procedure one can try to differentiate between different types of problems, like comprehension or retrieval problems for the respondent, or reading problems for the interviewer. This more detailed analysis is facilitated because all relevant tracks can be easily accessed again. The use of rather simple codes is appropriate for the more practical testing of questionnaires where interest lies foremost in the detection and reparation of problems.

Another more 'scientific' method of qualitative coding and analysis starts from the other side, with the most detailed observation and description, and generalizes step by step from there. Behaviour is described first in basic categories: the subject who initiates the behaviour, what this subject does (a state or an event) and qualifiers that describe in more detail the behavioural act. An example may clarify this procedure.

> *At question x John (subject) looks (state) flabbergasted (qualifier) at the interviewer (qualifier)*
> *Interviewer Mary (subject) says 'uhuh' (event) looking neutral (qualifier) in another direction (qualifier)*

---

[2] Programs for organizing digital files and behaviour coding are the Observer from Noldus and Interact from Mangold. Programs designed only for behaviour coding are a.o. EthoLog, Jwatcher of the Animal Behavior Laboratory of Macquarie University and Focal32 developed by W.L. Roberts. Will Dijkstra from the Free University of Amsterdam developed Sequence for coding and analyzing interviews. It only runs on Macintosh.

This way one gets as many codes as a full description of the act would need. So some categorization or generalization is needed to reduce the amount of codes. These categories fit more than one incident while leaving out some detail.

Reduction:

At question x
John > Respondent
Looks flabbergasted > does not understand anything
Interviewer Mary > Interviewer
Says 'uhuh' + two qualifiers > reacts in a paralinguistic way without steering

Further reduction

At question X
R. gives no immediate answer because of problem with understanding
I. gives neutral prompt.

Still further:

question X in some way problematic.
Qualification: I. makes maybe a mistake because neutral prompt does not solve the lack of understanding.

Depending on the kind of software used the program can aggregate the detailed codes into more encompassing ones by recode commands or the researcher has to do this manually in consecutive rounds of coding. This type of coding which starts from a thick description and gradually evolves into a data set suited for quantitative analysis is greatly facilitated by coding software. This kind of analysis is most appropriate for more fundamental research into the response processes.

4. Conclusion

The two most common types of pre-testing a questionnaire are some form of behaviour coding and 'cognitive' research. Both depend much on scrupulous observation of the behaviour of the respondent and interviewer. This observation is greatly facilitated by modern technology to record all that behaviour. There exists software which can organize those observational data in such a way that coding and analysis become more efficient. The scientific rigour and validity of both the 'quick and dirty' method of finding and solving problems and the more fundamental method of studying the response processes are considerably increased.

References

Australian Bureau of Statistics, Pre-testing in Survey Development: An Australian Bureau of Statistics Perspective., 2001: Population Survey Development. Australian Bureau of Statistics. www.sch.abs.gov.au/SCH

Beukenhorst, D./Giesen, D./de Vree, M., 2002: Computerized versus interviewer-guided evaluation of CASI questionnaires. Paper presented at QUEST workshop 2002 Washington DC

Burgess, M.J./Paton, D., 1993: Coding of Respondent Behaviour By Interviewers To Test Questionnaire Wording. www.amstat.org/sections/srms/Proceedings/y1993

Conrad, F./Blair,J., 1996: From Impressions to Data: Increasing the Objectivity of Cognitive Interviews. www.amstat.org/sections/srms/Proceedings/y1996

Geertz, C., 2000: Local Knowlegde: Further essays in interpretive Anthropology. Basic Books (originally published 1973).

Houtkoop, H., 2000: Interaction and the Standardized Survey Interview, Cambridge: Cambridge U.P.

Maynard, D. W./Houtkoop-Steenstra,H./Schaeffer, N.C./van der Zouwen, J. (eds.), 2002: Standardization and Tacit Knowledge. New York: Wiley & Sons

Miller, K., 2004: Implications of socio-cultural factors in the question response process. ZUMA special 9

Nadra,G./Blair, J./Conrad,F., 2003: Inside the Black Box: Analysis of Interviewer-Respondent Interactions in Cognitive Interviews. Paper presented at Federal Committee on Statistical Methodology Research Conference, nov 17-19 2003, Arlington

Presser, S./Blair, J., 1994: Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology, vol 2 (no 12): 73-104

Prüfer, P./Rexroth, M., 1996: Verfahren zur Evaluation von Survey-Fragen: Ein Überblick. ZUMA-Arbeitsbericht Nr. 96/05

Redline, C. D./Lankford,C.P., 2001: Eye-Movement Analysis: A New Tool for Evaluating the Design of Visually Administered Instruments (Paper and Web). Paper presented at the American Association of Public Opinion Research, Montreal, Canada

Roth, E./Patterson, E., 2000: Using Observational Study As A Tool For Discovery: Uncovering Cognitive And Collaborative Demands And Adaptive Strategies. Paper presented at 5th Naturalistic Decision Making Conference. www.csel.Eng.Ohio-state.edu/emily/NDMrothfinal

Schechter, S./Blair, J./Hey, J.V., 1996: Conducting Cognitive Interviews to Test Self-Administered and Telephone Surveys: Which Methods Should We Use? www.amstat.org/sections/srms/Proceedings/y1996

Stanley, J. S., 1996: Standardizing Interviewer Behavior Based on the Results of Behavior Coding, www.amstat.org/sections/srms/Proceedings/y1996

Contact

Dirkjan Beukenhorst
Statistics Netherlands
CBS
Postbus 4481
NL – 6401 CZ Heerlen
Netherlands
email: dbkt@cbs.nl

TESTING WEB SURVEYS

HELENA BÄCKSTRÖM & BIRGIT HENNINGSSON

1. Introduction

Web surveys are still a new technique at Statistics Sweden, though surveys conducted on the web are becoming more frequently used. At Statistics Sweden people work with web surveys at different departments and in different ways. This causes a variety of layouts. We recommend forming some sort of guidelines about layout and other features considering web questionnaires.

2. Tests we have done

We have been testing some different web questionnaires, in different ways. We would like to tell you about three of these tests and what methods we have been using throughout the testing.

2.1 Statistical Databases of Sweden

Background

Statistics Denmark has during a couple of years carried out a web survey to users of the Statistical databases in Denmark. From the results a lot of relevant conclusions have been drawn and proposals to measures have appeared. Statistics Sweden got interested in doing a similar survey.

The Measurement Lab was involved at a very early stage. We discussed what questions to ask and constructed the questions. We also evaluated the web questionnaire with a suitable method and on an adequate number of “test people”.

Design

The thought is that the layout of the questionnaire should be in accordance with the web site of Statistics Sweden, and with its databases as well.

- The same colours are used here as on Statistics Sweden's web site and in the databases. A squared background behind the question is also a well-known characteristic.
- The head of the page is equipped with the logotype of Statistics Sweden, the reason simply to let the respondent know who is behind the survey.
- There is only one question at the time on the screen. This makes it very easy to get an overview of the question.
- All text is written in the font "Arial".
- There are five different colours in use. In the head of the page there are two shades of yellow. The name of the survey is written in red above the question. The question is in bold, instructions in italic and the response alternatives in normal font.
- At the bottom of each page there are two buttons, "Previous" and "Next". It is easier for the eye to notice and click on the biggest button. We changed the Swedish word for "Next" to "Next question" to let the buttons be of the same size.

2.2 Database with Information about our Surveys

Background

Statistics Sweden is going to build up a special database with information about all our surveys. A web questionnaire is sent out to all the product managers for them to fill out data about the surveys they are responsible for. One of the goals with the questionnaire is to be able to follow the development of electronic collection at Statistics Sweden. The product managers are meant to contribute to the development work by answering the web questionnaire.

Design

- The questionnaire is confusing. There is no order in the questionnaire.
- The questionnaire consists of 13 different sections, each of them with different numbers of questions. You can go between the different sections by choosing "section" in the head, in a "pop-up" menu.
- You have to scroll to see all the questions in one section.
- The same data will be filled out on several different places. They have not taken advantage of the possibilities on the web.
- There are neither skip instructions nor any questions whether the demanded data are available for the survey or not. All people have to read all the text to be able to make a decision whether it is data they should fill out or not.

2.3 The harvest and the autumn sowing

Background

The questionnaire that is to be sent out to farmers consists of two tables with questions about the harvest and the autumn sowing during 2003. Questions are asked about how big the area is for a lot of different crops, how big the harvest was, how many percent water the harvest contained and how many crops that can be sowed in the autumn.

Design

- The questionnaire is a little more than one page. You have to scroll.
- At the bottom of the page all the buttons are collected like “Save”, “Print the instructions”, “Send” and “Log out”.
- You find the instructions at each question where they are needed. The instructions are provided to the respondent in form of Pop-up buttons, which facilitate the task for the respondent. There is no need for them to search through the instructions.
- Calculations are made automatically.
- Some figures are already printed from previous year. To visualize that the respondent can not change these figures, they are written in a light grey colour.
- One problem is that you can not read the “Comments” if you use a screen with 800X600 pixels.

3. Testing Methods

Statistical Databases

The method we used was a combination of ”**think-aloud**” interviews and **observations**. From registers of users of the databases, a suitable sample was made. As a total, five interviews were made and they took about half an hour each. The tests and the interviews were tape-recorded.

At first the interviewer introduced the test and the interview, i.e. she explained the background and the purpose of the evaluation, and told the test person how the test would be carried out.

The test person filled out the questionnaire on the screen while the interviewer was sitting nearby observing the process. In the meantime the test person “thought aloud”, i.e. he/she told the interviewer what he/she was thinking of when reading the questions and how he/she discussed to get an answer that finally was filled out.

Once the test person had filled out the questionnaire a couple of more questions were asked, according to a topic guide. The questions were about the length of the questionnaire, the time spent, the wording and contents of the questions, layout and so on.

Database with Information about our Surveys

The Measurement Lab was not involved when this questionnaire was made. The difficulties showed up when the product managers filled out the questionnaire. All the respondents are working at Statistics Sweden. Afterwards it was decided that we should find out about the problems.

We did nine **in-depth interviews** with persons from different departments. We went through the questionnaire page by page. During the interview the respondents told us about the problems and thoughts they had when they filled out the questionnaire and also during the test. We wanted to find out about problems with definitions, vagueness and misunderstandings as well as problems caused by the design of the web questionnaire.

The harvest and autumn sowing

We did not do any test. We only tried to fill out the answers on the screen. The client had made the questionnaire. We **worked together with the programmer** and discussed different solutions to find out the best result. The work was very interesting and we learned a lot from one another.

4. Our Experiences

Most of our own experiences come from questionnaires made for enterprises and municipalities. The international experiences we have seen are mostly from questionnaires to individuals. We have just started to send out web surveys to individuals. Still very few will answer on the web if they can choose a paper version.

As you have heard we have used **think-aloud** tests together with **observations**. We have also made **in-depth interviews** to get opinions and more understanding of how the respondents think while they are filling out the questionnaire. The test methods we have used so far seem to have given us a lot of information.

We would like to see more cooperation between the producer and the programmer. Conducting surveys on the web is still a new technique and there are many different ways to handle it. Discuss with one another and exchange experiences. Use the possibilities that the web offers you!

There are a lot of good things with the web. For the producer - the data entry is ready. But also for the respondent; sums and calculations can be done automatically. Another example is the possibility to click on a word and you will see its definition. You can also get help for editing but do not overuse this option, since it might irritate the respondents and make them not bother after a while.

Some other things to remember:

- You need a clear structure since the overview is difficult on the web.
- It must be easy to save and to print. If the respondent leaves the questionnaire – will he/she be able to automatically return to the question where he/she left off?
- Put the buttons “Save” and “Print” together – it makes them easier to notice.
- Not too much to read. You need as simple questions as possible.
- You have to be careful with colours. Some colours are already associated with certain characteristics.
- Scroll as seldom as possible. Scrolling down too far stops you from seeing the headlines.
- The respondents expect it to go very quickly to fill out the questionnaire.
- Do not overrate the competence of the respondent. The test will show you.
- Which is the best font? Ventura?

5. Checklist

We have an increasing amount of jobs about web surveys. The clients also want us to tell them about our experiences. And they want good advice. Therefore we started to make a checklist draft which serves as something to look at and to discuss with the client when we start a new job. See the Appendix.

Appendix

Checklist: Designing Electronic Questionnaires

From paper to electronic questionnaires

When changing over from paper questionnaire to an electronic one, not only is the questionnaire “translated” to the new technique but its possibilities are also used. This creates new and more options at the same time as it makes greater demands. Electronic questionnaires can be more or less intelligent. The following help might be built in:

- conditions for editing
- possibilities to correct errors found and to register other changes
- routines for summing up totals and to make other calculations
- automatically performed skip instructions
- information texts shown on the screen when clicking on the given question

To think about beforehand

1. Consider secrecy, passwords, coding. Report to information security officer.
2. The design has to take into consideration both how computers work and how respondents expect questionnaires to be filled out.
3. Take into account that the use of mixed modes probably will be necessary, e.g. complementing the web questionnaire with a paper questionnaire.
4. It is the respondent who should review his own data. The checks should focus on commonly occurring errors and be designed to facilitate the respondent's understanding of the question and give knowledge of what the answer stands for. All checks that are implemented in the questionnaire should be carefully tested by respondents.
5. Work method: the programmers and the producers have to collaborate. Write a specification of demands.
6. Follow SCB:s general directions for publications on the Web.
7. Write briefly! It is not preferred to read long texts on the computer screen.
8. Prepare to test the questionnaire in the environment of the respondents.

Technical environment

1. What similarities to ordinary software are desirable?
   - how to save
   - how to move on to the next page
   - how to deliver, send
   - how to print
2. It should be easy to **SAVE**.
   It should be possible for the respondent to fill some parts out and then resume where they left off, i.e. without having to start from the beginning of the survey again. Respondents should be informed about this at the start of the questionnaire.
3. Make it easy to PRINT
   a. An empty form– Respondents often need information before they answer
   b. A filled out form
4. Leave instructions at each question where it is needed by providing pop-up help buttons. Make sure the instructions are sufficient especially where input format is crucial. E.g.: “enter your date of birth using: MM/DD/YYYY format”.
5. Let the respondent scroll as seldom as possible
6. Use graphic symbols to tell respondents how much they have left of the questionnaire.

Questionnaire design

- The design has to be respondent friendly.
- Introduce with a **welcome page**, which motivates the respondent to start answering the questions.
- Start with a question fully visible on the first screen page. Make the question easy to understand and easy to answer.
- Introduce each question in a conventional way, to make it look like a question in a paper questionnaire.
- Take advantages of the **possibilities on the web**. Sometimes it can be better to adjust the paper questionnaire according to the web.
- Combine words, graphic language and figures to **support and guide** the respondent through the wanted answering order.

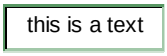
  - verbally - words
  - graphically - shapes, colours, figures etc.
  - numerically - numbers
- The respondent should be permitted to make the following actions when filling out the questionnaire:

  - Proceed to the next question,
  - Return to the previous question and
  - Quit.
- The respondent should always be able to proceed to the next question without being forced to answer every question on the way.
- Use a **logical question order** and take advantage of the opportunity of creating automatic skips.
- Try to make enough space for all answer alternatives on one screen page. If not – make double rows with clear instructions about how to navigate.
- **Radio buttons**  are appropriate for a relatively short list, with mutually exclusive items. When one item in a list is selected, all the others are unselected.
- **Check boxes**  are suitable for multiple responses, “check-all-that-apply” items.

Drop boxes

Pros: Drop boxes (drop-down menus) save space on screen; they are therefore suitable for long option lists, which allow type-ahead lookup (e.g. state, country).
Cons: Drop boxes imply more work for the respondent (at least three mouse clicks). The answer alternatives are often not visible until clicked on.

Text Field and Text Area



Be careful with open questions. Size and shape of the text field/text area should be adjusted for the length of the required answer. Provide sufficient space and always let the respondents know how much text they can fit in. Avoid horizontal scrolling! Use the text area for open-ended responses, general comments etc.

Size and colour

Use “Sans Serif” (e.g. Verdana, Arial, MS Sans serif) in a large font. Stick to one and the same colour. Never use more than four different colours on one screen. Be consistent!
Some colours are already globally associated with a certain characteristic:

Underlined text	clickable link
Blue	clickable link
Red	emergency messages or critical icons (red is associated with errors or warnings)

Consistency

Place the information texts at the same position on every screen and use the same format for the same information.

References

http://www.websm.org/topics.html
the best about electronic data collections and web surveys.

http://survey.sesrc.wsu.edu/dillman/

Couper, M. P., 2002: Web Survey Design: Survey Research Center, University of Michigan and Survey Joint Program in Survey Methology. Course in Stockholm 2002

Dillman, D., 2000: Mail and Internet Surveys: The Tailored Design Method. 2nd Edition John Wiley & Sons, Inc., New York: NY, Washington State University.

