Questionnaire Testing for Business Surveys at Statistics Canada: Practices and Considerations

Dave Lawrence, Statistics Canada

Background

Over the past 20 years, the Questionnaire Design Resource Centre (QDRC), Statistics Canada has tested a wide array of business and establishment survey questionnaires. What have we learned from our experiences? How have these studies helped us establish sound testing practices? In which areas do we need to improve or gain more expertise?

This paper provides a brief synopsis of business survey questionnaire testing conducted by the QDRC at Statistics Canada. Particular emphasis is placed on the cognitive interviewing practices we have developed and the challenges faced over time. The paper concludes by asking some questions about what the future holds for researchers with respect to issues such as electronic data reporting and usability testing.

A Brief History

Statistics Canada has been using cognitive methods to test survey questionnaires since 1987. Early testing efforts focused on household and social surveys such as the *Labour Force Survey, Survey of School Leavers* and the *1991 Census of Population.*

The idea of using qualitative methods to pre-test new or redesigned business survey questionnaires was new to many project managers since many business survey programs of this period had been established as long-standing, repeated surveys. However, in the early 1990's cognitive methods were slowly extended to the testing of business and agricultural survey questionnaires at Statistics Canada such as the *Survey of Employment, Payroll and Hours (1989 and 1990)*, the *Census of Construction (1990)* the *1992 Farm Financial Survey* and the *1996 Census of Agriculture*.

In recent years, there has been an increased emphasis among program areas and survey takers to conduct questionnaire testing of business surveys. Since 2004 almost half the annual questionnaire testing projects undertaken by the Questionnaire Design Resource Centre (QDRC) have been either business or agricultural surveys. This increased demand is largely a result of program initiatives such as improving efficiencies in data collection and processing as well as improving respondent relations. We have also observed a change in the types of survey instruments. New and emerging data needs as well as a broader scope of subject matter topics has demonstrated a shift to service-oriented business surveys from the 'more traditional' manufacturing and retail survey domains. Recent testing projects have included topics such as information and technology, research and development, the environment and globalization.

The remainder of this paper focuses on the practices, considerations and challenges associated with conducting cognitive interviews with business and agricultural surveys.

Project Initiation

The QDRC provides questionnaire design and testing services to the various subject matter clients at Statistics Canada. For business and agricultural surveys, these services typically involve:

- Performing a critical review of a draft (test) questionnaire
- Meeting with the client area to obtain a clear understanding of the information needs and data requirements for the survey and the target population
- Establishing testing priorities and a fieldwork strategy
- Planning and implementing the fieldwork
- Reporting findings, observations and recommendations

In most situations with business and agricultural surveys, the QDRC reviews draft questionnaires developed by subject matter experts and survey stakeholders. The QDRC is not typically involved in the conceptual design, unless requested by the subject matter area.

Methodology and Fieldwork Strategy

The QDRC most commonly uses *on-site cognitive interviews* when testing business or agricultural survey questionnaires with members of the target population. Focus group discussions are also used to explore conceptual development, feasibility studies, as well as data quality and product assessments. However, cognitive interviews remain the principal tool for testing and evaluating test questionnaires where the goal is to identify areas where measurement error might occur and to address ways to possibly reduce such errors.

The rationale and advantages of conducting on-site cognitive interviews when testing business and agricultural questionnaires include:

- *Type and complexity of queried information*. The ability of respondents to retrieve the requested information depends on how well they understand the survey questions and the data source. Access to one or more information sources such as administrative or financial records, or other persons from the organization may be required. This can be better observed and explored on-site.

- *Compatibility* of survey questions and response categories with record-keeping practices and operational activities. Again, the degree to which data queries coincide with record-keeping practices is easier to observe and probe when on-site.

- *Location*. Business people and farm operators are busy. Conducting on-site visits typically makes it easier to schedule appointments and ensure higher participation rates.

- *Issues of confidentiality or sensitivity*. Business and agricultural survey questionnaires frequently request detailed information about business operations and practices. Participants are frequently reticent to share and discuss this type of information in a group setting. Cognitive interviews are more appropriate for exploring and understanding how participants react to and respond to these types of data queries.

- *Person Most Knowledgeable*. More than one respondent may be required to complete the test document. Inviting more than one participant to the interview may be done at the time of recruitment. On occasion, after having reflected on the test material, a participant decides to invite colleagues to the meeting.

Part of the testing strategy is determining the number of test sites to be visited and how many interviews to conduct. These decisions are made during the planning stages and are directly related to characteristics of the target population. Practical constraints such as budget as well as time available for conducting the fieldwork also play a role in the planning process.

Historically, for many business surveys, the number of test sites was limited to two or three locations – typically larger metropolitan areas where sufficient numbers of representative sample units could be found. While much business survey questionnaire testing continues to be carried out in major economic regions, our questionnaire testing often goes further a field. Specialized business populations in the resource sector such as oil and gas producers, mining, and aquaculture require consultations to be carried out in more remote areas or various locales. Diversity of operations may also dictate that questionnaire testing be carried out at several test site locations across the country. Agricultural practices are known to be different based on attributes such as geography and type of farm operation. This has a direct impact on the number of test site locations as well as the number of actual interviews conducted.