SCB 2001: Ask the right questions: How to develop, test, evaluate and improve questionnaires

Contact

Helena Bäckström
Statistics Sweden
Box 24 300
SE – 104 51 Stockholm
Sweden
email: helena.backstrom@scb.se

Birgit Henningsson
Statistics Sweden
Research and Development Methodology
SE – 701 89 Örebro
Sweden
email: birgit.henningsson@scb.se

WITH REGARD TO THE DESIGN OF MAJOR STATISTICAL SURVEYS: ARE WE WAITING TOO LONG TO EVALUATE SUBSTANTIVE QUESTIONNAIRE CONTENT?[1]

JAMES L. ESPOSITO

1. Introduction

QUEST – the acronym that serves as the signature for our workshop series refers to **QU**estionnaire **E**valuation **St**andards. During each of the preceding three QUEST workshops (1997 through 2001), various attendees have written papers that address (directly or indirectly) the topic of standards for designing and evaluating survey questionnaires. One aspect of this discussion that has not been satisfactorily addressed is the point at which evaluation work actually begins. For example, does evaluation work begin formally when behavioral scientists commence cognitive testing on components of a draft questionnaire, or does/should evaluation work begin earlier with the observation and conceptualization "products" of subject-matter specialists? In the present paper, I present a framework that relates questionnaire design-and-evaluation processes to sources of measurement error and take the position that evaluation work should commence much earlier in the design (and redesign) process than has been the case historically. Though there are many excellent published works on questionnaire design and/or evaluation in the literature [Akkerboom and Dehue, 1997; Converse and Presser, 1986; DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Foddy, 1993; Forsyth and Lessler, 1991; Fowler, 1995; Goldenberg, Anderson, Willimack, Freedman, Rutchik, and Moy, 2002 (for

1 The views expressed in this paper are those of the author and do not reflect the policies of the Bureau of Labor Statistics.

an establishment survey perspective); Oksenberg, Cannell and Kalton, 1991; Platek, 1985; Schwartz and Sudman, 1996; Sudman and Bradburn, 1982; Turner and Martin, 1984; and Willis, Royston and Bercini, 1991], this paper draws primarily on ideas presented by QUEST authors over the course of the past three workshops.

2. Background: Questionnaire Design-and-Evaluation Models

As noted, various members of the QUEST group have presented papers that address the topic of standards for designing and evaluating survey questionnaires, usually by presenting a descriptive model of the design-and-evaluation process. Several of those models are summarized below.

Model One. At the first QUEST workshop, Lindstrom and Akkerboom (1997) presented a "Questionnaire Testing Model" that comprised four phases:

"Phase 1, *the definition/feasibility study*, is used to construct a prototype questionnaire and data collection design usually involving a go/no go decision for further development.

Phase 2, *the qualitative content test*, is a small-scale test used mainly to produce a less error-prone questionnaire draft.

Phase 3, *the quantitative content test*, is a small-scale test used mainly to produce a less error-prone administration of the draft questionnaire, focusing on operational conditions.

Phase 4, the *quantitative pilot study*, is a final medium-scale test of the whole design (1997, pp. 10-11)"

The fourth phase is followed by the implementation phase and then the survey proper. In related papers, various authors identify and describe methods appropriate at each phase of the design-and-evaluation process. For example, during the qualitative content test, Akkerboom and Dehue (1997, Table 1a, p. 129) suggest that researchers employ the following pretesting tools: "focus groups, observation (ordinary interviews), in-depth interviews, 1-to-1 meta-interviews (cognitive interviews if using cognitive stimuli), [and] expert reappraisal." Later, during the quantitative pilot test, they suggest the use of "analysis outcomes, experiments, evaluation questions, [and] other monitoring tools (e.g., re-interviews, focus groups/debriefings)."

Model Two. At that same initial QUEST workshop, Esposito (1997) presented a five-phase model of the questionnaire design-and-evaluation process that was intended to encompass both initial design work and redesign work (see Esposito and Rothgeb, 1997).

The model comprised five phases: (1) conceptualization; (2) operationalization; (3) *pretesting* (evaluation work conducted *prior to* survey implementation); (4) survey administration; (5) *quality assessment* (evaluation work conducted *after* survey implementation). Esposito and Rothgeb (1997, pp. 543-551) also describe various evaluation methods appropriate for both pretesting and quality assessment research.

Model Three. At the 1999 QUEST workshop, Lindström presented the details of a questionnaire development-and-evaluation model that comprised seven phases:

Phase 1. Defining the contents of the survey.
Phase 2. Developing a questionnaire "at the desk".
Phase 3. Testing the questionnaire.
Phase 4. Adapting to production.
Phase 5. Monitoring the initial production.
Phase 6. Evaluation of quality and identification of sources of error.
Phase 7. Declaration of Quality

In this paper, Lindström (1999) emphasizes communication among client, methodologist and producer, describes the specific tasks associated with each phase, and identifies various methods appropriate to each phase.

Remarks and observations. Though the three models identified above differ in various respects, they (and other models of the design-and-evaluation process) seem to have accepted a dubious assumption: That *formal evaluation work* need not commence until *after* a preliminary set of questionnaire items have been developed. Formal evaluations of early developmental activity (i.e., observation and conceptualization) – such as, (1) incorporating independent/ethnographic observations of the subject-matter domain to determine how representative a sponsor's observations of that domain might be; or (2) undertaking an examination of the process by which and by whom key survey concepts have been developed – are rarely incorporated as well-developed components in questionnaire design-and-evaluation frameworks. In fact, as the following quotations suggest, there does not appear to be a great deal of enthusiasm for presurvey evaluation work of any kind.

"Selling focus groups, cognitive interviews and other pretests to survey clients is difficult. When clients make contacts, they either seem to believe that they already know how questions should be asked, or that you should be able to suggest solutions to question problems straight away. … In either case, when clients call the survey organization, they expect them to start the data collection almost instantly (Haraldsen, 1999, p. 67)."

"The competence and ambitions of our clients varies a lot. Some clients with good knowledge of measurement difficulties and with ambitions of high reliability

questionnaires will discuss in great detail their goals and their means to achieve them. They will also find resources to pay for the necessary studies and are often skilled in evaluating and using them. [New paragraph.] Often the clients do not have the time, capacity or interest to perform a questionnaire test especially adapted to their survey. When they contact the Measurement Laboratory (ML) there is a very short period available before their survey production is planned to start. Frequently these clients have designed preliminary questionnaires on their own (Henningsson, 2001, p. 73)."

"… [We] spent much time and effort this summer negotiating with questionnaire sponsors regarding questions that, as we saw in the lab, did not necessarily coincide with the actuality of respondents' lives. That is, the questions were written with an intended research agenda in mind and neglected to account for the ways in which respondents, themselves, experienced and made sense of the phenomena (Miller, 2001, p. 92)."

To better understand the importance and timing of evaluation work and its role in minimizing measurement error, we need an expanded model of the questionnaire design-and-evaluation process.

3. A Framework Relating Questionnaire Design-and-Evaluation Processes to Sources of Measurement Error

All surveys are not created equal: The framework outlined below and described very briefly is intended more for consideration in the design/redesign of interviewer-administered surveys that have recognized and ongoing societal importance (e.g., national surveys of health, employment, economic activity, social conditions, income, et cetera). A more detailed description of the framework can be found in a two recent conference papers (Esposito, 2002 and 2003).

The framework comprises two explicit dimensions, plus the implicit dimension of time/change (see Table 1):

- (1) Eight design-and-evaluation phases (for both initial design *and* redesign efforts);
- (2) Five sources of measurement error; and
- (3) The dimension of time – coupled with the inevitability of social, cultural, and technological change.

Table 1: A Framework Relating Questionnaire Design-and-Evaluation (D-and-E) Processes to Sources of Measurement Error

			INTERDEPENDENT SOURCES OF MEASUREMENT ERROR (at P7 or RP7)				
			Questionnaire D-and-E Team		Information/Data Collection Context		
		INITIAL DESIGN	*Content Specialist (1)*	*Design Specialist (2)*	*Interviewer (3)*	*Respondent (4)*	*Mode (5)*
	P1	*Observation*	C_{11}	C_{12}	C_{13}	C_{14}	
	P2	Evaluation	C_{21}	C_{22}	C_{23}	C_{24}	
Questionnaire	P3	*Conceptualization*	C_{31}	C_{32}	C_{33}	C_{34}	
Design and	P4	Evaluation	C_{41}	C_{42}	C_{43}	C_{44}	
Evaluation	P5	*Operationalization*	C_{51}	C_{52}	C_{53}	C_{54}	C_{55}
Phases	P6	Evaluation	C_{61}	C_{62}	C_{63}	C_{64}	C_{65}
	P7	***Administration***	$\mathbf{C_{71}}$	$\mathbf{C_{72}}$	$\mathbf{C_{73}}$	$\mathbf{C_{74}}$	$\mathbf{C_{75}}$
	P8	Evaluation	C_{81}	C_{82}	C_{83}	C_{84}	C_{85}
		REDESIGN					
	RP1	*Observation*	C_{R11}	C_{R12}	C_{R13}	C_{R14}	
	RP2	Evaluation	C_{R21}	C_{R22}	C_{R23}	C_{R24}	
Questionnaire	RP3	*Conceptualization*	C_{R31}	C_{R32}	C_{R33}	C_{R34}	
***Redesign* and**	RP4	Evaluation	C_{R41}	C_{R42}	C_{R43}	C_{R44}	
Evaluation	RP5	*Operationalization*	C_{R51}	C_{R52}	C_{R53}	C_{R54}	C_{R55}
Phases	RP6	Evaluation	C_{R61}	C_{R62}	C_{R63}	C_{R64}	C_{R65}
	RP7	***Administration***	$\mathbf{C_{R71}}$	$\mathbf{C_{R72}}$	$\mathbf{C_{R73}}$	$\mathbf{C_{R74}}$	$\mathbf{C_{R75}}$
	RP8	Evaluation	C_{R81}	C_{R82}	C_{R83}	C_{R84}	C_{R85}

With regard to the first dimension, four core design phases/processes are specified:

- **P1:** *Observation.* The foundation upon which survey concepts are built. Quality threats: Preconceived ideas/theories. Limited field of observation.
- **P3:** *Conceptualization.* The process of simplifying/organizing domain-relevant observations. The substantive elements upon which questionnaire items and metadata are built. Quality threats: Preconceived ideas/theories.
- **P5:** *Operationalization.* The translation of domain-relevant concepts into questionnaire items and metadata. Quality threats: Inadequate design skills.
- **P7:** *Administration.* Gathering self-report data by means of an interviewer-administered questionnaire. Quality threats: Sources of measurement error. Inadequate resources (staff and funding).

And four accompanying evaluation phases or processes:

- **P2:** Assessment of Observation Phase
- **P4:** Assessment of Conceptualization Phase
- **P6:** Assessment of Operationalization Phase
- **P8:** Assessment of Administration Phase

With regard to the second dimension, which draws largely on the work of Groves (1987, 1989), five sources of measurement error are specified (for details on the first two sources, which differ from Groves, see Esposito, 2002 or 2003):

- **S1:** Questionnaire: *Content Specialist* [subject-matter experts within a particular domain (e.g., health; labor-force dynamics; income and wealth; demographics)]
- **S2:** Questionnaire: *Design Specialist* (professionals who, in collaboration with content specialists, design questionnaires and develop ancillary metadata)
- **S3:** Interviewer
- **S4:** Respondent
- **S5:** Mode

Several additional aspects of the framework are worthy of note:

"First, it is presumed that design-and-evaluation work can and often does overlap across phases and that movement between certain phases (P1 through P6) is bidirectional and potentially iterative.

Second, the phrase "interdependent sources of measurement error" has been adopted to reflect the view that measurement error – and accuracy, too – is presumed to be the outcome of collaborative/interactive processes involving the various sources of error identified in Table 1. Within a given data-collection context, measurement error is presumed to be a byproduct of role- and task-specific activities … that manifest themselves during the survey administrative phase (P7 or RP7). Various role- and task-

specific activities that are performed inadequately at prior design-and-evaluation phases (P1 through P6) can be viewed as *precursors* to measurement error.

Third, the actual performance of role- and task-specific activities – represented as generically-labeled cell entries (e.g., C_{12}) – is presumed to vary across [questionnaire] design-and-evaluation efforts. Whether a particular cell has an entry or not would depend on whether specific cell-related activities were conducted. For example, if content specialists are not involved in pretesting work conducted during the initial questionnaire design, then cell C_{61} would be left blank. Empty cells are problematic in that they represent activity or knowledge gaps that are apt to increase the locus and magnitude of measurement error.

And lastly, as noted, social, cultural and technological change also plays a crucial role in the measurement process. Unless continuously monitored and accounted for by content and design specialists, rapid change within a given target domain can have a substantial effect on measurement error (Esposito, 2003, p. 55)."

4. Discussion

In this closing section, I will address two issues that have relevance to the question asked in the title of this paper. First: Should evaluation work begin sooner than phase six (P6), the phase during which draft questionnaires are most often pretested? I would say "yes", because *technically* well-designed questionnaire items (e.g., simple/familiar wording; good structure; acceptable working-memory demands), while necessary if high-quality survey data are to be obtained, provide no guarantee that measurement error will be minimized. At the earliest stages of the development/design process, we must seek to establish domain-relevant *grounding* for all of the substantive concepts mentioned in our draft questionnaires – and we need to evaluate that foundational work, whatever the source (see suggestions below). In framework-specific terms (Table 1), we can see that there are many threats to measurement accuracy situated upstream in the early phases of the design-and-evaluation process (e.g., at P1 and P3). Evaluation work conducted during P6 cannot be expected to identify (e.g., using standard pretesting techniques) or successfully remove/avoid (e.g., via modifications to item wording or questionnaire structure) all of these potentially damaging threats. The evaluation process must begin sooner. Postponing evaluation work until P6 could prove unwise for other more pragmatic reasons. For example, sponsors/clients, wanting "hard evidence" of problems, could choose not to implement specific research recommendations due to reservations about the use of qualitative evaluation techniques (e.g., focus groups; cognitive interviews); or they could choose not to implement some recommendations due to the sheer number or

magnitude of problems detected and/or because of insufficient time or funding (see Rothgeb, Loomis and Hess, 2001). Again, the sooner problems are identified, the better the chances that they might be considered and resolved.

Regarding the second issue alluded to above: What actions might practitioners take to assure themselves that questionnaire development/design activities have been properly grounded, empirically and conceptually? Here are some suggestions:

- Request that sponsors/clients provide documentation to support their observations and conceptualizations of the target domain (see Table 1, phases P1 and P3, respectively). In the absence of sufficient metadata – and in collaboration with content specialists, we should consider gathering empirical/behavioral data that might be used to support and/or to evaluate prior observational and conceptual work, and we should make these metadata available to sponsors/clients in a timely fashion.
- During the early stages of questionnaire design, request information on the source of draft questionnaire items and carefully examine available documentation/metadata. For items taken from preexisting questionnaires and incorporated into a new questionnaire, obtain whatever documentation/metadata might be available to assess the empirical and conceptual foundations of these items, and make note of substantial conceptual disparities among draft items and transplanted items.

Because the nature of work differs at different points in the questionnaire design-and-evaluation process (e.g., domain-relevant observations and conceptualizing in natural contexts versus more restricted behavioral observations/activities in laboratory, office or field-based contexts), the research methodologies used in the early phases of that process (e.g., P2 and P4) will tend to differ from those used later (e.g., P6 and P8; for thoughtful discussions of these methods, see DeMaio et al., 1993; Forsyth and Lessler, 1991). So, to the extent that practitioners choose to act on the suggestions offered above, we must expand our research repertoires to incorporate a variety of ethnographic, sociological and social psychological methods (e.g., Beebe, 2001; Gerber, 1999; Glaser and Strauss, 1967/1999; Hox, 1997; Webb, Campbell, Schwartz, and Sechrest, 1966). One might also expect that taking action on the suggestions offered above will have a greater likelihood of success: (1) if communications between content and design specialists begins at the early planning stages of the design-and-evaluation process, rather than later (Rothgeb, Loomis and Hess, 2001), and (2) if every participating group – sponsors/clients; content, design and production specialists; interviewers and respondents – understands the essentially collaborative nature of this process.

References

Akkerboom, H., and Dehue, F., 1997: "The Dutch Model of Data Collection Development for Official Surveys." International Journal of Public Opinion Research, 9, 126-145.

Beebe, J., 2001: Rapid Assessment Process: An Introduction. Walnut Creek, CA: AltaMira Press.

Converse, J.M., and Presser, S., 1986: Survey Questions: Handcrafting the Standardized Questionnaire. Newbury Park CA: Sage.