The QDRC, in consultation with the client, establishes fieldwork dates for each test site before recruitment begins. Since most on-site cognitive interviews occur during normal business hours, only 3 or 4 interviews a day are scheduled. Most interviews take about 60 or 70 minutes to complete. Another 30-45 minutes is allowed for traveling to the next location. Evening interviews are rare; however, some agricultural interviews are conducted at this time of day. Where possible recruiters will try to schedule interviews in the same general area each day to minimize travel times between appointments.

At Statistics Canada, questionnaire testing for national surveys almost always requires that a survey instrument be tested in both official languages – English and French. This factor also influences the number of test site locations.

Recruitment and Scheduling Considerations

Recruitment criteria

Unlike recruitment for social or household survey testing, centralized list frames are used to pre-identify and target possible participants for testing business and agricultural surveys. Due to the confidential nature of these registers, only Statistics Canada employees are permitted to access the lists for recruitment purposes. Further, in accordance with Statistics Canada's confidentiality requirements, all recruitment must be carried out on the premises. When the test site locations are established, recruitment of study participants is selective based on specific characteristics of interest.

For example, criteria for questionnaire testing with businesses may include:

- Size of business (number of employees, revenues)
- Corporate level (establishment, location, enterprise)
- Industry type (e.g., North American Industry Classification System (NAICS))
- Geographic location

Testing agricultural questionnaires may consider:

- Farm size (based on income, number of employees, etc.)
- Farm type (e.g., crops, livestock-type, greenhouse)
- Operating arrangements (corporation, partnership, etc.)
- Age of main operator

The recruitment specifications (number of test sites, number of interviews, site-specific recruitment lists and screening criteria) are determined in consultation with the client from which a detailed screening questionnaire is developed. Recruiters use this questionnaire to verify the screening criteria when setting up interview appointments. Recruitment typically starts 2 to 3 weeks before the fieldwork. A confirmation call is made approximately 24 hours before each scheduled appointment.

Recruiters

A key difference between recruiting for social or household surveys compared to business surveys is that in many cases potential participants for business survey testing may already be respondents to other government or private sector surveys. Experience has shown that these participants will ask specific questions about the purpose and nature of the testing study. They often want to know exactly what will be expected of them in terms of their time and effort before agreeing to participate in an on-site visit. For these reasons, recruiters are provided with detailed background information about the survey as well as how the collected information will be used.

Further, only small teams of experienced trained recruiters - often only one or two - are used on business and agricultural surveys. This practice facilitates an improved control over recruitment and helps ensure the most appropriate respondents are recruited for questionnaire testing.

When a participant is successfully recruited, specific address and personal contact information is confirmed with the participant; as this may differ from what is contained on the recruitment listing. Also, depending on the type of the business to be visited (e.g., farm, manufacturing plant, mining operation), there may be special access requirements for security or safety reasons that the interview team must adhere to when on-site. These details must always be clarified and confirmed during recruitment.

Recruitment and Scheduling Challenges

Willingness to participate

Completing survey forms or participating in questionnaire testing is not a priority for most respondents to business or agricultural surveys. Achieving the desired number of interviews per test site can be more difficult than recruiting from the general population for social survey testing. Experience has demonstrated that it may require a recruiter to make as many as 10 to 20 contacts with potential participants to confirm one appointment. Since most business surveys are mandatory, sometimes a respondent’s willingness to participate hinges on a sense of compliance or obligation.

In certain situations – in particular with large corporate entities or with new survey subject matter – we have found it quite useful to solicit the help of industry or agricultural associations to endorse the questionnaire testing study as well as explain to their members the rationale and potential benefits of participating.

Given the importance that very large establishments and agricultural operations have on collected survey information, it is often worthwhile to consider requesting collaboration from these entities as part of the overall testing strategy. At Statistics Canada, this involves coordinating recruitment and test site visits with a special unit mandated for such respondent relations.

Completeness and accuracy of recruitment lists

Recruiters use comprehensive lists that include information such as business name, telephone number, location, and a contact name when calling business units. Even if a contact name is available, this may or may not be the desired person for testing a specific questionnaire. A recruiter often has to confirm this information at the time of the call. To mitigate some of these issues, or in the absence of an actual contact name, the recruiter must be able to ask for the ‘best person to complete the survey’. For example, rather than establishing contact with the comptroller or Chief Executive Officer, this may involve determining the person who can best answer questions about a specific subject matter area.

This process may involve the recruiter having to consult with the cognitive interviewer or the subject matter expert during the recruitment process. Telephone number tracing, numbered companies, complex operating arrangements, multiple operations, are just some of the other challenges faced by the business and agriculture survey recruiter. The prevalence of these situations will impact the efficiency of and time spent on recruitment.

Gate keepers

Who the desired participant is for a study may impact the ability of a recruiter to establish direct contact. Potential participants such as executives or company presidents will almost always have executive assistants who screen access to and the availability of their supervisors.

Frame updates

As a result of the recruitment process, a recruiter may encounter 'new frame information' such as updates to telephone numbers, addresses or contact names. Further, information related to corporate issues such as company structure, out of scope units or out of business units may be valuable to survey designers and methodologists. Where possible, this information should be provided to the client.

Overlap detection

During recruitment, it is also important to be mindful of test site locations and testing conducted with small or specialized target populations to ensure participants in recent or previous questionnaire testing projects are not unknowingly re-contacted for a new study.