DeMaio, T., Mathiowetz, N., Rothgeb, J., Beach, M.E., and Durant, S., 1993: Protocol for Pretesting Demographic Surveys at the Census Bureau. Washington, DC: U.S. Bureau of the Census.

Esposito, J.L., 2003: "A Framework Relating Questionnaire Design and Evaluation Processes to Sources of Measurement Error." Proceedings of the 2003 Research Conference of the Federal Committee on Statistical Methodology, Statistical Policy Working Paper Number 37. Washington, DC: U.S. Office of Management and Budget. [Available at: http://www.fcsm.gov/reports/ (at some point in 2004).]

Esposito, J.L., 2002: "Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study." Paper presented at International Conference on Questionnaire Development, Evaluation and Testing (QDET) Methods, Charlestown, SC.

Esposito, J.L., and Rothgeb, J.M., 1997: "Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment." In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), Survey Measurement and Process Quality. New York: Wiley, 541-571.

Esposito, J.L., 1997: "The Survey Design/Redesign Process: A Confluence of Art, Science and Negotiation." Proceedings of the 1997 QUEST Workshop. Örebro, Sweden: Statistics Sweden.

Foddy, W., 1993: Constructing Questions for Interviews and Questionnaires. Cambridge, UK: Cambridge University Press.

Forsyth, B.H., and Lessler, J.T., 1991: "Cognitive Laboratory Methods: A Taxonomy." In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), Measurement Errors in Surveys. New York: Wiley, 393-418.

Fowler. F.J, 1995: Improving Survey Questions: Design and Evaluation. Thousand Oaks, CA: Sage.

Gerber, E., 1999: "The View from Anthropology: Ethnography and the Cognitive Interview." In M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds.), Cognition and Survey Research. New York: Wiley, 217-234.

Glaser, B.G., and Strauss, A.L., 1967/1999: The Discovery of Grounded Theory. New York: Aldine de Gruyter.

Goldenberg, K.L., Anderson, A.E., Willimack, D.K., Freedman, S.R., Rutchik, R.H., Moy, L.M., 2002: "Experiences Implementing Establishment Survey Questionnaire Development and Testing at Selected U.S. Government Agencies." Paper presented at International Conference on Questionnaire Development, Evaluation and Testing (QDET) Methods, Charlestown, SC.

Groves, R.M., 1989: Survey Errors and Survey Costs. New York: Wiley.

Groves, R.M., 1987: "Research on Survey Data Quality." Public Opinion Quarterly, 51, S156-S172.

Haraldsen, G., 1999: "Identifying Clients Needs for Questionnaire Development by a Pre-Survey Questionnaire." Proceedings of the 1999 QUEST Workshop. London: Office of National Statistics.

Henningsson, B., 2001: "An Enlightened Client." Proceedings of the 2001 QUEST Workshop. Washington, DC: U.S. Census Bureau.

Hox, J.J., 1997 : "From Theoretical Concept to Survey Question." In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), Survey Measurement and Process Quality. New York: Wiley, 47-69.

Lindström, H.L., 1999: "Client – Methodologist – Producer Communication to Promote Questionnaire Development and Evaluation." Proceedings of the 1999 QUEST Workshop. London: Office of National Statistics.

Lindström, H.L. and Akkerboom, H., 1997: "Terminology and Abbreviations for the Workshop." Proceedings of the 1997 QUEST Workshop. Örebro, Sweden: Statistics Sweden.

Miller, K., 2001: "Making the Sponsor – Respondent Link in Questionnaire Design." Proceedings of the 2001 QUEST Workshop. Washington, DC: U.S. Census Bureau.

Oksenberg, L., Cannell, C., and Kalton, G., 1991: "New Strategies for Pretesting Questionnaires." Journal of Official Statistics, 7, 349-365.

Platek, R., 1985: "Some Important Issues in Questionnaire Development." Journal of Official Statistics, 1, 119-136.

Rothgeb, J.M., Loomis, L.S., and Hess, J.C., 2001: "Challenges and Strategies in Gaining Acceptance of Research Results from Cognitive Questionnaire Testing." Proceedings of the 2001 QUEST Workshop. Washington, DC: U.S. Census Bureau.

Sudman, S., and Bradburn, N.M., 1982: Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass.

Schwarz, N., and Sudman, S. (eds.), 1996: Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Turner, C.F., and Martin, E., 1984: Surveying Subjective Phenomena, Volume 1. New York: Russell Sage Foundation.

Webb, E.J., Campbell, D.T., Schwartz, R.R., and Sechrest, L., 1966: Unobtrusive Measures: Non-reactive Research in the Social Sciences. Chicago: Rand McNally.

Willis, G.B., Royston, P., and Bercini. D., 1991: "The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires." Applied Cognitive Psychology, 5, 251-267.

Contact

James L. Esposito
Bureau of Labor Statistics
Postal Square Building
2 Massachusetts Ave., NE
Washington, DC 20212
U.S.A.
email: Esposito.Jim@bls.gov

IMPLICATIONS OF SOCIO-CULTURAL FACTORS IN THE QUESTION RESPONSE PROCESS

KRISTEN MILLER

A primary advantage of survey research methodology lies with its potential to depict characteristics of large and diverse populations as well as to make comparisons between distinctive subgroups within those populations. Respondents' differing conceptual and linguistic abilities, however, can present potential barriers to the comparability of survey data. Those with little formal education and who have little familiarity with the survey process are likely to be unsure of the overall intent of survey questions and the intended meaning of specific words. Similarly, respondents' cultural orientation toward particular concepts may vary from that of questionnaire designers; consequently, respondents could experience confusion or miss the intended meaning of those questions. If survey administration and questionnaire design is not adapted to cultural or socio-economic subgroups, data quality and the ability to make accurate representations of subpopulations are compromised.

Theoretical models depicting the phases of question response, however, are primarily informed by psychological, not sociological, principles. These models primarily focus on factors internal to the individual and highlight such concepts as perception, event memory, semantic memory, and means of computation. Socio-cultural conditions are typically regarded as tangential to these cognitive processes; the processes are seen as occurring outside and independently of social and cultural context. Question response models, therefore, characteristically treat question response as a universal process that is generalizable to all survey respondents regardless of social position.

Indeed, the problem of response error in estimates of cultural minority groups or poorer, less educated subpopulations has historically received relatively little attention in the field of survey methodology. The issue is notably absent in literature pertaining to survey measurement error. Only recently has the issue of standards for translations been raised by federal surveys. Federal research review boards have recognized the differing

intellectual abilities of respondents and, consequently, require nation-wide surveys to present their introductory materials at lower reading levels; respondents' clear understanding of their rights as survey participants is seen as an ethical obligation for government researchers. This level of consideration, however, has not been given to the readability of the survey questions themselves and the relationship to response error, though there is little doubt that the use of shorter, simpler words would likely reduce respondent burden and response error. Readability, however, represents only one dimension of difficulty that economically disadvantaged or culturally varied respondents could experience in the question-response process. Providing accessible reading levels does not address respondents' familiarity (or lack of familiarity) with survey protocols, their abilities to understand and respond to a question within the provided format, or the differing interpretations of key terms.

Through analysis of cognitive interviews conducted in 4 different racial and ethnic communities, this paper will illustrate how social and cultural factors can impact the question-response process and, in turn, the comparability of survey data. This paper is based on a study which is currently being conducted in Northwest Ohio's Latino community and suburban and rural poor Anglo communities by the Cognitive Methods Staff at the National Center for Health Statistics. In part, the current study is a follow-up study to previous work conducted January 2002 in rural Mississippi among participants who were primarily poor and had little formal education. The current work expands on the Mississippi project by more closely examining the role of culture, language and socio-economic factors in the question response process for general health surveys. The paper will describe preliminary findings comparing cognitive interviews in these three distinct communities along with interviews that were also conducted in Hyattsville, MD at the National Center for Health Statistics. In this discussion, the paper will illustrate that psychological models of the question-response process are not fully comprehensive; social context impacts the processes by which respondents answer survey questions and, as such, can impact the quality and usefulness of survey data. The paper ultimately calls for greater attention to the relationship of response error and respondent social location.

Methods

This paper is based on two cognitive interviewing projects: one conducted in rural Mississippi in January 2002 and the other conducted in Hyattsville and Northwest Ohio in Summer 2003. Interviews conducted in Mississippi were conducted for the Joint Canada and United States Health Survey and were based on a general health questionnaire for that survey. Interviews conducted in Summer 2003 were based on a collection of general

health, diet, exercise and wealth/income questions that were taken from various national surveys. All of the Latino interviews conducted in Northwest Ohio were conducted in Spanish; translations of the questions were taken from the translations used in their respective fielded survey. Along with the standard objectives of cognitive testing, the primary goals of both projects were to 1) identify the interpretive dimensions of each question, 2) identify various patterns of cognitive processing and 3) identify any indications that data from the various groups would not be comparable. While analysis is currently being conducted on interviews conducted this summer, analysis of the Mississippi interviews utilized grounded theory and the constant comparative method, a common technique for analysis of qualitative data. Patterns of interpretation and cognitive processing are currently being coded in the Ohio and Hyattsville interviews and will eventually undergo quantitative analyses to further investigate comparative differences between the 3 social groups.

Cognitive Interviewing Methods

The purpose of cognitive interviewing is not to obtain survey data or controlled experimental data, but rather to obtain information about the processes individuals use to answer survey questions as well as to identify potential problems that might lead to survey response error. As such, the methods provide insight into: 1) the ways in which cognitive tasks posed by a question are handled by respondents (i.e. comprehension of the question, retrieval of information, and formation of the answer), and 2) whether the answer given by the respondent represents what the question intended.

Data collection for cognitive methods differs dramatically from that of survey methods. While survey interviewing must adhere strictly to scripted questionnaires, cognitive interviewing uses the scripted survey question as only a starting point to begin a more detailed examination of the question response process. Additional probe questions, whether they are concurrent or retrospective, elicit the ways in which participants interpret key concepts, their abilities to recall the requested information, and the appropriateness of response categories. Because the interviews generate narrative responses rather than statistics, results are analyzed using qualitative techniques. This type of in-depth analysis can reveal potential response errors in survey questions and, as a result, can help to improve the overall quality of surveys and survey estimates.

Interview Participants

Twenty-one interviews were conducted in southern, rural Mississippi; 16 were conducted in participants' homes and 5 were conducted in a private room of a community center. All of the participants had telephones in their homes, yet lived in a very rural environment –

many lived down dirt roads that were several miles off main thoroughfares. Additionally, the social services for the area were relatively scant, and most of the very poor were dependent on a local church outreach group to provide food and medicines. All but 6 participants were African American. Most of the participants (15 of them) were in their 50s or 60s. Five were in their 30s or 40s, and one woman was 74 years of age. Of the 21 interview participants, 5 were employed in blue collar or service positions. The remaining participants were either retired, on disability, or unemployed. While we were unable to discern their actual incomes (a few participants, clearly the poorest, were unable to even give an estimate), it was clear from their living conditions that most participants were very poor. Many lived in mobile homes or in houses with only one or two rooms; two of the participants did not have indoor plumbing until the year 1999. As for education, two of the participants had at least some college education and six others had graduated from high school. However, thirteen had not graduated high school, though most had reached the ninth, tenth, eleventh, or even twelfth grades. Two participants did not reach high school; one had reached the first grade, the other had reached the fourth grade.

To date, 52 interviews have been conducted for the Northwest Ohio/Hyattsville project. Twelve interviews were conducted in Hyattsville at the National Center for Health statistics and 40 were conducted in Northwest Ohio. Of the Ohio interviews, half were conducted English and the other half were conducted in Spanish. Though a few Hyattsville participants were economically disadvantaged, Northwest Ohio participants were poorer and had the least amount of education. Many of the interviews conducted in Spanish were among first generation immigrants from Mexico who spoke little to no English. Unlike the Mississippi participants, the Northwest Ohio participants had access to social service programs. For example, many of the Mexican participants were taking English classes in a larger metropolitan area, and many of the Anglo participants were receiving services from the local county health department. The following charts outline the demographic characteristics of participants from the four interview groups:

Southern Mississippi Participants, January 2002

	Race/Ethnicity	Income	Education	Gender
Mississippi (English) 21 Participants	White = 6 Black = 15	<20K = 19 20-30K = 2	Elementary = 2 Some H.S. = 13 H. S. Grad. = 6 Some College = 2	♀ = 15 ♂ = 6

Northwest Ohio & Hyattsville Participants, Summer 2003

	Race/Ethnicity	Income	Education	Gender
Hyattsville (English) 12 Participants	White = 7 Black = 5	11-20K = 4 21-30K = 1 31-50K = 3 51-80K = 1 61-80K = 0 81K+ = 3	Elementary = 0 Some H.S. = 2 H. S. Grad. = 4 Some College = 6	♀= 7 ♂= 5
NW Ohio (English) 20 Participants	White = 20 Black = 0	0-10K = 8 11-20K = 7 21-30K = 5	Elementary = 2 Some H.S. = 13 H. S. Grad. = 3 Some College = 2	♀ = 12 ♂ = 8
NW Ohio (Spanish) 20 Participants	Mexican = 11 Puerto Rican= 1 Mex. Am. = 8	0-10K = 7 11-20K = 5 21-30K = 3 31-50K = 4 51-80K = 1	Elementary = 2 Some H.S. = 12 H. S. Grad. = 4 Some College = 2	♀ = 16 ♂ = 4

Findings: Comparisons Among the Cognitive Interview Groups

Three primary themes form the basis for a preliminary comparison of the 4 different interview groups: 1) Lacking knowledge of question-response protocol and respondents' expected role, 2) Responding outside the expected cultural frame of reference, and 3) Responding outside questions' systems of knowledge. Unlike the other three groups, most Mississippi participants lacked even a general knowledge of the survey process and were unaware of the protocol for their role as the respondent. As such, Mississippi participants, unlike any of the other participants, needed to be instructed and guided by interviewers in order to participate in the question-response process. Latino respondents, on the other hand, (especially those who were first generation Mexican immigrants) differed in their cultural orientation toward various questions – most markedly questions regarding meals and diet. Their differing cultural frame of reference caused differing interpretations and confusion in the response process to these questions and, in many cases, lead to response error. Finally, Mississippi and both Northwest Ohio groups (in comparison to the Hyattsville interviews) had difficulty responding to questions that were written through foreign systems of knowledge – in this case from a medical perspective. Though Mississippi and Latino responses generated more errors, Northwest Ohio Anglo participants also had difficulty reporting whether they had specific chronic conditions and whether they had received mammogram or PSA tests.

It is important to note that, while all three comparative themes involve cognitive processes, they are rooted specifically within sociological processes. It is the social and cultural position of the respondent that informs the cognitive processes. This is seen most clearly in the comparison of differing social and cultural groups. The remainder of the paper will more fully describe these themes, drawing comparisons of the question-response process between the 4 interview groups.

Lacking Knowledge of Question-Response Protocol and Respondents' Expected Role

It has long been recognized, within the field of survey research, that the interview is best conceptualized as a social interaction, bound by social norms and patterns of expectations; the survey interview is viewed as a social system, involving two roles (the interviewer and the respondent) who are united in a common task. This conceptualization of the survey process, one that lays out a relatively well-executed interaction between the respondent and interviewer (essentially strangers bound together by a shared purpose), lacks recognition that the interaction is dependent upon a pre-existing familiarity with surveys and the question-response process.

While this paradigm, in all likelihood, is an excellent depiction of the vast majority of survey interviews, it does not so clearly depict interactive patterns found in many of the Mississippi interviews. To be sure, the interviews illustrate that the normative patterns which frame a successful survey interview are *not* rooted within a universal system of knowledge. While some of the Mississippi participants were familiar with the survey concept, a few had never before participated in a survey, were not entirely sure what surveys were used for and had no previous knowledge of questionnaire design or format. Consequently, these participants were unable to engage in the interview with the kind of ease typical of survey respondents. One 60 year old woman, for example, struggled with the process throughout the entire interview because she was operating with the impression that her answers were to be previously conceived; she did not realize that she was expected to formulate her response *as* she was being asked the question. Though the interviewer provided instruction as well as positive reinforcement, she continuously expressed an inability to answer even general questions about her own health:

> PARTICIPANT: I had never been asked these questions before. That's the reason I really don't know how to answer these questions. I'm doing the best I can.
>
> *INTERVIEWER: Oh, you're doing a great job. You're doing fine.*
>
> PARTICIPANT: I'm doing the best I can…. because, like I say, some questions you all [are] shooting out here to me… I have never heard before.

As the passage illustrates, the participant was not aware of what was expected of her in the role of respondent. While providing an impromptu response (even if it is not quite accurate) is in the purview of "being a survey respondent," this woman did not know this prior to the interview.

Another critical expectation which was also unmet by many Mississippi participants is that respondents should understand that they are to ultimately produce an answer that will fit within a provided response category. Additionally, they should understand, if their "real answer" does not squarely match the provided category, they can "make do" and adjust so that their answer is categorizable. Of course, this issue is a common consideration in questionnaire design; it is typical for a respondent to protest that no category adequately represents their experience. Some Mississippi participants, however, were unaware of this expectation. Because they did not understand the mechanics of survey research, they considered any type of answer (as long as it answered the question) as suitable. This is illustrated in an interview with a 34 year old woman who, throughout the course of the interview, increasingly became upset each time the interviewer asked for clarification or refinement of her initial response so that it could be appropriately categorized within the survey format. The discord finally came to a head when she was asked the question, "When was the last time you had a pap smear?":

> *INTERVIEWER: Okay. And when was the last time? Less than one year ago…*
>
> PARTICIPANT: Last year.
>
> *INTERVIEWER: Was it less than one year ago or one to two years ago?*
>
> PARTICIPANT: Last year!
>
> *INTERVIEWER: So, does that [mean]…*
>
> PARTICIPANT: A year ago!

To ease the situation, as well as to obtain a code-able response, the interviewer was compelled to explain the fundamentals of questionnaire design. The interviewer needed to convince the participant that she was not being rude (which was the woman's understanding), but was merely following a set of instructions that were given to her by someone else…

> *INTERVIEWER: I got these ridiculous categories. Look what I have. [The interviewer shows her the sheet of paper with the written categories] I have less than one year ago and one year to less than two years ago. So, how…*
>
> PARTICIPANT: [Now understanding, the woman kindly pats the interviewer on the leg and interrupts] Put less than two years ago, then.

Now, understanding that there were pre-written response categories (response categories that were written by someone else and were, consequently, not changeable), the participant learned an aspect of the role of survey respondent – to provide a code-able answer. This lesson in questionnaire format was pivotal in the interview; from that point, the interaction was much more pleasant and easy-going.

Similarly, those unfamiliar with the format of survey questions had particular difficulty with scale questions that used generic, essentially non-descript, response categories, such as "mild," "moderate," "severe," and "extreme," but relied on incremental order to convey meaning. One 59 year old man, for example, had difficulty responding to such a question, contending that "to me, mild and moderate are about the same thing." It was also the case that one Northwest Ohio woman (of all Ohio interviews) did not know the meaning of the term "moderate" and, like Mississippi participants, did not intuit the meaning from the increasing order and, consequently, ignored that response category. Once the term was explained to her, she recognized that "moderate" was the most accurate response and requested to have her original answer changed. Most noticeably, the Mississippi woman who earlier had worried that she would not be able to answer the survey's questions also struggled with these response categories…

> *INTERVIEWER: Overall, in the last 30 days, how much difficulty did you have with work or household activities? Would you say none, mild, moderate, severe or extreme?*
>
> PARTICIPANT: I guess that's one I can't hardly get because my housework… is hard.
>
> *INTERVIEWER: Do these just – these categories just not fit you?*
>
> PARTICIPANT: It's the categories, you know, that I can't, you know, get them right, so I really can't just get them right.
>
> *INTERVIEWER: If you were to describe, in your own words, how much trouble you have with work and household activities, how would you describe that?*
>
> PARTICIPANT: Okay. In my own words, okay, when I get ready to do something, I can't do it. If I get ready to dust, now, I can do it if I sit down on the floor, scooting around. I can do it that way.

Because the woman was unable to make sense of the categories, the interviewer needed to explain the ordering of the categories; she needed to make explicit the implied meaning of the categories.

> *INTERVIEWER: Well, if this is like a little trouble, [interviewer puts her hand down toward the floor] and this is kind of like middle trouble [interviewer raises her hand to waist-level], and this is like a lot of trouble [interviewer raises her hand again to*

head-level], where are you? [Indicating again with her hand,] Are you near the top, are you near the bottom, are you kind of in the middle?

PARTICIPANT: I'm here, about [She puts her hand at head-level].

INTERVIEWER: You're like a lot of trouble.

PARTICIPANT: Right.

INTERVIEWER: So, it's difficult?

PARTICIPANT: It's difficult.

For many of the Mississippi participants, like the woman in the above passage, once they were taught about the survey interaction, the question-response process became much more straight-forward. This was not the case, however, for a few participants. These participants did not grasp the formality of the question-response process, and could answer questions only if they were restated conversationally. This 68 year old man with a first grade education, for example, was unable to provide health information through a structured survey question:

INTERVIEWER: The next few questions are about limitations in your daily activities caused by a health condition or problem. Do you have any difficulty hearing or seeing?

PARTICIPANT: What was that?

INTERVIEWER: Do you have any difficulty hearing or seeing?

PARTICIPANT: I don't understand.

INTERVIEWER: Okay.

PARTICIPANT: I don't understand, that's why I want my wife to take it. You can ask her the same thing, and you can get all that wrote down.

The interviewer chose to ignore the respondent's request to have his wife serve as a proxy. Instead, he restates the question so that it is communicated conversationally. In this less structured format the participant is able to clearly understand the question and relay rather detailed health information….

INTERVIEWER: Can you hear all right?

PARTICIPANT: I can hear a little bit, but not too much. I hear sometimes like if you talk real plain. Some people talk real plain to me, and I can understand them pretty good.

INTERVIEWER: How about seeing, can you see okay?

PARTICIPANT: I got a cataract on this eye. I can't see out of this eye. I can see out of this eye. I can see, but I can't see real good, there's like a skim over it. It's dim.

INTERVIEWER: Do you have any trouble walking?

> PARTICIPANT: I can walk all right. When I walk my back will hurt me. If I walk a little piece, a couple times like from here outdoors I'd have to sit down because all across my back here and my side it would be hurting so bad I would have to sit down at the end.

As these excerpts have illustrated, the survey process necessarily holds expectations for the respondent. Examining the ways in which respondents struggle with the interaction points to the types of expectations inherent in the question response process that typically go unnoticed (e.g., knowing that responses are impromptu, knowing to provide a code-able answer, or discerning the meaning of questions in a structured format). Survey interviews, as they are conceptualized as social interactions with normative patterns of expectations, are necessarily bound within a system of knowledge. Those respondents who do not have access to that particular system of knowledge will struggle in the interaction, will need to be educated about what is expected of them, or may simply not be able to complete the interaction within the standardized format required of survey design.

This type of difficulty was not seen to this degree in the Northwest Ohio group, among both the English and Spanish speaking participants. In only one English interview (a man in his 40s with Down's syndrome) and in one Spanish interview (a woman in her 50s who had spent her life as a homemaker) did Northwest Ohio participants need to be guided through the protocol of the question-response interaction. And, as described before, in only one English interview did an Ohio participant fail to intuit the meaning of vague Likert scale response categories. Because we had originally conceptualized this problem as being related to income and education-level, we were somewhat surprised to not observe the same degree in Northwest Ohio interviewees who had similar levels of formal education. Upon further reflection, however, we believe that Northwest Ohio participants, unlike the Mississippi participants, had much more exposure to other systems of knowledge and, consequently were much more adaptable and resourceful in new interactional settings. We expect that subsequent analysis of the Ohio interviews will illuminate more detail regarding this hypothesized relationship.

Responding Outside the Expected Cultural Frame of Reference

From the very beginning of Spanish interviews, it was clear that some translated survey questions caused interpretation difficulties for Latino participants. That is, particular words were translated literally from English and, because of cultural differences, did not convey the same meaning. For example, the phrase *frijoles con chile* was intended to mean chili beans, but was interpreted by most Latino participants as beans with hot sauce. Additionally, some words varied by particular region (e.g., Puerto Rican Spanish uses

nami for yam, while Mexican Spanish uses *camote*) or were more formal forms of Spanish (e.g., the word *fiambre* for lunchmeat). Consequently, these terms were not always understood by Latino participants. Similarly, some words in Spanish had more than one meaning and could easily be taken out of context. For example, the word *comida* can mean meal, food, and the name of a meal – like the Anglo word for dinner. Consequently, the question "Did you eat a morning meal?" was translated as *"¿ Ayer comío Ud. la comida de la mañana?"* but understood by participants as "Did you eat your dinner in the morning?"

In addition to recognizing the language differences, the Latino interviews revealed cultural differences (unrelated to translation problems) that impacted the question response process. For example, traditional Mexican meal patterns differ from the meal patterns of Anglo-Americans. Although customs around food or eating are tied to the social class of its consumer, as well as the time of year, several of Mexican meal patterns are an institutional part of the culture. These traditional Mexican meal patterns continue through migration to the United Sates and have managed to survive, to varying degrees, depending on acculturation to the working and eating patterns of western culture. (See Appendix A for a more detailed description of the meal patterns of many Mexican families.) Consequently, food and diet questions – that were structured in the questionnaire by the Anglo-American meals of breakfast, lunch and dinner – created difficulty for most of the Mexican participants who used their own cultural meal pattern as a frame of reference for responding. For example, this passage illustrates difficulty because of the confusion over the multiple meanings of the word "comida" but also because of the differing meal patterns:

INTERVIEWER:	*Digame, Ayer comio usted la comida de la manana?* Tell me, did you eat a meal in the morning?
PARTICIPANT:	No.
INTERVIEWER:	*Cuando digo "la comida de la manana" que viene a su mente. Teine un nombre esa comida de la manana?* When I say morning meal, what comes to your mind? Does it have a name, that morning meal?
PARTICIPANT:	*Pues, un nombre del estillo de la comida* Well, a name for the type of food?
INTERVIEWER:	*Tiene otro nombre Usted para distingir esta comida a otros comidas?* Do you have a name to distinguish this meal from other meals?
PARTICIPANT:	*No yo le diria que seria el mismo nombre pero no lo uso asi.* No, I would say that it is the same name but I do not use it that way.

INTERVIEWER: *Y la manana para Usted , que quiere decir, que tanto tiempo, de que horas a que horas*
And morning for you, what does it mean, what time frame or from what hour to what hour is it?

PARTICIPANTS: *Pues en la manana el desayuno es a las nueve.*
Well in the morning "el desayuno" is at nine .

INTERVIEWER: *So, el desayuno, lo nombra el desayuno, es a las nueve ?*
So, "el desayuno, you name it "desayuno", is at nine?

PARTICIPANT: *Si, por que you no doy el que le dicen….como le dicen…..... Braaq faat*
Yes, because I don't serve, what they call…… how do they say…….. Braaq faat

INTERVIEWER: Breakfast?

PARTICIPANT: *Si.*
Yes.

INTERVIEWER: *No hace breakfast sino que hace desayuno?*
You don't make breakfast, but you make "desayuno"?

PARTICIPANT: *Si, yo desayuno, asi estoy acustumbrada…doy mi desayuno y mi comida y en la cena como algo mas liviano.*
Yes, I have "desayuno," that is how I am accustomed… I serve "desayuno," and than my "comida" and for "cena," I eat something a lot lighter.

Even more understated, but perhaps more consequential for health surveys, we found that the concept of health, itself, differed dramatically from Latino and Anglo participants. While most of the Anglo participants (77%) conceptualized health as a physical phenomena, most of the Hispanic participants (90%) used a far more comprehensive conceptualization of health, incorporating emotional and spiritual dimensions. For example:

INTERVIEWER: *Diria que su salud en general, es excelente, muy buena, buena, regular o mala?*
Would you say your health in general is excellent, very good, good, fair or bad?

PARTICIPANT: *Excelente.*
Excellent.

INTERVIEWER: *Y porque dices excelente ?*
And why do you say excellent?

PARTICIPANT: *Porque me siento bien hasta ahorita, no tengo ningun ilimento fisico, este, tengo ganas de vivir cada dia. Me levanto con ganas de seguir adelante y evocar a mis hijos que sean Buenos hijos.*

Because I feel fine up to now, I don't have any physical limitations, and I have the desire to live each day. I get up with the desire to keep going and evoke my children to be good children

INTERVIEWER: *Y porque no " muy buena en lugar de excelente.*
And why not, very good in place of excellent?

PARTICIPANT: *Porque si diria no muy buena, me sentiria enferma. Y es todo a lo contrario, me siento bien. Fisicamente me siento bien, Moralmente tambian.*
Because if I said no very good, I would feel ill.
And it is totally contrary, I feel good. Physically I feel good, morally as well.

INTERVIEWER: *Cuando dises Moralmente que quires decir?*
When you say Morally what are you saying?

PARTICIPANT: *Quiero decir que moralmente, que yo me siento, o sea , mi espiritu, me siento bien con migo misma y con las personas que me rodian.*
What I mean to say is morally, I feel, that is my spirit, I feel good with myself and with the persons that surround me.

Upon reflection, it is not surprising that Latino participants – especially those who were raised in Mexico – would more closely associate health with spirituality. The tradition of Mexican medicine, *curanderisimo,* is directly connected with ritual and a more holistic sense of well-being. It is interesting to note that the two Latinos who did not hold a comprehensive view of health were second generation Mexican Americans and, consequently, had assumed more Anglo cultural customs. At this point, the extent to which differing conceptions of health may impact the quality of survey data for this general health question (and possibly other subjective health questions) is unclear. We hope that subsequent analysis of the Ohio and Hyattsville interviews will more fully illuminate this issue.

Responding Outside Questions' Systems of Knowledge

The final comparative theme regards respondents' abilities to form an answer to a question that is rooted within a knowledge system existing entirely outside their own frame of reference. That is, the question addresses a matter that, in no way crosses the respondents' own personal knowledge; the words or language used in the question is not what they normally use to describe their experience. As in the case of these particular questionnaires, the system of knowledge that respondents were required to address was medical knowledge, and the clearest example of this was in the chronic condition section of the questionnaire. In that section respondents are asked about various health conditions

in which they have been diagnosed. The particular intent of this set of questions is to track *doctor diagnosed conditions*, and respondents are required to report information told to them by their doctor. However, as is evident from the following passages, respondents – especially those with limited access to good health care or those who are unable to retain what was told to them by their doctor – are not always able to accurately report this information. This theme was present among all interview groups, though appeared less frequently among Hyattsville participants who, for the most part, had adequate health resources. For example, this 30 year old Mississippi man, like several participants from Mississippi and Ohio, confused the condition of "chronic bronchitis" with the condition of "acute bronchitis"…

> *INTERVIEWER: Do you have chronic bronchitis?*
>
> PARTICIPANT: I think I do. I'm not for sure.
>
> *INTERVIEWER: Okay. What -- tell me what you're thinking.*
>
> PARTICIPANT: As far as what? About the bronchitis?
>
> *INTERVIEWER: Yes.*
>
> PARTICIPANT: Well, I was just trying to think of, you know, you know, I think I've – like when I go to the doctor, I got a cold and, you know, I'm diagnosed I got bronchitis.

A similar problem occurred with the Mississippi man who was recently diagnosed with chronic bronchitis. He knew that he definitely had chronic bronchitis, but when he was asked if he was diagnosed with asthma, he also answered affirmatively because he was under the impression that he was taking asthma medication:

> *INTERVIEWER: Do you have asthma?*
>
> PARTICIPANT: I guess I do. He's [the doctor's] got it down as that acute… [The subject is trying to remember the exact diagnosis and has trouble pronouncing the name]… ex-car-bor-ation…. is what I'm trying to say, chronic bronchitis. Now he has given me asthma medications, inhalers, as you can see over there on the counter, about five or six different types. All that has to do with the chronic bronchitis. It's acute, it's severe….
>
> *INTERVIEWER: So, I guess I'm a little bit confused about – and maybe it's because I don't understand the medical terms, but that – is there a difference between asthma and chronic bronchitis?*
>
> PARTICIPANT: I could not answer that. I don't know. Wheezing of the chest is what I have. What it does is when you – smoking doesn't help, of course, we all know that. When I catch a cold, I am susceptible to pneumonia almost immediately.…

INTERVIEWER: Okay. Now, do you remember specifically if your doctor said you had asthma?

PARTICIPANT: No. He gave me asthma medication….I know for a fact, I had the… [respondent has trouble articulating the actual diagnosis] … acute blah, blah, blah chronic bronchitis.

INTERVIEWER: So, this question right here, do you have asthma? Is that a tricky question for you to answer?

PARTICIPANT: No, it's not a tricky question. It's just I can't answer it honestly. I can't say yes, I can't say no, because I don't know.

INTERVIEWER: Okay.

PARTICIPANT: I don't want to lie to you in this interview. All I can do is tell you the truth.

INTERVIEWER: Okay.

SUBJECT: That's all I can do.

INTERVIEWER: Okay.

SUBJECT: I can't answer that truthfully.

By far, the biggest problem in the chronic conditions section was the various heart conditions: coronary heart disease, angina, heart attacks, and congestive heart failure. These questions contained words that were filled with medical jargon, and while participants would know that they had problems with their heart or that they indeed had heart disease, they were uncertain if they had "coronary heart disease." For example, when asked, "Do you have coronary heart disease?," one woman responded:

PARTICIPANT: I have an enlarged heart. I don't know whether that would be [coronary heart disease] – it is a disease though. It is, but I don't know about that.