Timing and location

The time of year that on-site cognitive interviews are scheduled can be challenging both for recruiters and researchers. Fiscal year-ends, peak production cycles or other economic situations may affect participants' willingness to schedule interviews. On occasion, field work can be modified to better accommodate participants' schedules. For example, in early 2009, a study of small businesses involved with information and communications technology was rescheduled (and reduced in size) due to the economic climate and business priorities of small operators at that time.

On other occasions the fieldwork schedule can be less flexible. For example, the Census of Agriculture is conducted every five years during the month of May – a very busy time for Canadian farm operators. Although, difficult for recruitment, questionnaire testing is frequently carried out during this month in order to replicate as closely as possible the actual data collection period when testing with farm operators.

Another factor that influences scheduling fieldwork is the nature of the target population itself. Businesses whose activities are specialized or unique may require traveling extensive distances to conduct a small number of cognitive interviews. For example, testing survey questionnaires about greenhouse gas emissions, refining petroleum, or aquaculture, has involved extensive travel for QDRC consultants to 'non-traditional' and sometimes remote locations.

Fieldwork

Interview teams

At the QDRC, typically no more than two or three cognitive interviewers will work on a given questionnaire testing project. Small teams help facilitate the overall planning and scheduling as well as help maintain consistent interview protocols and techniques. Cognitive interviewers assigned to a specific testing project work together with the client to plan the overall testing strategy. This practice allows consultants to familiarize themselves with the relevant subject matter, concepts and terminology prior to commencing the fieldwork.

Subject matter clients are encouraged to travel on-site and directly observe questionnaire testing. Subject matter specialists are often not questionnaire design experts; however, by observing the cognitive interviews, they can see firsthand what works and what does not work with their test instrument. In addition, business and agricultural surveys typically involve complex concepts or data elements required for specific policy programs or economic models. It is beneficial to the overall success of the interviews to have an observer who is very knowledgeable of all facets of the survey and who can appropriately address specific queries from participants.

The number of observers is typically limited to a maximum of two per interview. In practice, there is often just one observer.

When conducting on-site cognitive interviews with businesses, interviewers and observers typically take hand-written notes as opposed to audio-recording the sessions. Some business participants have been leery of recording conversations when discussing business affairs. Given the potential negative influence on participants' willingness to speak freely and share information, we usually opt not to record.

Cognitive interviews in practice

Having assessed the overall test objectives and reviewed the test questionnaire, the cognitive interviewers work with the client to develop an interviewer guide. The guide is essentially a list of probes used by the interviewer to explore pre-identified issues or potential problem areas with the test questionnaire. Due to varying business structures and operating arrangements, a complete and comprehensive set of probes often cannot be created before the fieldwork begins.

The QDRC often advises the client to conduct two or more rounds of cognitive interviewing; or if appropriate, some preliminary focus group discussions followed by one or more rounds of cognitive interviews. Where possible, the interview teams will first conduct several cognitive interviews locally, before traveling to other test sites. This preliminary testing can help improve probes as well as identify unanticipated issues or problems. This practice also allows clients to observe at minimal cost. Conducting local testing first permits preliminary debriefing meetings with the client and stakeholders. If needed, adjustments can be made to the interviewer guide or test materials before travel to other test sites is initiated.

Verbal probing and think aloud interviews

When testing business and agricultural survey questionnaires, our approach has typically been that of using verbal probes as opposed to think-aloud interviews.

When interviewing business survey participants, verbal probes help maintain control of the interview as well as permit the consultant to manage the allotted time for test items and probing. Time is money for business respondents. More than once, an interview team has arrived at a scheduled interview only to be told by the participant that the scheduled 60 minutes has been reduced to 40 because of a more pressing business matter.

For self-complete, one-on-one interviews, participants are always invited to 'think-aloud' as they complete the test questionnaire. A small number of participants seem to be able to do this naturally. Others try to 'think-aloud' as they start the questionnaire but often they are unable to maintain this practice without constant prompting from the interviewer. This general lack of proficiency, increased burden and the lack of 'training time' in an on-site setting tends to diminish the utility of the think-aloud in practice.

Verbal probes are structured in the sense that the cognitive interviewers know the issues to be explored. The wording for these probes may not be literally scripted; however, an objective line of query is typically pre-defined and discussed between consultants.

As noted above, varying corporate structures and business operations often make it difficult to create a complete set of probes *a priori*. Situations do arise in the field where an interviewer decides to explore an issue that arises during the session. This may occur either when an interviewer decides to explore a concept that had not been pre-identified, or it may happen due to a response or query from the participant. Where possible, the interviewer will note these occurrences and apprise his or her colleagues in the field.

Concurrent versus retrospective probing

In the early years of testing business survey questionnaires, the QDRC sometimes sent advance materials and questionnaires to test participants before the scheduled on-site interview. Participants were encouraged to review the questionnaire and complete it before the appointment. The on-site interview would then proceed with retrospective probing. The rationale for this approach was that on-site interview time was limited. Since data queries often involved accessing external files or records, it was felt that participants might not have sufficient time to truly reflect on the scope of the data queries in the allotted interview time.

While some participants did complete the questionnaire prior to the cognitive interview, we found that more frequently the document was not completed. It had often been relegated to a lower position in the ‘In basket’, or at best, the material had only been quickly perused. Further, it was felt that some observational information with respect to how participants react to the test document was missing from this approach.