INTERVIEWER: And you know that's a disease?

PARTICIPANT: Yes.

INTERVIEWER: But, you don't know if it qualifies – if it counts as coronary heart disease?

PARTICIPANT: No.

When asked the same question, another woman responded:

PARTICIPANT: I know I have heart disease, but I don't know – I don't know what you call it, but I know he said I had a bad heart.

INTERVIEWER: Okay. So you – tell me what you know. You know you have a bad heart?

PARTICIPANT: He said one of the valves wasn't pumping fast or something. Needs to open and close better, and when I walk a lot or I even try to run, it pumps, you know, my heart starts to beating real fast… That's what he said. He gave me some nitroglycerine pills.

INTERVIEWER: Okay. So, you don't know if you have coronary heart disease?

PARTICIPANT: No, I don't.

In another interview, a man experienced the same type of difficulty:

PARTICIPANT: I really don't understand it, but I do have a problem with my heart. Sixty percent of it is closed, because they had to go in so they can see what was wrong, and they found out it was sixty percent closed, you know….

INTERVIEWER: Okay. Did your doctor tell you that you had coronary heart disease? Did he use those words?

PARTICIPANT: I don't know for sure what he used on that, but I knew he said I had heart problems.

INTERVIEWER: You just - you know you have a heart problem?

PARTICIPANT: I think it's on that thing there, too. I think what you said was that coronary whatever you want to call it, yes. I would say yes.

What this man refers to in this passage ("it's on that thing there") is a report from his doctor that describes his entire medical condition. From this report, we were able to learn that this participant was diagnosed with final stage emphysema, chronic bronchitis and congestive heart failure. He could not read the report and was unable to accurately report his conditions:

INTERVIEWER: Do you have congestive heart failure?

PARTICIPANT: No. I never had failure.

INTERVIEWER: Okay, you would say no to that?

PARTICIPANT: Yes.

INTERVIEWER: Okay. What does that mean to you? Congestive heart failure, what do you think that means?

PARTICIPANT: It means just like if you, you know, having a heart attack or something like that….. To me, you know, some people have heart pain and it doesn't kill them, but you know, I never had that problem.

INTERVIEWER: It's when your heart stops pumping, is that what you mean?

PARTICIPANT: And then they bring in - you know how they shock you and bring you back or something. I never had that problem.

These questions ask respondents about matters in which they do not know the answers. The answers to the questions lie outside their own system of knowledge; they are being asked to be informants for their doctors, (that is, other knowledge systems) and those who have little education have a particularly difficult time making that leap.

Those respondents who were economically well off were more likely to have health insurance, received more attention from their doctor, and consequently, were more likely to understand their medical condition. Interestingly, one woman from Ohio who had been diagnosed with congestive heart failure was able to describe her disease to great detail though she had only a high school education and was living in extreme poverty. The most telling difference between her situation and the others who were unable to correctly respond to this question was that she was under the care of the county health department and receiving many social services, including weekly visitations of a nurse practitioner who was able to educate her about the disease.

The inability for participants to provide answers brings attention to the fact that survey questions often ask respondents to provide information that is outside their personal knowledge base. In these two particular questionnaires, the problem was found not only in the chronic condition section, but also in questions about medical tests ("Have you ever had a PSA test?" "A mammogram?"), in causal questions ("Which one of the following is the best description of the cause of this condition: accident, existed from birth or genetic, work conditions, disease or illness, aging, or emotional or mental health problem or condition?"), and in medication questions ("In the past month did you take tranquilizers such as Valium?," "In the past month did you take an anti-depressant?") Participants who were taking Xanax for a variety of reasons ("for nerves," "to help me quit smoking," "to calm me so I can sleep") could name the actual medication, but could not properly classify the drug as a tranquilizer, an anti-depressant or a sleeping pill.

Conclusion

Conducting cognitive interviews with small subsections of the US population serves two functions for survey research. On a practical level, cognitive analysis of interviews with different cultural and racial/ethnic groups suggests several guidelines for improving questionnaires so that survey research can advance the quality of estimates for these subpopulations. On a more theoretical level, however, identifying various ways in which participants from various social groups were unable to negotiate the survey interaction provides clearer insight into the intricacies of the question-response process – interactive aspects that are otherwise invisible. Because a number of the Mississippi participants had never participated in a survey and were entirely unfamiliar with the expected patterns of interaction, these interviews help to articulate basic expectations of the survey

respondent. These interviews suggest that respondents' particular social location does influence how respondents make sense of and answer survey questions. Fully understanding this relationship between socio-cultural phenomena and the response process is vital for health disparities research as well as survey research occurring within international and multi-cultural contexts.

Appendix A: Mexican Meal Patterns

Desayuno	from the Latin term "dis-lunare" which means to break the fast, is consumed after a night's fasting, after one awakens in the morning. It often consists of something light such a cup of coffee and some bread, tortilla or hot cereal.
Almuerzo	follows the *desayuno* a heavier breakfast meal, which may consist of chorizo with eggs, fried potatoes, refried beans, tortillas, fruit, coffee, milk or juice. Not to be confused with western culture's lunch, this meal is consumed before 11:00 a.m Depending on a person's schedule, one does not necessarily have to have a desayuno before having almuerzo. A mother may fix a big almuerzo before the children go off to school while the father would have a light desayuno before going of to work. Yet in the summer months or on weekends, when school was out families might have the large Almuerzo later in the morning after having had a cup of coffee, milk, juice and some pan dulce shortly after awakening. On Sundays, women might prepare a large pot of "*Caldo*"chicken or beef vegetable soup with corn tortillas which the family would have for Almuerzo after attending early morning mass.
Comida	served any time in the afternoon between noon and 4 p.m., is the heaviest meal of the day. This large meal may consist of a "*sopa*" soup, a main dish or "*guisado"*, such as beef, chicken or pork with gravy accompanied by beans and rice, tortillas; a fruit beverage followed, perhaps, by *"postre"*, a desert. The comida is typically consumed at noontime during the school year, at 1 or 2 during the summer months except on Sundays where it might be late in the afternoon, before 4.
Cena	a lighter meal consumed at night. It may consist of milk, hot chocolate with some bread, a warm flour tortilla with some refried beans, some fruit or left overs from the Comida. If the amuerzo and comida were substantial and one had a merienda, the cena might be bypassed all together.
Merienda	is gathering moment with the family where they converse and partake of "pan dulce" sweet bread or pastries with coffee, Mexican chocolate or some other "*antojitos,"* whim such as " *champurrado*" a hot chocolate pudding. It was the custom in some families to have "merienda" at 3 in the afternoon. The clock could be set by this ritual. At 3 p.m., a mother would start the coffee or make Mexican chocolate and aunts, cousins and married siblings would arrive with a bag of "pan dulce" from the Mexican bakery and would visit as coffee, milk or chocolate are served with sweet bread. This merienda ritual resembles the European afternoon tea more than a Western afternoon snack.

Contact

Kristen Miller, PhD
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782, U.S.A.
email: ktm8@cdc.gov

COGNITIVE LABORATORY EXPERIENCES AND BEYOND: SOME IDEAS FOR FUTURE RESEARCH

GER SNIJKERS

1. Introduction

In the literature on questionnaire design and survey methodology, pre-testing is mentioned as a way to evaluate questionnaires (i.e. investigate whether they work as intended) and control for measurement errors (i.e. assess validity). As the American Statistical Association puts it (ASA, 1999, p. 11): "The questionnaire designer must understand the need to **pretest**, **pretest**, and then **pretest** some more." Clark and Schober (1992, p. 29) indicate why this need to pre-test: "Surveyors cannot possibly write perfect questions, self-evident to each respondent, that never need clarification. And because they cannot, the answers will often be surprising."

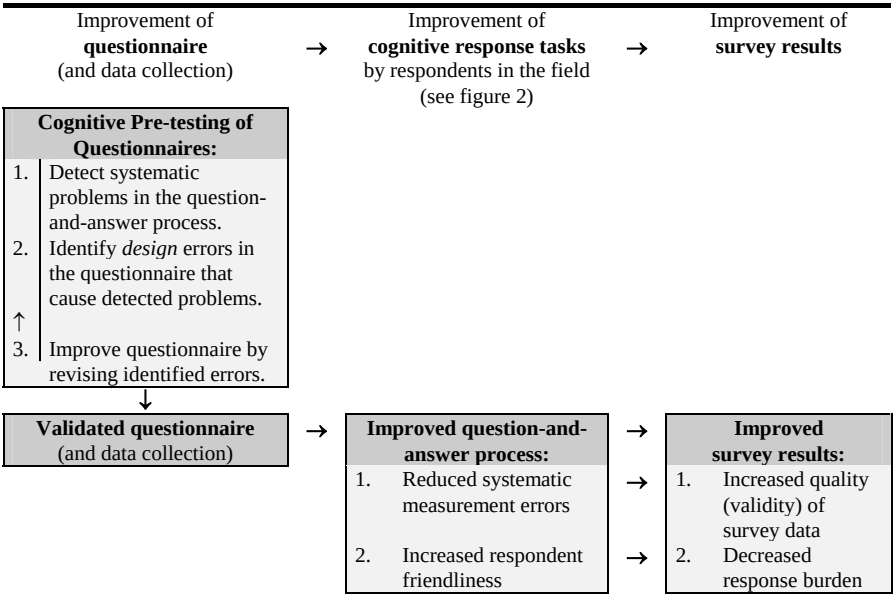
In the every-day practice of survey design, however, pre-testing and its results are not always accepted. A general feeling towards pre-testing is expressed by Converse and Presser (1986, pp. 51-52): "Pretesting a survey questionnaire is always recommended – no text in survey methods would speak against such hallowed scientific advice – but in practice it is probably often honored in the breach or the hurry. There is never the money nor, as deadlines loom, the time, to do enough of it. There is a corollary weakness that the practice is intuitive and informal. There are no general principles of good pretesting, no systematization of practice, no consensus about expectations, and we rarely leave records for each other. How a pretest was conducted, what investigators learned from it, how they redesigned their questionnaire on the basis of it – these matters are reported only sketchily in research reports, if at all. Not surprisingly, the power of pretests is sometimes exaggerated and their potentials often unrealized."

It has almost been twenty years since this text has been written. Although progress has been made, still several aspects of pre-testing as mentioned by Converse and Presser need to be addressed. In this paper, some personal ideas for future research will be discussed. These ideas have been presented at the 2003 QUEST Workshop in Mannheim. These ideas will be discussed in the sections 3 and 4, followed by a conclusion in section 5. But first, the general framework of pre-testing will be presented briefly in section 2.

2. The aim of cognitive pre-testing

Cognitive pre-testing is not an end in itself; it is aimed at improving the data quality, by improving the questionnaire. By means of small-scale pre-testing the questionnaire is validated, i.e. errors in the questionnaire that cause systematic errors in the question-and-answer process of the respondent in an interview setting are detected, explained and improved (in an iterative process). In this way, the questionnaire will be adapted to the question-and-answer process and becomes easier to answer, within a shorter period of time, and will be more respondent-friendly. Thus, resulting in reduced measurement errors, i.e. increased quality of survey data, and reduced respondent burden. This is the CASM[1] paradigm (see figure 1).

Figure 1: The CASM paradigm: Validating questionnaires and improving survey results by cognitive pre-testing



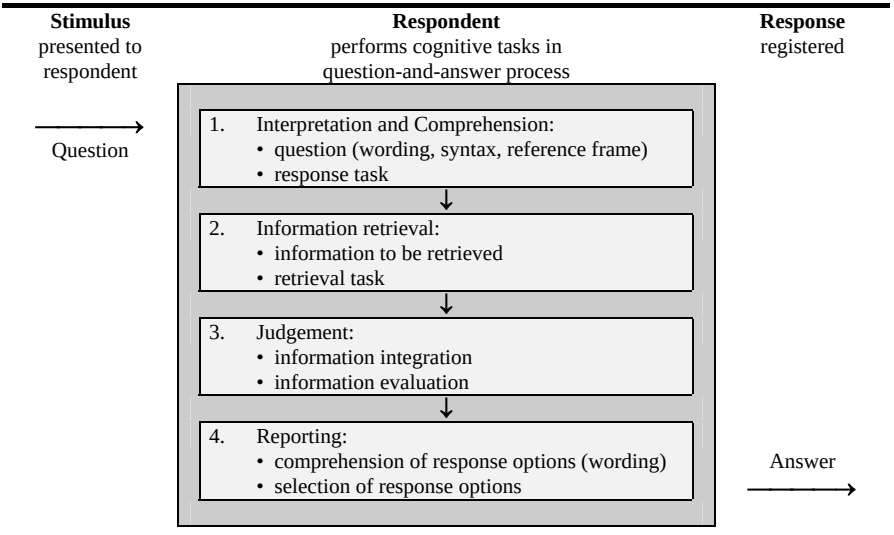
From: Snijkers (2002, p. 225).

1 CASM means Cognitive Aspects of Survey Methodology. For more information on CASM see Jabine et al. (1984), Hippler et al. (1987), Sirken et al. (1999), and Presser et al. (2004).

The question-and-answer process has been modelled by Tourangeau and Rasinki. In 1988 they presented a 4-stage model: comprehension, retrieval, judgement and reporting (see figure 2). This model "offered a view of the survey respondent as a question-and-answer system that carried out a series of mental operations, such as comprehension of what was required in response, retrieval of relevant information from memory, and decision-making to arrive at and provide answers to the survey interviewer's inquiry." (Jobe & Mingay, 1991, p.178.) According to Jobe and Mingay (1991, p. 178), "modelling the respondent's mental operations represented a vast change over the simple stimulus-response conception of respondent behaviour, that from the beginning of modern survey-taking governed the principles employed in designing survey instruments."

Methods to research the question-and-answer process, used in pre-test research are, among others: expert (re-)appraisal, focus groups, in-depth interviews (including thinking aloud and follow-up probing, meaning-oriented probing, paraphrasing, targeted test questions, and vignettes), and behavioural coding. These methods have been applied in cognitive laboratories to evaluate questionnaires to be used mostly in face-to-face interviews.

Figure 2: The question-and-answer process within the stimulus-response model of survey responding



From: Snijkers (2002, p. 7)

3. Needed Research

Now that we know the objectives of pre-testing, we can look at research that is needed in order to meet these objectives better. According to me, future research that is needed should address the following issues:

1. Assess the effectiveness of pre-test methods.
2. Develop empirically based guidelines for questionnaire design.
3. Adapt pre-test methods to new survey design issues, like Web surveys and Business surveys.

In this section these research issues will be discussed.

3.1. Assessing the effectiveness of pre-testing methods

Research with regard to assessing the effectiveness of pre-testing methods incorporates a number of issues that need more research. These issues deal with:

- reproducibility of methods,
- standardisation of terminology,
- new methods to research the question-and-answer process, and
- validity of methods.

As for reproducibility of pre-testing methods, Converse and Presser (1986, see section 1), already said what is needed: “general principles of good pretesting”, and “systematization of practice”. This is confirmed by Willis et al. in 1999 (p. 137), who discussed systematic schemes for describing the practise of cognitive interviewing methods. They concluded that “(…) no such schemes exist for use in cognitive interviewing research.”

Research concerning reproducibility of pre-testing methods is aimed at developing *Current Best Practices*. This concerns questions like:

- What pre-testing methods do we use, in what situations?
- How have they been applied? E.g.:
  - What probes are used in what situations, and what kind of findings do they produce? What are good probes?
  - Can pre-test methods be used in other modes, like the telephone and for pre-testing self-administered questionnaires?
- What findings do these methods result in?
- Who has to conduct pre-test research?
- How can these methods be improved?

Basically, aiming at Current Best Practices means a systematic description of the methods that currently are being used (see e.g. DeMaio et al., 1993, and Snijkers, 2002, chapter 4) and standardisation of these methods. To achieve this goal e.g. at the QDET Conference[2] a number of short courses were presented (Willis & Forsyth, 2002; Gerber, 2002; Mathiowetz, 2002). Most papers presented at the 2003 QUEST Workshop addressed this issue, describing own experiences in conducting pre-test studies and their results.

In order to get Current Best Practices, it is necessary for all pre-test researchers to speak the same language. Thus, also *standardisation of terminology* is needed. At the QUEST meeting, a good start has been made by Esposito (2003), who presented a draft lexicon of concepts.

Developing Current Best Practices is purely descriptive: how methods are being applied, and what terminology is being used. However, during the QUEST meeting we also concluded that most pre-testing methods address the first step in the question-and-answer process. These methods are not satisfactory regarding the investigation of the retrieval and judgement steps. Methods that help us to investigate these steps in more detail, should be developed.