This strategy has evolved over time. Currently, QDRC interviewers tend to more commonly use concurrent probing when testing business and agricultural survey questionnaires. Concurrent probing here is interpreted as probing that is conducted <u>at the time</u> the test questionnaire is administered or completed. Whether the actual probing is carried out at specific intervals during the interview, or following the completion of the questionnaire depends on factors such as the overall length of the questionnaire or homogeneity of data queries.

Occasionally, participants will request to see a copy of the test questionnaire before the scheduled appointment. Although this practice is usually not part of the proposed field methodology, an advance copy is sent in advance if their participation hinges on first seeing the material. In these situations, the actual cognitive interview may proceed with retrospective or concurrent probing depending on whether the questionnaire has been completed in advance of the interview.

Some agricultural surveys have a mail-out, CATI follow-up data collection methodology. In these situations, we may wish to ‘mirror’ the actual data collection as closely as possible. Recruiters inform participants that they will receive a questionnaire package in the mail. Participants are asked to follow the instructions and to complete the questionnaire prior to the on-site visit. At the time of the interview, we typically find that some questionnaires have been fully or partially completed and other times nothing has been done at all. The interviewer will adapt to the specific situation and conduct the interview as a retrospective or concurrent session.

Looking Ahead

This paper has provided a brief overview of practices and challenges of conducting business and agricultural survey questionnaire testing at Statistics Canada, specifically looking at on-site cognitive interviewing. Many of the planning and implementation practices discussed have evolved from conducting numerous studies with a wide array of subject matter areas and target populations.

While we continually try to improve our practices and cognitive interviewing skills, there are a few areas where we need to focus our efforts. We can learn both from our in-house practices as well as those of our international counterparts:

- At this point in time, the QDRC typically is involved with reviewing and testing draft questionnaires developed by others. Earlier involvement of questionnaire design experts in the conceptual development of test questionnaires would improve the initial quality of these test instruments.

- The QDRC has not yet had much exposure to usability testing of web-based surveys. This is an emerging area at Statistics Canada where we need to learn and develop sound techniques for testing and evaluating survey instruments.

- While we strive to produce better questions, we often have no real proof. Cognitive testing provides us with information and feedback from which we make recommendations or suggestions for improvements. These techniques alone do not validate the process. We should be constantly open to considering new or alternative measures of success in evaluating questionnaires.

Relations between Cognitive Problems and Statistical Bias

Gustav Haraldsen, Statistics Norway

In this paper I will raise two questions. First I will ask to what extent qualitative methods like cognitive interviewing disclose questionnaire problems that actually have a significant effect on the survey results. Secondly I would like to turn this question the other way around, and ask to what extent the statistical bias recognised in surveys capture systematic errors detected in qualitative tests. The topic will be discussed with reference to examples from some qualitative and quantitative tests that were carried out in a Norwegian survey about working conditions. There both different terms and different response scales used in the questions were first addressed qualitatively during the questionnaire development and subsequently tested in quantitative experiments and re interviews during the field period.

Concurrent Cognitive Interviewing and Usability Testing

Jennifer Hunter Childs, Jennifer C. Romano, Erica Olmsted-Hawala, and Nathan Jurgenson

U.S. Census Bureau

Author Note

Jennifer Hunter Childs, Jennifer C. Romano, Erica Olmsted-Hawala, and Nathan Jurgenson, Statistical Research Division, U.S. Census Bureau. Correspondence concerning this article should be addressed to Jennifer Hunter Childs, Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233. E-mail: jennifer.hunter.childs@census.gov

Disclaimer: This report is released to inform interested parties of research and to encourage discussion.

Any views expressed on methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau. Preliminary findings from this paper were presented at the May 2009 Questionnaire Evaluation Standards Conference in Bergen, Norway and at the May 2009 American Association for Public Opinion Research Conference in Hollywood, Florida, USA.

Abstract

Survey pretesting timelines often involve cognitive interviewing that precedes, and is very distinct from, usability testing. In this paper, we make the case that conducting cognitive and usability testing concurrently with usability and cognitive interviewing experts working together would be more comprehensive and less labor-intensive approach to pretesting. By testing the same questionnaire concurrently with respondents and interviewers (the users in this case), potentially problematic question wording and survey design can be more efficiently identified in a way that can be used to improve the questionnaire for both the respondent and the interviewer.

Prior to the 2006 Census Test, the U.S. Census Bureau conducted separate rounds of cognitive and usability testing on an interviewer-administered Nonresponse Followup questionnaire in preparation for the 2010 Census. In doing the testing separately, we learned that, in addition to usability issues, usability testing also identified cognitive question wording issues. Similarly, while examining question wording, cognitive interviewing also identified poor usability features. Though many techniques used in these two types of studies are similar, usability researchers are not typically trained specifically in survey methodology, and survey methodologists are typically not trained in human-computer interaction. Usability researchers are perfectly capable of recommending changes in wording, but their recommendations may not be informed by research known to the survey methodologists who conduct cognitive interviewing. Similarly, survey methodologists may make recommendations for improving usability that are not backed up by research in human-computer interaction. While well intentioned, such uninformed recommendations could do more harm than good.

In 2008, the Census Bureau cognitive and usability labs concurrently conducted 40 cognitive and 20 usability interviews to test a revised, paper-based Nonresponse Followup questionnaire and presented results and recommendations from both types of testing together. When testing was conducted concurrently, staff from both labs, representing both specialties, were at the table at the same time, creating a more efficient methodology. By examining this case study, this paper discusses what can be gained by conducting these studies concurrently rather than conducting them independently (as is typically done). This paper also provides examples of findings and recommendations that are possible through this joint research and the synergy from having members from both research teams involved.