Furthermore, it seems useful to incorporate in our toolbox other research methods, i.e. combining cognitive and non-cognitive methods that investigate the quality of questionnaires. A combination of results from several research approaches will result in even more information on the quality of questionnaires and will help to improve the crafting of survey questions. Here, methods like split-ballot MTMM (multi-trait multi-method) experiments (Saris, 1998; Saris et al., 2002) and interaction analysis (Maynard et al., 2002) can be mentioned.

The last aspect concerning the effectiveness of pre-test methods discussed here is the validity of pre-test research. This aspect deals with the question as posed by Groves in 1996 (p. 401-402): “How do we know what we think they think is really what they think?”, while discussing the usefulness of cognitive research:

1. Is there evidence that a discovered ‘problem’ will exist for all members of the target population? Is evidence sought that different problems exist for groups for whom the questions are more salient, more or less threatening, more or less burdensome?

2 International Conference on Questionnaire Development, Evaluation, and Testing Methods, November 14-17, 2002, Charleston, South Carolina: www.jpsm.umd.edu/qdet (Presser et al., 2004).

2. Do multiple measures of the same component of the question-answer technique discover the same problem (that is, exhibit convergent validity)?
3. When the problem is 'fixed', does replication of the techniques show that the problem has disappeared?
4. When the problem is fixed, does application of other techniques discover any new problem?
5. Is there evidence that the fixed problem produces a question with less measurement error than the original one?

These kinds of questions require experimental designs with contrasts of new and old questions and explicit measures of accuracy. Such studies are common to survey methodology for studies of measurement error and are needed to demonstrate the validity of pretesting techniques. At CASM II, Schwarz (1999, p. 71) also quoted Groves and was surprised to see that "in the light of the extensive applied work done in cognitive laboratories, (…) a systematic evaluation of the practical usefulness of cognitive laboratory procedures is still missing."

Research to assess the validity of pre-testing methods has been done by Fowler (2002, 2003). At the QUEST meeting he discussed the results of split-ballot experiments. Other research in this category has been presented by Rothgeb (2003) at the QUEST meeting. She discussed a vehicle for question testing in a field environment and conducting split-sample field experiments to compare different questionnaire designs: The Questionnaire Design Experimental Research Survey (QDERS). Yet another way to research the accuracy of measurements nowadays, is offered by the use of register data, and confront these data to survey data gathered in such experiments. In my opinion, more research in this field is needed, since it is essential with regard to the aim of pre-test research (as described in section 2, figure 1): Are pre-tested and accordingly revised questionnaires more valid measuring instruments, and do they produce better data, than questionnaires that have not been tested?

3.2. Developing guidelines for questionnaire design

In 1999, Willis et al. raised the following question: "What have we learned in general about questionnaire design, based on the thousands of cognitive interviews that have been conducted, that can be used to inform the crafting of survey questions?" The research that is mentioned in this section is aimed at answering this question.

Since the beginning of CASM and the development of cognitive laboratories, lots of questions have been pre-tested all over the world. This means that lots of situations have been encountered in which the question-and-answer process has been problematic. And

lots of recommendations have been presented to improve the questions. However, these situations and recommendations have not been systematically described. What is needed is a systematic review and description of these situations, findings and recommendations. On the basis of pre-test research *empirically based guidelines for questionnaire design* can be developed.

In the literature on questionnaire design lots of guidelines are presented (see e.g. Dillman, 2000; Czaja & Blair, 1996; Fowler, 1995; Foddy, 1993; Converse & Presser, 1986). However, in my view, they are not a precise enough tool for survey practitioners to develop good questions. And, sometimes guidelines are contradictory to each other. Still, a lot of practice and hands-on experience is needed to craft questionnaires. Thus, questionnaire design still is an art (Payne, 1980).

For instance, a common guideline is that question wording should be simple and as short as possible (Dillman, 2000). On the other hand Fowler (1995, p. 103) states that "a survey question should be worded so that every respondent is answering the same question." And "wording of the questions must constitute a complete and adequate script such that, when interviewers read the question as worded, respondents will be fully prepared to answer the question." However, a common dilemma in questionnaire design is: When to leave the interpretation of the question to the respondent (and have a simple and short question) and when to make it precise (and consequently have a long question containing a definition)? The present guidelines won't help in this situation.

The development of guidelines for questionnaire design should start with the development of a *question database*. This database should include question wordings and meta-information (like the origin of the question, the questionnaire it comes from, pre-test results, recommended improvements) regarding these questions. In stead of designing new questions from scratch all over again, survey practitioners may design questionnaires by selecting questions from this database. And, more information on the measurement instrument becomes available, making meta-data from several pre-test studies comparable, and having more indications on the quality of the questionnaire. Then, questions like "What kind of questions (with what characteristics) in what situations result in what kind of problems in the question-and-answer process, and how should these questions be reworded?" can be answered.

3.3. Adapting pre-test methods to new survey design issues

Traditionally, pre-testing methods are oriented at evaluating questionnaires for face-to-face interviews. However, since interviewer-administered face-to-face surveys are becoming too expensive and cheaper modes are being used more and more, pre-testing

methods for these modes should become available. This includes pre-testing methods for telephone surveys. A new mode that has a lot of attention nowadays is the Internet.

The Internet is a very cheap and easy to use mode. Questionnaires can be easily developed and put on the net. However, like with face-to-face surveys, when questionnaires are not properly developed, the quality of the data can be questioned. Therefore, *pre-testing methods for Web surveys* should be developed. Already Fowler (1995) and Dillman (2000) present ways for pre-testing self-administered questionnaires. Also pre-testing methods can be combined with usability testing (Couper, 2000). At the QUEST workshop in Mannheim Bäckström and Henningsson (2003) discussed this issue. They presented a checklist for designing electronic, self-administered questionnaires.

Apart from adapting pre-testing to new modes, pre-testing methods should also be adapted to 'new' populations. A population that needs more attention in questionnaire design are establishments. Over the years this population has been given attention with regard to questionnaire design and pre-testing (Phipps et al., 1993; Cox et al., 1995; ASA, 2000). However, during the QUEST meeting it became obvious that pre-testing questionnaires for business surveys still needs more attention. A number of papers addressed this issue. Giesen (2003) discussed an extensive program for pre-testing the Dutch Production Survey at statistics Netherlands. Jones (2003) discussed a framework for reviewing data collection instruments in business surveys at the British Office for National statistics. Response burden was discussed by Haraldsen (2003). He presented a conceptual model for response burden in business surveys.

4. More Needed Research

There is one aspect that is not addressed thus far. That is the underlying model of the question-and-answer process. In section 3, it was assumed that the response process is stable in time and stable among different socio-cultural groups in society. These issues will be addressed in this section.

4.1. The time dimension

Assuming that the response process in stable over time, means that once a questionnaire has been pre-tested and adapted according to the findings, it can be used over time. However, as we all know, language changes over time. And consequently, a questionnaire that once was approved to be a good measuring instrument needs improvements within, say, 5 to 10 years.

That the interpretation of questions may change over time, especially is a problem for continuous surveys. One way of dealing with this problem is by saying that the wording of a question Q in year T is the same stimulus in year T+n. However, in that case, although the wording has not changed, question Q may turn out to be a different stimulus, since it will be interpreted differently. Another way of looking at this problem is by making the questions comparable in concept: question Q has to measure concept C, and how can that be operationalised, i.e. how can question Q be reworded in such a way that the same concept is measured in year T+n? Now, what we need to find out is, when question Q does not measure concept C anymore.

To deal with the time dimension, we need *methods to continuously monitor the quality of questionnaires*. As for continuous surveys, and for re-use of questions in a question database (see subsection 3.2), pre-testing once is not enough. As Converse and Presser (1986, p. 51) indicate: "… the meaning of questions can be affected by the context of neighboring questions in the interview." And furthermore, "language constantly changes" making question wordings subject to changing interpretations. Fowler and Cannell (1996) argued that behaviour coding might be used in this way. Also split-ballot experiments can be used in this way. However, more research addressing this issue is needed to develop efficient monitoring methods.

4.2. The sociological dimension

Another aspect that needs to be researched is the way the response process differs among socio-cultural groups. In pre-test studies volunteering respondents are selected in such a way that people with different backgrounds are selected, e.g. with regard to gender, age, level of income, level of education, race, etc. Like with the time dimension, the interpretation of questions may not be the same for different groups in society. However, in general, one questionnaire is crafted for the whole of the sample (within one language group).

This is the 'one-size-fits-all' approach (Snijkers & Luppes, 2002). This approach has been improved by the Tailored Design Method (Dillman, 1978, 2000). Also Brög (2000) has developed a respondent-oriented design for surveys. His starting point is: "The researcher must adjust to the respondent, not the respondent to the researcher." This approach means tailoring the questionnaire.

At the 2003 QUEST meeting, difficulties in the response process and the interviewer-respondent interaction that are encountered while pre-testing a questionnaire with different cultural and racial/ethnic groups, have been discussed by Miller (2003). She concludes that the question-and-answer model should be extended with sociological

factors: “Fully understanding this relationship between socio-cultural phenomena and the response process is vital for (…) survey research occurring within international and multi-cultural contexts.”

5. Conclusions

In sections 3 and 4 a number of research issues have been discussed. However, it may be clear that not al issues can be addressed. So, we need to prioritise: what is most urgent? According to me, the following issues should be given most attention:

1. More split-ballot experiments, on a continuous basis (see subsection 3.1).
2. The development of guidelines for questionnaire design, following from a question database (see subsection 3.2).
3. The development of Current Best Practices, starting with detailed descriptions of pre-test methods, the application of these methods, and terminology (see subsection 3.1).
4. The development of methods to continuously monitor the quality of questionnaires (see subsection 4.1)
5. Adapting pre-test methods to Web surveys, including pre-testing self-administered questionnaires and usability testing (see subsection 3.3).
6. Pre-testing business surveys (see subsection 3.3).

In this list of research issues, attention is given to the following aspects concerning pre-test research:

1. Improvement of the pre-test methods,
2. Improvement of the results of pre-test research (the recommendations),
3. Adapting pre-test methods to new survey design issues.

However, since we cannot control changes in language and society over time, we should tailor questionnaires and continue with pre-testing, pre-testing and pre-testing.

References

ASA, 1999: Designing a Questionnaire. ASA series: What is a Survey? American Statistical Association, Alexandria, VA, www.amstat.org.

ASA, 2000: ICES-II. Proceedings of the Second International Conference on Establishment Surveys. Survey Methods for Businesses, Farms and Institutions. Invited papers. June 17-21, Buffalo, New York. American Statistical Association, Alexandria, VA, www.amstat.org.

Bäckström, H./Henningsson, B., 2003: Testing Web Surveys. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Brög, W., 2000: The New KONTIV Design: A Total Survey Design for Surveys on Mobility Behaviour. In: ASA (American Statistical Association, Alexandria, VA, www.amstat.org), Proceedings of the Second International Conference on Establishment Surveys. Survey Methods for Businesses, Farms and Institutions. Invited papers, pp. 353-360. 17-21 June, Buffalo, New York.

Clark, H./Schober, M., 1992: Asking Questions and Influencing Answers. In: Tanur (ed.), Questions about Questions: Inquiries into the Cognitive Bases of Surveys, pp. 15-48. Russel Sage Foundation, New York.

Czaja, R./Blair, J., 1996: Designing Surveys. A Guide to Decisions and Evaluation. Sage, London.

Converse, J.M./Presser, S., 1986: Survey Questions. Handcrafting the Standardized Questionnaire. Quantitative Applications in the Social Sciences Series, No. 63, Sage, Beverly Hills.

Couper, M.P., 2000: Usability Evaluation of Computer-Assisted Survey Instruments. Social Science Computer Review, Vol. 18, No. 4, pp. 384-396.

Cox, B.G./Binder, D.A./Nanjamma Chinnappa, B./Christianson, A./Colledge, M.J./Kott P.S. (eds.), 1995: Business Survey Methods. Wiley, New York.

DeMaio, T./Mathiowetz, N./Rothgeb, J./Beach, M.E./Durant, S., 1993: Protocol for Pretesting Demographic Surveys at the Census Bureau. Census Bureau Report. U.S. Department of Commerce, Bureau of the Census, Washington, DC.

Dillman, D.A., 1978: Mail and Telephone Surveys: The Total Design Method Wiley, New York.

Dillman, D.A., 2000: Mail and Internet Surveys. The Tailored Design Method. Wiley, New York.

Esposito, J., 2003: A Lexicon of Questionnaire Evaluation Terminology: Concepts and Working Definitions. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Foddy, W., 1993, Constructing Questions for Interviews and Questionnaires. Theory and Practice in Social Research. Cambridge University Press, Cambridge.

Fowler, F.J., 1995: Improving Survey Questions. Design and Evaluation. Applied Social Research Methods Series, Vol. 38, Sage, London.

Fowler, F.J., 2002: Getting Beyond Pretests and Cognitive Interviewing: The Case for More Split-Ballot Pilot Studies. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), 14-17 November, 2002, Charleston, South Carolina.

Fowler, F.J., 2003: More on the Value of Split Ballots. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Fowler, F.J./Cannell, Ch.F., 1996: Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. In: Schwarz & Sudman (eds.), Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research, pp. 15-36. Jossey-Bass, San Francisco.

Gerber, E., 2002: Cognitive Interviewing. Short course presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), 14-17 November, 2002, Charleston, South Carolina.

Giesen, D., 2003: Evaluation Plan for the Dutch Structural Business Statistics Questionnaire: Using Output to Guide Input Improvements. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Groves, R.M., 1996: How do we know what we think they think is really what they think? In: Schwarz & Sudman (eds.), Answering Question. Methodology for Determining Cognitive and Communicative Processes in Survey Research, pp. 389-402. Jossey-Bass, San Francisco.

Haraldsen, G., 2003: Searching for Response Burden in Focus Groups with Business Respondents. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Hippler, H.J./Schwarz, N./Sudman, S. (eds.), 1987: Social Information Processing and Survey Methodology. Springer-Verlag, New York.

Jabine, T.B./Straf, M.L./Tanur, J.M./Tourangeau, R. (eds.), 1984: Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines. Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. National Academy Press, Washington, DC.

Jobe, J.B./Mingay, D.J., 1991: Cognition and Survey Measurement: History and Overview. Applied Cognitive Psychology, Special Issue on Cognition and Survey Measurement, Vol. 5, No. 3, pp. 175-192.

Jones, J., 2003: Improving Business Survey Data Collection Instruments: Governance and Methodologies. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Maynard, D.W./Houtkoop-Steenstra, H./Schaeffer, N.C./van der Zouwen, J, 2002: Standardization and Tacit Knowledge. Interaction and Practice in the Survey Interview. (Wiley, New York.)

Mathiowetz, N.A., 2002: Behavior Coding: Tool for Questionnaire Evaluation. Short course presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), 14-17 November, 2002, Charleston, South Carolina.

Miller, K., 2003: Implications of Socio-Cultural Factors in the Question Response Process. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Payne, S.L. 1980 (originally published in 1951): The Art of asking Questions. Princeton University Press, Princeton, New Jersey.

Phipps, P.A./Butani, S./Chun, Y.I., 1993: Designing Establishment Survey Questionnaires. BLS Statistical Notes No. 35. U.S. Bureau of Labor Statistics, Washington, DC.

Presser, S./Rothgeb, J./Couper, M./Lessler, J./Martin, E./Martin, J./Singer E. (eds.), 2004: (forthcoming), Methods for Testing and Evaluating Survey Questionnaires. Wiley, New York.

Rothgeb, J., 2003: A Valuable Vehicle for Question Testing in a Field Environment: The U.S. Census Bureau's Questionnaire Design Experimental Research Survey. Paper presented at the 2003 QUEST workshop, 21-23 October, ZUMA, Mannheim, Germany.

Saris, W.E., 1998: The split-ballot MTMM experiment: An alternative way to evaluate the quality of questions. Research paper, University of Amsterdam, Amsterdam.

Saris, W.E./van der Veld, W./Gallhofer, I./Corten, I., 2002: A Scientific Approach to Questionnaire Development. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), 14-17 November, 2002, Charleston, South Carolina.

Schwarz, N., 1999, Cognitive Research into Survey Measurement: Its Influence on Survey Methodology and Cognitive Theory. In: Sirken et al. (eds.), Cognition and Survey Research, pp. 65-75. Wiley, New York.

Sirken, M.G./Herrmann, D.J./Schechter, S./Schwarz, N./Tanur, J.M./Tourangeau R. (eds.), 1999, Cognition and Survey Research. Wiley, New York.

Snijkers, G., 2002, Cognitive Laboratory Experience: On Pre-testing Computerised Questionnaires and Data Quality. Ph.D thesis. Utrecht University, Utrecht, and Statistics Netherlands, Heerlen.