Key words: Cognitive interviewing, usability testing, survey pretesting

Concurrent Cognitive Interviewing and Usability Testing

Survey methodology literature describes how cognitive interviewing and usability testing share many of the same techniques and have many of the same origins in cognitive psychology (Couper, 1999; Groves et al., 2004; Willis, 2005). Yet there is little evidence in the published literature of researchers using these methods concurrently (Willis, 2005). Survey pretesting timelines typically involve cognitive interviewing that precedes, and is very distinct from, usability testing.

Quite commonly, separate labs exist that specialize in *either* cognitive interviewing methods or usability testing. Cognitive interviewing labs generally specialize in understanding and analyzing surveys from the respondent's perspective. Usability labs, in the survey industry, focus on the user of the survey, which would be the interviewer in interviewer-administered surveys. In conducting testing separately at the U.S. Census Bureau, we have learned that in addition to usability issues, usability testing also identifies cognitive question wording issues. Similarly, while examining question wording, cognitive interviewing also identifies poor usability features. Though many techniques used in these two types of studies are similar, usability researchers are not typically trained specifically in survey methodology, and survey methodologists are typically not trained in human-computer interaction. While usability researchers are perfectly capable of recommending changes in wording, their recommendations may not be informed by research known to the survey methodologists who conduct cognitive interviewing. Similarly, survey methodologists may make recommendations for improving usability that are not backed up by research in human-computer interaction. While well intentioned, such uninformed recommendations could do more harm than good. For example, a question wording problem could be identified through usability testing, and in the absence of the knowledge of previous cognitive interview findings, the usability researcher may recommend alternative wording that has tested poorly in previous rounds of cognitive testing.

In order to take advantage of the strengths of both methodologies, the Census Bureau is now attempting to coordinate cognitive and usability testing and to consider how the synchronization will function in practice. In this paper, we present a case study in which we took a more comprehensive and less labor-intensive approach to pretesting by conducting both cognitive and usability testing at the same time on the 2010 Census Nonresponse Followup (NRFU) questionnaire. By concurrently testing the same survey with different sets of potential respondents and interviewers (the users, in this case), problematic question wording and issues with elements of the survey design were more efficiently identified and corrected, leading to an improved questionnaire for both the respondent and the interviewer. This paper outlines the benefits of conducting these tests concurrently by involving members of both the cognitive and usability labs as well as, the types of findings that were possible through this joint research and the synergy from having members from both research teams involved.

Background

Interestingly, though usability and human factors researchers sometimes reference the "cognitive interview," they do not often cite the same body of research that those in the field of cognitive interviewing typically cite (for usability studies that use a different frame of reference for the term "cognitive interview," see Howard & MacEachren, 1996; Roth & Harrower, 2008; Partala & Kangaskorte, 2009). Similarly, it is not uncommon for cognitive interviewing reports to talk about the "usability" of a questionnaire, often meaning understandability rather than usability as it is understood in the context of human factors research. However, in survey research cognitive interviewing and usability testing most often refer to two distinct types of testing.

The Census Bureau has two laboratories for pretesting population surveys, one dedicated to cognitive interviewing and the other dedicated to usability testing. The two staffs usually work separately, and thus, similar work that could be conducted jointly is often conducted independently.

Cognitive Interviewing

Cognitive interviewing is a well-known and commonly-practiced form of pretest by which social scientists usually conduct a semi-scripted interview with individual respondents in order to understand how respondents comprehend and answer the survey questions (see DeMaio & Rothgeb, 1996; U.S. Census Bureau, 2003; Willis, 2005). Interviewers probe respondents – either concurrently with the administration of the survey questions, or retrospectively, after the survey has been completed – to assess how well respondents understand the questions and concepts being measured, as well as the accuracy of their responses for themselves and their households. Some researchers also include "think-aloud" protocols in which respondents are asked to think out loud while filling out a self-administered questionnaire or while deciding how they will answer questions in an interviewer-administered survey (Beatty, 2004; Willis, 2005).

Generally, in cognitive interviewing, the focus is on respondents' understanding of question wording and their ability to map a response to the proposed response format. Typical results from cognitive interviewing show where survey respondents have difficulties understanding or responding to question wording or response categories, and thus, where revisions may be required. In a self-administered survey (either paper- or Web-based), there is some focus on navigation, but that is often secondary to the goal of evaluating question wording.

Cognitive interviewing is often used to determine the final question wording prior to field testing or fielding the survey.

Cognitive interviewing facilities at the Census Bureau. At the Census Bureau, the cognitive lab is composed of psychologists, sociologists, anthropologists, demographers, and sociolinguists. Cognitive interviewing at the Census Bureau dates back to the 1980s and the Cognitive Aspects of Survey Methodology (CASM) movement (DeMaio, 1983). The lab focuses largely on the four stage comprehension model proposed by Tourangeau, Rips, and Rasinski (2000). Cognitive interviewing seeks to measure accuracy at each of the four stages in the model – comprehension, retrieval, judgment and response.

Techniques commonly used in the Census Bureau's cognitive lab include concurrent think aloud, concurrent and/or retrospective probing, emergent and expansive probing, retrospective debriefing, and vignettes (hypothetical situations). The concurrent think-aloud method involves respondents verbally reporting their thought processes while they are completing a survey (Ericsson & Simon, 1993). According to Ericsson and Simon (1993), if the respondent is doing a problem-solving task to answer a question, this method may be successful; however, if the survey questions ask for basic information that the respondent already knows, a think-aloud protocol may not gather informative data, and may, in fact, actually disrupt the thought processes.

Concurrent probing involves administering survey questions one at a time, followed immediately by probes about different aspects of the survey question. An alternative method, retrospective probing, involves administering the entire survey questionnaire, or sections of it, without interrupting the respondent, and then probing at the end of the designated set of questions (Willis, 2005). Both types of probing can include paraphrase probes (e.g., "In your own words, what was that question asking?"); meaning probes (e.g., "What does 'race' mean to you in this question?"); process probes (e.g., "Tell me how you came up with your answer."); and expansive probes (e.g., "Tell me more about that."). Probes can be pre-determined and scripted, or they can be emergent, based on information that has been revealed in the interview (see Willis, 2005 for details on these different techniques).

Using a retrospective debriefing technique, the cognitive interviewer probes the respondent after he or she has finished filling out the entire questionnaire, or has completed the entire interview. The debriefing could ask questions about the questions that were in the survey, similar to the types of probes mentioned above, or it could ask expansive probes to further understand the respondent's own situation, in order to assess whether or not he or she answered the survey questions correctly.

Vignettes are hypothetical situations often used to cover situations that may not occur (or may occur at a very low frequency) within the population being interviewed. In vignettes, short stories are told or facts are presented to the respondent and the respondent is asked how he or she would respond to the survey question if he or she was in the situation presented in the vignette (Bates & DeMaio, 1989; Gerber, 1994; Gerber, Wellens, & Keeley, 1996).

In summary, the cognitive lab examines respondent understanding of and ability to respond to survey questions.

Usability Testing

Usability is defined as the extent to which a given product can be used with efficiency, effectiveness, and satisfaction (ISO 9241-11, 1998). The goal of usability testing is often to improve the product, so that people who use the product can do so quickly and easily (Dumas & Redish, 1999; Rubin, 1984). Usability testing within the survey industry is usually associated with Web sites and Internet self-administered data collection instruments, but usability testing can also be conducted on an interviewer-administered survey, in which case the interviewer is the user (Hansen, Couper, & Fuchs, 1998; Couper, 2000).

When conducting a usability test of a Computer-Assisted Telephone Interview (CATI) or Computer-Assisted Personal Interview (CAPI) survey, the participant (or user) is given a limited amount of training on how to administer the survey and then is asked to play the part of an interviewer. The participants administer the survey to respondents in order to assess the usability of the survey. The respondent role is often played by researchers who use prearranged respondent scripts to test specific interfaces in the survey. The goal of the usability test is to evaluate how usable the instrument is, i.e., whether it is intuitive enough for someone with limited training to be able to navigate without many problems. For survey usability testing, the focus is on the users' interaction with the questionnaire and whether they can complete the survey (either as a respondent in a self-administered format or as an interviewer in an interviewer-administered format). Usability testing also assesses the visual design (i.e., the look and feel) of the survey, with the goal of improving visual design and navigation.

Usability testing facilities at the Census Bureau. The usability lab at the Census Bureau is composed primarily of psychologists and human factors and usability specialists. Usability testing at the Census Bureau is based on principles of user-centered design (UCD), such that it takes into account the user's needs, wants, and capabilities, and can take place at various stages of survey development (Mayhew, 1999; Nielsen, 1993a). Usability testing often seeks to measure accuracy (whether people can complete a task or survey correctly, given verifiable information), efficiency (the time it takes to complete a task or survey), and participants’ self-rated satisfaction (with the entire Web site or survey, and/or with particular parts of it).

Techniques commonly used in the usability lab are similar to the methods used in the cognitive lab. The usability lab uses concurrent think aloud, concurrent and/or retrospective probing, retrospective debriefing, and satisfaction questionnaires. Concurrent think aloud, as previously described, is used to understand the participants’ thought process as they complete a task or survey (Ericsson & Simon, 1993). Usability practitioners use variations on the traditional think-aloud protocol with some new emerging protocols, such as the speech-communication protocol (Boren & Ramey, 2000; Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010). Concurrent probing is typically used when a product is in development and is not completely functional. For example, for a low-fidelity usability study of a paper prototype, if a participant said they would click on a link, the researcher would ask, “What do you think would happen when you click on the link?” (see Romano, Olmsted-Hawala & Murphy, 2009). Retrospective probing can be used at any stage of development and is used to inquire about things participants said and did during testing. The retrospective probes are asked at the end of the session. Similarly, retrospective debriefing occurs at the end of the session when participants are asked questions about specific aspects of the survey and things they liked or disliked about it. Prior to debriefing, after completing the session, participants often complete a satisfaction questionnaire in which they rate their satisfaction with the survey as a whole and with various aspects of it.

A typical usability study involves participants working with a Web site or a survey as they “normally” would and thinking aloud as they do so. If they become quiet, the researcher probes by saying, for example, “Tell me your thoughts,” “What are you thinking?”, or “Keep talking.” Even a quietly affirming “umm hum” by the researcher works to keep participants verbalizing their thoughts. The goal is for the participant to keep a running commentary during the session so researchers can hear participants' reactions as they use the site. For a Web site usability test, participants are given specific tasks to do on the site. Tasks are designed based on typical tasks that users would do on the Web site (e.g., information-seeking tasks; search-related tasks). For a survey, participants are given a scenario, such as pretending they received a letter in the mail asking them to complete the survey. Upon completion of the session, participants fill out a satisfaction questionnaire, and then the researcher asks additional debriefing questions.