Snijkers, G./Luppes, M., 2000, The Best of Two Worlds: Total Design Method and New Kontiv Design. An operational Model to improve Respondent Co-operation. In: Snijkers, Cognitive Laboratory Experience: On Pre-testing Computerised Questionnaires and Data Quality, chapter 9. Ph.D thesis. Utrecht University, Utrecht, and Statistics Netherlands, Heerlen.

Tourangeau, R./Rasinksi, K.A., 1988: Cognitive Processes underlying Context Effects in Attitude Measurement. Psychological Bulletin, Vol. 103, No. 3, pp. 299-314.

Willis, G. B./DeMaio, Th. J./Harris-Kojetin, B., 1999: Is the Bandwagon headed to the Methodological Promised Land? Evaluating the Validity of Cognitive interviewing Techniques. In: Sirken et al. (eds.), Cognition and Survey Research, pp. 133-153. Wiley, New York.

Willis, G.,/Forsyth, B., 2002: Methods for Questionnaire Appraisal and Expert Review. Short course presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), 14-17 November, 2002, Charleston, South Carolina.

Contact

Ger Snijkers
Statistics Netherlands
NL - Heerlen
The Netherlands
email: GSKS@CBS.nl

SUMMARY OF WRAP UP DISCUSSION

FLOYD JACKSON FOWLER, JR.

The final session of the workshop was devoted to two topics: what is the state of our current knowledge about how to evaluate questions and what are the priorities for needed research? The following is a brief summary of the conclusions from that session.

State of current knowledge

Although there is still much that we need to learn, progress has been made since our last meeting in documenting the value of question testing and in producing empirically-based generalizations about how to conduct testing.

1. The study conducted by Forsyth, Rothgeb, and Willis (2004) clearly demonstrated that pretesting does identify real question problems. Moreover, Fowler (2004) reports data that strongly support the position that problems identified in cognitive testing can have important effects on the data, the quantitative estimates made from surveys.

2. Expert reviews, cognitive testing, and field pretests with behavior coding all have contributions to make to the evaluation of questions. No one method is adequate to identify all of the various kinds of problems that questions can have. The fact that each method has the ability to identify some kinds of problems and not others should not be considered "flaws" in the method. Rather, we need to recognize that the approaches are complementary and that multiple methods should be used to comprehensively evaluate questions.

3. Evaluating the usability of survey instruments is as important as evaluating the question wording. The best techniques for evaluating usability and question wording are different, again leading to the conclusion that multiple testing approaches are needed.

4. With respect to the particular techniques for doing cognitive testing, DeMaio and Landreth's research (2004) supports the following conclusions:

    a. Listening to tapes of cognitive interviews to help prepare reports of the results increases the number of issues identified and almost certainly increases the value of the cognitive testing.

b. Cognitive interviews are done by researchers with Ph.D.s, specialists in question design, and regular interviewers with special training. There is no evidence that interviewers with any one of these backgrounds is consistently superior to the others.

c. Cognitive interviews can rely on think-aloud techniques and probes that are asked after questions have been answered. Sometimes probes are scripted; at other times the interviewers create their own probes or use a combination of scripted and ad hoc probes. There is no evidence that any of these variations is consistently superior to the others.

Needed Research

The list of things that are known about question evaluation is shorter than the list of areas in which further research is needed. There are many aspects of how to test questions about which we need to learn more:

1. There is much to learn about how to use each of the methods of question evaluation that were discussed:

   a. We need descriptive studies of how organizations are currently doing the various kinds of testing. We know there is variation, but better documentation of the variation that currently exists would help to define a research agenda for how best to study evaluation techniques.

   b. We also need to do more studies documenting the strengths and limitations of the various approaches to evaluating questions and survey instruments.

   c. For cognitive testing, we need critical evaluation of the various kinds of probes that are used, to learn which are the best and most productive

   d. We also need more research on how best to use the results from cognitive testing to improve questions and data quality: How interviewers should record and report what happens (audiotapes, videotapes, write ups of individual interviews, summaries across several interviews); who (i.e. people with what sort of credentials or role in the survey) are the best people to view or listen to the actual interviews; what is the best process for taking the results of the cognitive interviews and using them to revise question wording?

   e. Behavior coding has concentrated on four or five behaviors (question reading, probing, requests for clarification, interruptions, and inadequate answers) to provide evidence of question problems. Research to identify other significant behaviors and to document the relationships between behaviors and undesirable question features will add to the value of behavior coding.

   f. Expert reviewers need better empirically-based generalizations about how (and how much) observable question characteristics affect usability and data quality.

2. There is good reason to think that testing techniques may need to be adapted to the requirements of certain kinds of surveys and populations. For example:
   a. Think-aloud interviews clearly seem to work better with some respondents than others
   b. There is anecdotal reason to think that the best cognitive probes may vary from respondent to respondent. We need studies to document and understand those relationships
   c. The best approaches to testing questions for surveys of individuals may differ from the best approaches to testing questions and instruments for establishment surveys.
   d. We have much to learn about how mode of administration affects the best way to evaluate questions and instruments. This is not just a matter of making sure that usability is evaluated. It is specifically addressing the fact that the way a question is understood and answered may vary by whether the respondent hears it on the telephone, reads it from a written page or reads it on a computer screen. Testing methods must be adapted to detect the problems that are specific to the mode of administration. At the moment, we do not have generalizations about how to adapt our testing to the mode of administration.
   e. Usability testing has emerged as one of the most underdeveloped aspects of survey instrument evaluation. Its importance has risen because computer-based data collection is on the rise. It is understood that respondent willingness to complete surveys that are self administered, on paper or on a computer, is affected by how easy it is to do. More studies are needed about how to efficiently and effectively identify problems that real respondents will have when they try to complete a survey.
3. To do these studies, we need to develop some new tools.
   a. One of the major barriers to studying question evaluation strategies is the absence of good measures of success. The studies to date have largely counted and compared the number of "problems" identified by testing, with no real way of knowing whether a “problem” actually had any effect on the survey estimates. Split-ballot tests are one approach to assessing the effects of alternative question wording on data, though knowing which question is “best” depends on having a strong theory or independent data if results are different. An even better design is to have measures in the survey instrument or external data to directly assess the validity of the answers to alternative questions. To date such validation studies have been rare, but we need them to evaluate the effectiveness of our testing and the questions changes that they produce.
   b. We need a better typology of question problems – one specifically geared to question evaluation techniques. Tourangeau’s (1984) sorting of issues into comprehension, retrieval, coding and answering was a start. However, we need a

more refined system that would enable researchers to match the kind of problem they are potentially concerned about with a particular approach to testing or probing in a cognitive interview.

c. We also need a typology of usability problems. Following instructions, seeing and understanding all the relevant parts of a question, knowing how to use computer aids, and simply knowing how to navigate through a set of questions are all part of usability. We need a parsimonious and useful list of usability issues, so we can be sure we are using testing procedures that will identify problems of the various types that matter.

d. The importance of usability needs to be assessed. It is a reasonable hypothesis that problems with usability adversely affect comprehension directly, as well as affecting willingness or ability to answer questions (leading to item nonresponse). However, those relationships have not been documented.

e. Finally, we need a better set of generalizations about features of questions that adversely affect measurement. Question testing identifies question or instrument features we think are problems. However, if we have stronger, evidence-based generalizations about what question features to avoid, we would have fewer bad questions to start with and the effectiveness of improving questions that were found wanting via testing would be greatly enhanced.

References

Forsyth, B., Rothgeb, J., and Willis, G., 2004: Does Pretesting Make a Difference? in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

Fowler, F. J., 2004: Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies, in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

DeMaio, T., and Landreth, A., 2004: Do Different Cognitive Interview Methods Produce Different Results? in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

Tourangeau, R., 1984: Cognitive Science and Survey Methods. pp. 73-100 in: T. Jabine/ M. Straf/J. Tanner/R. Tourangeau (eds.) Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines Washington: National Academy Press.

Contact

Floyd Jackson Fowler, Jr.
Center for Survey Research
University of Massachusetts, Boston
100 Morrissey Blvd.
Boston, MA 02125 -USA
email: floyd.fowler@umb.edu

Agenda for the QUEST 2003 Conference

October 21-23, 2003

ZUMA (Centre for Survey Research and Methodology)

B2,1

D-68072 Mannheim

Phone: 0049 (0)621 1246-0

Tuesday, October 21, 2003

9:15 – 9:20 a.m.	Welcome	Margrit Rexroth, ZUMA, Germany
9:20 – 9:30 a.m.	Opening Remarks	Ingwer Borg, Programme Director of ZUMA, Germany
9:30 – 10:00 a.m.	Introductions of Attendees	Peter Prüfer, ZUMA, Germany
10:00 – 10:30 a.m.	General Remarks about QUEST, Objectives of QUEST 2003	Jack Fowler Center for Survey Research, University of Massachusetts, Boston, U.S.A.
10:30 – 10:45 a.m.	Break	
10:45 – 11:15 a.m.	Computer Assisted Pretesting of CATI – Questionnaires	Frank Faulbaum, Department of Empirical Social Research, University of Duisburg-Essen, Germany
11:15 – 11:30 a.m.	Discussion of Invited Paper	QUEST Participants
11:30 – 11:50 a.m.	Concepts and Terminology	James L. Esposito Bureau of Labor Statistics, U.S.A.
11:50 – 12:30 a.m.	Discussion about “Concepts and Terminology”	QUEST Participants
12:30 – 2:00 p.m.	Lunch at ZUMA Foyer, B2,15	

Tuesday, October 21, 2003 (Continued)

	Topic 1: Current State of Knowledge	
2:00 – 3:45 p.m.	**Paper Session I** (15 min. each)	**Moderator: Deirdre Giesen** **Statistics Netherlands**
	Cognitive Model of the Question-Answering Process and Development of Pretesting	Anja Ahola Statistics Finland
	Discussion (10 minutes)	
	Our Experience from Measurement Tests in Developing Countries	Gunilla Davidsson and Birgit Henningsson Statistics Sweden
	Discussion (10 minutes)	
	New Pretesting Methods at Statistics New Zealand (tentative) Discussion (10 minutes)	Denise Grealish Statistics New Zealand
	I Can't See You: Conducting Cognitive Interviews Over the Telephone Discussion (10 minutes)	Carol Cosenza Center for Survey Research, University of Massachusetts, Boston, U.S.A.
3:45 – 4:00 p.m.	Break	
	Topic 2: Methods and Modes	
4:00 – 5:15 p.m.	**Paper Session II** (15 min. each)	**Moderator: Jennifer Rothgeb** **U.S. Census Bureau, Washington**
	Testing Survey Questionnaires: Recent Experiences and Developments	Paul Kelly Statistics Canada
	Discussion (10 minutes)	
	More about Split Ballots Discussion (10 minutes)	Jack Fowler Center for Survey Research, University of Massachusetts, Boston, U.S.A.
	Paraphrasing can be dangerous – sometimes	Peter Prüfer and Margrit Rexroth ZUMA, Germany
	Discussion (10 minutes)	
5:15 p.m.	Adjourn	
6:30 p.m.	Complementary Dinner hosted by ZUMA	We have organized a group bus to a restaurant in the Palatinate, famous for its wine and food. Meeting point: ZUMA Foyer, B2,15

Wednesday, October 22, 2003

	Topic 2: Methods and Modes	
9:30 – 11:10 a.m.	**Paper Session III** (15 min. each)	**Moderator: Debbie Collins National Center for Social Research, U.K.**
	Testing Web Surveys Discussion (10 minutes)	Helena Bäckström and Birgit Henningsson Statistics Sweden
	Exploring the Data in an Experiment of Alternative Cognitive Interviewing Methods Discussion (10 minutes)	Terry DeMaio and Ashley Landreth Center for Survey Methods Research, U.S. Census Bureau
	Questionnaire Evaluation of the Dutch Production Surveys: Using Output to Guide Input Improvements	Deirdre Giesen Statistics Netherland
	Discussion (10 minutes)	
	The Use of Internal Administrative Sources in the Design & Evaluation of Data Collection Instruments	Jacqui Jones Office for National Statistics, U.K.
	Discussion (10 minutes)	
11:10 – 11:25 a.m.	Break	
11:25 – 12:45 a.m.	**Paper Session IV** (15 min. each)	**Moderator: Birgit Henningsson Statistics Sweden**
	A Valuable Vehicle for Question Testing in a Field Environment – The Census Bureau's Questionnaire Design Experimental Research Survey (QDERS)	Jennifer Rothgeb U.S. Census Bureau
	Discussion (15 minutes)	
	Testing Automatic Coding Questions in the Field Discussion (15 minutes)	Rachel Vis Statistics Netherlands
	Searching for Response Burdens in Focus Groups with Business Respondents	Gustav Haraldsen Statistics Norway
	Discussion (15 minutes)	
12:45 – 2:00 p.m.	Lunch at a restaurant, near ZUMA	

Wednesday, October 22, 2003 (Continued)

	Topic 3: The Computer as a Questionnaires Evaluation Tool	
2:00 – 3:00 p.m.	**Paper Session V** (15 min. each)	**Moderator: Ger Snijkers** **Utrecht University**
	Using the Computer as Recording Device in CASI-Survey Cognitive Testing	Øyvind Brekke Statistics Norway
	Discussion (15 minutes)	
	The Observer, a Software and Hardware Configuration for Coding and Analysis of Behaviour	Dirkjan Beukenhorst Statistics Netherland
	Discussion (15 minutes)	
3:00 – 3:15 p.m.	Break	
	Topic 4: Theoretical Advances in Questionnaire Design and Evaluation	
3:15 – 5:15 p.m.	**Paper Session VI** (15 min. each)	**Moderator: Gustav Haraldsen** **Statistics Norway**
	Theoretical Advances in Questionnaire Design and Evaluation	Debbie Collins National Centre for Social Research, U.K.
	Discussion (15 minutes)	
	With Regard to the Design of Major Statistical Surveys: Are We Waiting Too Long to Evaluate Substantive Questionnaire Content?	James L. Esposito Bureau of Labor Statistics, U.S.A.
	Discussion (15 minutes)	
	Integrating Socio-Cultural Factors into the Question Response Model Discussion (15 minutes)	Kristen Miller National Center for Health Statistics, U.S.A.
	Paradigms of Cognitive Interviewing Practice, and Their Implications for Developing Standards of Best Practices	Paul Beatty National Center for Health Statistics, U.S.A.
	Discussion (15 minutes)	
5:15 p.m.	Adjourn, Free Time for Personal Use	

Thursday, October 23, 2003

	Final Session	**Moderator: Jack Fowler Center for Survey Research, University of Massachusetts, Boston, U.S.A.**
9:30 – 9:45 a.m.	Summary Remarks	Jack Fowler
9:45 – 10:30 a.m.	Reflection of the Current State of Knowledge about how to Evaluate Survey Questions	Jack Fowler
	Discussion	
10:30 – 10:45 a.m.	Break	
10:45 – 11:45 a.m.	What are the Highest Priorities for Additional Research on Question Evaluation?	Jack Fowler
	Discussion	
11:45 – 12:15 a.m.	Preliminary Plans for QUEST	Jack Fowler
12:30 a.m.	Little Lunch at ZUMA Foyer, B2,15	
Adjourn		
1:30 p.m.	Planning Committee Meeting at ZUMA	**Moderator: James L. Esposito Bureau of Labor Statistics, U.S.A.**

LIST OF CONTRIBUTORS

Ahola, Anja
Statistics Finland
Social Statistics
FIN – 00022 Statistics Finland
email: anja.ahola@stat.fi

Bäckström, Helena
Statistics Sweden
Box 24 300
SE – 104 51 Stockholm
Sweden
email: helena.backstrom@scb.se

Beatty, Paul
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782
U.S.A.
email: pbb5@cdc.gov

Beukenhorst, Dirkjan
Statistics Netherlands
CBS
Postbus 4481
NL – 6401 CZ Heerlen
The Netherlands
email: dbkt@cbs.nl

Cosenza, Carol
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
U.S.A.
email: Carol.Cosenza@umb.edu

Davidsson, Gunilla
SCB Statistiska centralbyrån
Statistics Sweden
Box 24 300
SE – 104 51 Stockholm
Sweden

email: gunilla.davidsson@scb.se

DeMaio, Terry
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: Theresa.J.DeMaio@census.gov

Esposito, James L.
Bureau of Labor Statistics
Postal Square Building
Washington, DC 20212
U.S.A.
email: Esposito.Jim@bls.gov

Faulbaum, Frank
Lehrstuhl für Sozialwissenschaftliche Methoden/Empirische Sozialforschung
Sozialwissenschaftliches Umfragezentrum
Universität Duisburg-Essen, Standort Duisburg
Lotharstraße 65
D – 47048 Duisburg
email: faulbaum@uni-duisburg.de

Fowler, Floyd Jackson Jr.
Center for Survey Research
University of Massachusetts, Boston
100 Morrissey Blvd.
Boston, MA 02125
U.S.A.
email: floyd.fowler@umb.edu

Giesen, Deirdre
Statistics Netherlands
Kloosterweg 1
NL – 6412 CN Heerlen
The Netherlands
email: igin@cbs.nl

Grealish, Denise
Statistics New Zealand
Questionnaire Design Consultancy
PO Box 2922
Wellington
New Zealand
email: denise.grealish@stats.govt.nz

Haraldsen, Gustav
Statistics Norway
P.P. box 8131 Dep
N – 0033 Oslo
Norway
email: GHa@ssb.no

Henningsson, Birgit
Statistics Sweden
Research and Development Methodology
SE – 701 89 Örebro
Sweden
email: birgit.henningsson@scb.se

Jones, Jacqui
Office for National Statistics
Room D136
Government Buildings
Cardiff Road
UK – Newport
Wales
United Kingdom
email: Jacqui.Jones@ons.gsi.gov.uk

Landreth, Ashley,
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: ashley.denele.landreth@census.gov

Miller, Kristen
National Center for Health Statistics
Off. of Research & Methodology
3311 Toledo Road
Hyattsville, MD 20782
U.S.A.
email: ktm8@cdc.gov

Prüfer, Peter
ZUMA
Postfach 12 21 55
D – 68072 Mannheim
email: pruefer@zuma-mannheim.de

Rexroth, Margrit
ZUMA
Postfach 12 21 55
D – 68072 Mannheim
email: rexroth@zuma-mannheim.de

Rothgeb, Jennifer M.
U.S. Census Bureau
SRD/Center for Survey Methods Research
Washington, DC 20233-9100
U.S.A.
email: jennifer.m.rothgeb@census.gov

Snijkers, Ger
Statistics Netherlands
Postbus 4481
NL – 6401 CZ Heerlen
The Netherlands
email: GSKS@CBS.nl

Vis, Rachel
Statistics Netherlands
Kloosterweg 1
NL – 6412 CZ Heerlen
The Netherlands
email: RVCS@cbs.nl

ZUMA-Nachrichten Spezial

Die Reihe ZUMA-Nachrichten-Spezial dient dazu, den Forschungsstand größerer Arbeits- oder Forschungsbereiche bei ZUMA zu dokumentieren oder die Ergebnisse von Konferenzen und Symposien vorzustellen (http://www.gesis.org/publikationen/Publikationen/zeitschriften/ZUMA_Nachrichten_spezial/). Bisher sind acht Bände erschienen.