The goal of usability testing is often to see how users interact with a product, if it "works" for the user as intended, and how it could be improved. Often, developers do not take users' capabilities and limitations into account when developing a site or survey, and they miss important flaws that can only be discovered through user interaction. Through usability testing, researchers observe users interacting with the site or survey and identify parts of the design that can be improved.

Case Study: United States 2010 Census

As a part of the decennial census operations, the Census Bureau sends census forms to all known housing units in the country. The Census Bureau attempts to send an interviewer to every housing unit that does not return a census form by mail. The interviewer asks the household to participate in the census via an in-person interview. This personal visit is a part of the NRFU operation, in which the Census Bureau collects basic data on each housing unit (e.g., whether the unit is occupied, whether the unit is owned or rented) as well as some basic demographic data about each person who lives in the household (e.g., names, ages, races). The NRFU survey uses flashcards (also known as showcards, or in this case an Information Sheet) to assist respondents in answering particularly long or complex questions. See Figure 1 for the NRFU Information Sheet used in this study.

The Census Bureau had originally planned to collect NRFU data using a CAPI survey in 2010. Separate cognitive and usability tests were conducted (Childs, 2008; Olmsted & Hourcade, 2005; Olmsted, Hourcade & Abdalla, 2005), as this was the typical questionnaire-development procedure at that time. An early version of the survey underwent usability testing, and cognitive interviewing was conducted first with a paper script of the question wording, then later with a programmed hand-held computer (after usability testing was complete). The usability findings primarily focused on issues with data entry, navigation and interviewer tasks (e.g., handling flashcards) but also mentioned some instances of problematic question wording. These wording problems provided an early glimpse of many issues that were later explored through cognitive interviewing. For example, Olmsted and colleagues (Olmsted & Hourcade, 2005; Olmsted, Hourcade & Abdalla, 2005) found that reading topic-based questions over and over in their entirety, for each household member, was repetitive and burdensome for the interviewer. This finding was then replicated in cognitive interviewing and shown also to be burdensome for respondents (Childs, 2008).

The cognitive interview findings focused on a few navigational issues beyond those found in the usability testing, but findings were primarily on question wording problems and interviewer tasks. Separate testing yielded some unique results with each method and some redundant results. Additionally, in each test, researchers were working with only partial expertise and knowledge of the subject area from which recommendations to solve identified problems could be generated. Researchers from the usability lab had expertise with visual design and navigation, but not with question wording. Researchers from the cognitive lab had expertise with question wording, but not with visual design or navigation. Therefore, each test resulted in recommendations that may have been suboptimal because both the usability and cognitive aspects were not examined at the same time.

After the NRFU survey had gone through the original development process, the Census Bureau changed its design such that the 2010 NRFU data would be collected via an interviewer-administered paper-and-pencil instrument (PAPI), rather than via the CAPI instrument. Because of this late-breaking change in plans and the limited availability of time, concurrent cognitive and usability testing of the interviewer-administered questionnaire was essential. Born out of necessity, a new method of concurrent cognitive and usability testing emerged at the Census Bureau.

Method

Researchers from both the cognitive and usability labs worked together on the test design for each study and the reporting of joint findings and recommendations from the cognitive and usability testing. Forty cognitive interviews and 20 usability sessions were conducted with different respondents using the NRFU questionnaire. This test is larger than most usability studies, and than many cognitive interview studies conducted of the general population (for information on sample sizes in usability and cognitive interview studies, see Nielsen, 2003; Nielsen & Landauer, 1993; and Willis, 2005). It was conducted this way because of the short amount of time available for testing, and the desire to be as comprehensive as possible.

The cognitive interviewing sessions focused on respondent comprehension, accuracy, and the ability to answer questions, given participants' own situations. Thus, the participant, or subject, of the cognitive interview was the respondent. During the cognitive interviews, first, a researcher administered the NRFU questionnaire, playing the part of a field interviewer.

Researchers asked respondents to pretend they were in their own homes, and interviewers stood in front of respondents – simulating a doorstep interview. During the interview, researchers asked respondents to report any difficulty they were having while attempting to understand or answer any of the questions. Otherwise, the interview proceeded without interruption as if it were being administered in the field. Following the NRFU interview, a second, different cognitive researcher conducted a retrospective debriefing. During the debriefing, the researcher walked the respondent through each of the previously administered NRFU questions and responses and probed the respondent about the meaning of key questions and terms and about potentially difficult or sensitive questions. These key questions had been identified through prior experience with these questions (for a summary of recent pretesting with this questionnaire see Childs, 2008). Cognitive interview sessions examined accuracy at each of the four stages of survey response – comprehension, retrieval, judgment and response.