* * *

ZUMA- Nachrichten Spezial Band 1

Text Analysis and Computers

Hrsg. von Cornelia Züll, Janet Harkness und Jürgen H.P. Hoffmeyer- Zlotnik

Mannheim, ZUMA, 1996, 132 Seiten, ISBN 3-924220-11-5

Das Heft entstand im Anschluß an eine internationale Tagung zur computerunterstützten Textanalyse, bei der sich Wissenschaftler aus den verschiedensten Disziplinen trafen. Die hier abgedruckten Papiere der eingeladenen Hauptredner dokumentieren den Forschungsstand in vier Bereichen: Computer-Assisted Content Analysis: An Overview *(E. Mergenthaler*); Computer-Aided Qualitative Data Analysis: An Overview (*U. Kelle*); Machine-Readable Text Corpora and the Linguistic Description of Language (*Chr. Mair*); Principle of Content Analysis for Information *Retrieval (J. Krause*). Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 2 (vergriffen)

Eurobarometer. Measurement Instruments for Opinions in Europe

Hrsg. von Willem E. Saris und Max Kaase

Mannheim: ZUMA 1997, ISBN 3-924220-12-3

In der Empirischen Sozialforschung finden in Europa Telefoninterviews anstelle von face to face-Interviews zunehmende Verbreitung. Im Rahmen der zweimal jährlich für die Europäische Kommission in Brüssel durchgeführten Repräsentativbefragungen in den Mitgliedsländern der Europäischen Union, den sogenannten Eurobarometern, ergab sich für die Erhebung vom Frühjahr 1994 (EB 41.0) die Möglichkeit, durch eine zeitgleich mit einem weitgehend identischen Fragenprogramm stattfindende Telefonbefragung in den damaligen zwölf Mitgliedsländern der EU, systematisch Effekte der unterschiedlichen Stichprobenansätze und Erhebungsmethoden zu untersuchen. Dabei konnte das Analysespektrum noch durch eine Telefon-Panelkomponente in dreien der zwölf EU-Länder für das face to face-Eurobarometer erweitert werden. Die Beiträge im vorliegenden Buch untersuchen auf dieser Grundlage methodische und methodologische Fragestellungen, die insbesondere für die

international vergleichende Sozialforschung, aber auch für die Markt- und Meinungsforschung in Europa von großer Bedeutung sind. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

*** * ***

ZUMA-Nachrichten Spezial Band 3
Cross-Cultural Survey Equivalence.
Hrsg. von J. Harkness
Mannheim: ZUMA 1998, 187 Seiten, ISBN 3-924220-13-1

This volume, the third in the ZUMA-Nachrichten-Spezial series on methodological issues in empirical social science research, is devoted to issues of cross-cultural methodology. The focus is on issues of equivalence, the key requirement in cross-national and cross-cultural comparative research. As the contributions indicate, equivalence is, however, better thought of in terms of equivalencies - in social science surveys and in other standardised instruments of measurement. Contributors come from different countries and continents and from widely differing research backgrounds, ranging from linguistics to survey research and ist methodologies, to cultural anthropology and cross-cultural psychology. They are: Timothy P. Johnson, Fons J.R. van de Vijver, Willem E. Saris, Janet A. Harkness and Alicia Schoua-Glusberg, Michael Braun and Jacqueline Scott, Ingwer Borg: Peter Ph. Mohler, Tom W. Smith and Janet A. Harkness.

*** * ***

ZUMA-Nachrichten Spezial Band 4 (vergriffen)
Nonresponse in Survey Research
Hrsg. von A. Koch und R. Porst
Mannheim: ZUMA 1998, 354 Seiten, ISBN 3-924220-15-8

This volume, the fourth in the ZUMA-Nachrichten Spezial series on methodological issues in empirical social science research, takes up issues of nonresponse. Nonresponse, that is, the failure to obtain measurements from all targeted members of a survey sample, is a problem which confronts many survey organizations in different parts of the world. The papers in this volume discuss nonresponse from different perspectives: they describe efforts undertaken for individual surveys and procedures employed in different countries to deal with nonresponse, analyses of the role of interviewers, the use of advance letters, incentives, etc. to reduce nonresponse rates, analyses of the correlates and consequences of nonresponse, and descriptions of post-survey statistical adjustments to compensate for nonresponse. All the contributions are based on presentations made at the ‘8th International Workshop on

Household Survey Nonresponse'. The workshop took place in September 1997 in Mannheim, Germany, the home base of the workshop host institute, ZUMA. Twenty-nine papers were presented and discussed, of which twenty-five are included here.

*** * ***

ZUMA-Nachrichten Spezial Band 5
A review of software for text analysis
Alexa Melina & Cornelia Zuell
Mannheim: ZUMA 1999, 176 Seiten, ISBN 3-924220-16-6

The book reviews a selection of software for computer-assisted text analysis. The primary aim is to provide a detailed account of the spectrum of available text analysis software and catalogue the kinds of support the selected software offers to the user. A related, more general, goal is to record the tendencies both in functionality and technology and identify the areas where more development is needed. For this reason the presented selection of software comprises not only fully developed commercial and research programs, but also prototypes and beta versions. An additional aspect with regards to the kinds of software reviewed is that both qualitative and quantitative-oriented types of research are included. Depending on research purposes and project design the text analyst can profit from available tools independently of their orientation. The following fifteen programs are reviewed: AQUAD, ATLAS.ti, CoAN, Code-A-Text, DICTION, DIMAP-MCCA, HyperRESEARCH, KEDS, NUD*IST, QED, TATOE, TEXTPACK, TextSmart, WinMAXpro, and WordStat and the criteria and methodology used for selecting them are delineated. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/ zuma_nachrichten_spezial/).

*** * ***

ZUMA-Nachrichten Spezial Band 6
Sozialstrukturanalysen mit dem Mikrozensus
Hrsg. von Paul Lüttinger
Mannheim: ZUMA 1999, 402 Seiten, ISBN 3-924220-17-4

Im Oktober 1998 veranstaltete die Abteilung Mikrodaten von ZUMA die Konferenz "Forschung mit dem Mikrozensus: Analysen zur Sozialstruktur und zum Arbeitsmarkt", an der vorwiegend Nutzer des Mikrozensus teilahmen. Hauptziel dieser ersten Nutzerkonferenz war es, ein Forum für den Informationsaustausch zwischen den Datennutzern und den statistischen Ämtern zu schaffen. Die mehr als 20 Vorträge gingen deutlich über die von den statistischen Ämtern veröffentlichten Standardergebnisse zum Mikrozensus hinaus und sind weitgehend in diesem Band ZUMA-Nachrichten Spezial abgedruckt. Die Autoren sind: Walter Müller; Karl Brenke; Esther Hansch und Michael-Burkhard Piorkowski; Friedhelm

Pfeiffer; Jürgen Schupp, Joachim Frick, Lutz Kaiser und Gert Wagner; Elke Wolf; Dietmar Dathe; Bernd Eggen; Erich Stutzer; Carsten Baumann; Susanne von Below; Thomas Bulmahn; Martin Groß; Reiner H. Dinkel, Marc Luy und Uwe Lebok sowie Wolfgang Strengmannn-Kuhn. Der Band ist als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 7

Social and Economic Analyses of Consumer Panel Data
Georgios Papastefanou, Peter Schmidt, Axel Börsch-Supan, Hartmut Lüdtke, Ulrich Oltersdorf (Eds.)
Mannheim: ZUMA 2001; 212 Seiten; CD-Rom

Eine von der Abteilung Einkommen und Verbrauch von ZUMA organisierte Arbeitsgruppe hat sich mit datentechnischem Handling und Analysepotential von komplexen Verbraucherpaneldaten, am Beispiel des ConsumerScan Haushaltspanels der Gesellschaft für Marktforschung (GfK, Nürnberg) beschäftigt und die Ergebnisse in einem Symposium im Oktober 1999 vorgestellt. Die überwiegende Zahl der vorgetragenen Arbeiten, die man als Werkstattberichte ansehen kann, sind in diesem Band abgedruckt. Neben einem detallierten Einblick in die Praxis und das Datenerhebungsprogramm von Verbraucherpanels, wie sie z.B. bei der Marktforschungen der GfK unterhalten werden, enthält der Band z.B. Untersuchungen zu Fragen der Flexibilität von Preisbildungsvorgängen, des Lebensstils im alltäglichen Konsums, der Gesundheitsorientierung im Konsumverhalten, der Umweltorientierung und ihrer Umsetzung im Kauf alltäglicher Haushaltungsprodukte. Der Band enthält eine CD-ROM mit Dokumenten und Codebüchern der aufbereiteten ZUMA-Verbraucherpaneldaten 1995. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/ publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 8

Von Generation zu Generation
Hrsg. von Jan van Deth
Mannheim: ZUMA 2002, 68 Seiten, ISBN 3-924220-23-9

Aus Anlass der Ehrung von Prof. Dr. Max Kaase, Prof. Dr. Walter Müller und Prof. Dr. Hansgert Peisert für ihre langjährige und richtungsweisende Mitarbeit in der Mitgliederversammlung des ZUMA e.V. fand am 14. Juni 2002 eine wissenschaftliche Tagung statt. Der Band enthält Beiträge von Jan van Deth, Hubert Feger, Jürgen Rost, Erwin K. Scheuch, Andreas Diekman und Hans-Dieter Klingemann. Die Beiträge sind auch online verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/.)

ZUMA-Nachrichten Spezial Band 9

QUEST 2003

Questionnaire Evaluation Standards

Peter Prüfer, Margrit Rexroth, Floyd Jackson Fowler, Jr. (Eds.)

Mannheim: ZUMA 2004, 216 Seiten, ISBN 3-924220-27-1

This volume, the ninth in the ZUMA-Nachrichten Spezial series on methodological issues in empirical social science research takes up issues of question and questionnaire evaluation. The papers in this volume discuss practical as well as theoretical aspects of questionnaire evaluation. All contributions are based on presentations made at the fourth QUEST (Questionnaire Evaluation Standards) conference which took place from October 21 - 23, 2003 at ZUMA in Mannheim. There were 26 attendees from 9 countries representing 14 organizations: Bureau of Labor Statistics, USA, Center for Survey Research, University of Massachusetts, USA, Institut für Demoskopie Allensbach, Germany, National Center for Health Statistics, USA, National Center for Social Rerearch, U.K., Office of National Statistics, U.K., Statistics Canada, Statistics Finland, Statistics Netherlands, Statistics New Zealand, Statistics Norway, Statistics Sweden, U.S. Census Bureau, ZUMA, Germany. This volume can be downloaded as a PDF file (http://www.gesis.org/publikationen/ zuma_nachrichten_spezial/).

DURCHWAHL-RUFNUMMERN (STAND: MÄRZ 2004)

Sie erreichen die Mitarbeiter von ZUMA unter der Nummer 0621-1246-(Durchwahlnummer); die Zentrale unter 1246-0. Sie ist von Montag bis Donnerstag von 8.30 bis 17.00 und freitags von 8.30 bis 15.30 besetzt. Die mit (S) bezeichneten Mitarbeiterinnen nehmen Sekretariatsaufgaben wahr.

DIREKTION

Prof. Dr. Peter Ph. Mohler (Direktor) 173
Carol Cassidy (Stellv. Direktorin) 146
Margit Bäck (S) 172
Elisabeth Bähr (S) 172
Maria Kreppe-Aygün 184

INTERNE INFRASTRUKTUR

Verwaltung
Dipl.-Kfm. Jost Henze 161

Information & Kommunikation
Dipl.-Soz. Kerstin Hollerbach 174

EDV-Infrastruktur
Carol Cassidy 146

Datenbanken
Joachim Wackerow 262

WISSENSVERMITTLUNG & BERATUNG

Wissenschaftlicher Leiter
Prof. Dr. Ingwer Borg 151

Projektberater
Dr. Wolfgang Bandilla 136
Dr. Michael Braun 176
PD Dr. Jürgen H.P. Hoffmeyer-Zlotnik 175
Dipl. Soz. Rolf Porst 228
Dr. Beatrice Rammstedt 155
Christa v. Briel (S) 231
Dagmar Haas (S) 152
Patricia Lüder (S) 221

Pretesting
Dipl.-Psych. Peter Prüfer 227
Margrit Rexroth, M.A. 230

Textanalyse, Vercodung
Alfons J. Geis, M.A. 222
Patricia Lüder (S) 221

Statistik
PD Dr. Siegfried Gabler 281
Dr. Sabine Häder 282
Dipl.-Math. Michael Wiedenbeck 279

Telefonumfragen
Dipl.-Soz. Michael Schneid 209
Dipl.-Soz. Angelika Stiegler 208

Online-Umfragen
Dipl.-Sozialw. Wolfgang Neubarth 205

Elektronische Handbücher ZIS/EHES
Dr. Angelika Glöckner-Rist 171

Computerunterstützte Textanalyse, Textpack, NSD-stat
Cornelia Züll 147
Juliane Landmann 144

DAUERBEOBACHTUNG

Wissenschaftliche Leiterin
Prof. Dr. Ursula Hoffmann-Lange 247

ALLBUS
Dipl.-Soz. Achim Koch *280*
Dipl. Soz. Michael Blohm 276
Dipl. Soz. Alexander Haarmann 273
Dipl.-Soz. Martina Wasmer 273
Julia Khorshed (S) 274

International Social Survey Programme (ISSP)
Dr. Janet Harkness 284
Sabine Klein 272
Evi Scholz 283

German Micro Data Lab
Prof. Dr. Ursula Hoffmann-Lange *247*
Dipl.-Soz. Bernhard Schimpl-Neimanns 263
Dipl.-Soz. Jeanette Bohr 261
Dipl.-Sozialwiss. Helga Christians 266
Dipl. Soz. Nadia Granato 264
Dr. Annette Kohlmann 253
Dr. Paul Lüttinger 268
Dr. Georgios Papastefanou 278
Dr. Heike Wirth 269
Irene Fischer (S) 265

METHODENFORSCHUNG & ENTWICKLUNG

Wissenschaftlicher Leiter
N.N.

Soziale Indikatoren
Dr. Heinz-Herbert Noll 241
Dipl.-Soz. Regina Berger-Schmitt 248
Dr. Caroline Kramer 244
Dr. Stefan Weick 245
Margit Bäck (S) 242

DRITTMITTELPROJEKTE

Nina Rother (PIONEUR) 285
Lars Kaczmirek (WebSM) 206