While the cognitive interviews focused on the respondent's perspective, the usability sessions focused on the interviewer's experience. The participant, or subject, in the usability study was the interviewer. The usability sessions began with a 45-minute interviewer-training session. A research assistant (trainer) read an abbreviated training script, which was taken from the actual interviewer-training manual, verbatim, so that all participants were exposed to the same information that an actual interviewer would be. At the end of the training, the participant (interviewer) completed two practice exercises, similar to those that would be completed in the actual NRFU training. The trainer corrected the participant as necessary, as would be done in actual training. Following this training activity, the participant conducted four interviews, interviewing a different research assistant who read from four different pre-scripted scenarios that were designed to test different situations that an interviewer might face. These scenarios were based on prior experience with these questions (Childs, 2008; Olmsted & Hourcade, 2005; Olmsted et al., 2005; Olmsted, 2004). Following these exercises, participants were asked debriefing questions about their experiences with the NRFU questionnaire. The usability sessions measured accuracy (whether participants could accurately complete each question for each scenario), satisfaction (self-rated satisfaction ratings with using the questionnaire) and efficiency (how long it took to complete each interview).

Results

Overall, we found combining cognitive and usability testing to be a beneficial strategy. The results presented here provide examples of ways that we benefitted from the joint methodology above and beyond separate testing.

Information Sheet Findings

The flashcard used in the PAPI NRFU test was in the form of an Information Sheet that the respondent could keep. Figure 1 shows the Information Sheet. This had recently been changed from a different format (a booklet), and the joint cognitive and usability test was the first opportunity to examine how the new format would work – both for interviewers and for respondents.

In the cognitive interview, for the most part, respondents answered comprehension-related probes indicating that they understood the Information Sheet as it was intended and were able to use it for all relevant questions. Based on the cognitive interviewing, one minor recommendation was made to change the wording in one of the Lists on the Information Sheet and that change was implemented for the 2010 Census. In the usability sessions, interviewers were also successful in administering the Information Sheet and reading the associated questions. Based on findings from both the cognitive and usability sessions, several recommendations were made for training interviewers on administering the Information Sheet. For example, in cognitive interviewing, respondents reported that they did not feel like they had enough time to read List A. To address this, we recommended adding a note to the interviewer training suggesting that the interviewer give the respondent a moment to read List A after the Information Sheet is handed to him or her, prior to asking the question about it.

Through joint usability and cognitive testing, we concluded that the Information Sheet is a format that is usable by both interviewers and respondents and thus is recommended for use in the 2010 Census. Through only cognitive interviewing, we would not have known whether or not the new format would work well with interviewers. With only usability testing, we may not have realized that respondents needed additional time to process the information on the Information Sheet prior to answering the questions. By developing recommendations jointly, we were able to use information gained from experiences with the interviewers to help solve the respondent problem of not having enough time to process information.

Relationship Question Findings

Joint cognitive and usability testing allowed us to see both how respondents would answer questions about themselves and their households and how interviewers would deal with responses that did not exactly map onto the response categories. For example, the census question on relationship uses a list on the Information Sheet to display all 13 response categories to the respondent (see Figure 1 for List B). The question was worded as follows: “Please look at List B on the Information sheet. How is (NAME) related to (PERSON 1)?”

In the cognitive interviews, respondents often reported simply “son” or “daughter” and had to be probed by the interviewer to clarify whether the child was biological, adopted, or a stepchild. A few respondents struggled to find response categories that matched members of their households – e.g., a daughter's fiancé, a step-great-great-grandchild, a common law partner, and foster children. Although most of these respondents chose a category, it was not always the category that the Census Bureau would have used to classify them.

In the usability sessions, participants were given three test scenarios, and they were scored for their accuracy based on how the Census Bureau would categorize particular relationships. Table 1 shows participant accuracy across all cases. Respondents were better able to classify a nanny as an "other nonrelative" than a foster daughter (who technically fits into the same category), but accuracy was still low. The half-sister relationship was also misclassified (classified as something other than "brother or sister") in a majority of situations.

Based on these problems identified through cognitive and usability testing, we recommended that during training, it should be reinforced that interviewers need to ask the respondent to pick the most appropriate category from List B and should probe for "biological, adopted or stepson (or daughter)." Additionally, training should include examples like those that were problematic in testing (e.g., foster child, half-sister, common law partner).

The joint cognitive and usability testing allowed us to see both where respondents had difficulty mapping their households' situations onto the response categories, and also where interviewers had difficulty making that assessment. The combination of these problems allowed us to conclude that training should include problematic examples, as well as directions for the interviewer to ask the respondent to choose from the list of categories when they respond with a relationship that is not listed.

Hispanic-Origin Question Findings

The question on Hispanic origin demonstrated a situation in which respondents had difficulty with a concept but in which interviewers were better able to deal with the question when using pre-scripted scenarios similar to the actual respondent scenarios. This indicated that the presence of the interviewer could perhaps solve some of the problems that respondents had with this question.

The Hispanic origin question asks the respondent to look at the Information Sheet (List C on Figure 1) and answer the question "Are you of Hispanic, Latino, or Spanish origin? In the cognitive interviews, some Hispanic respondents answered with "Latino" or "Spanish" without providing their countries of origin (as List B suggests they should do). In other cases, respondents described both their race and their origin in response to this question. Additionally, some respondents were unable to respond to this question unassisted. However, with probing by the interviewer, they were able to provide an answer that was determined to be correct, in most cases, based on respondent debriefings in the cognitive interviews.

In the usability sessions, the test scenarios provided several different responses that respondents often give to this question, and the participants (interviewers) were remarkably accurate in their reporting of respondents' answers. Table 2 shows accuracy across five different origins (one of which was non-Hispanic) that were reported.

Because respondents had difficulty in the cognitive interviews but participants playing the role of interviewers in usability testing were able to successfully navigate the question, we focused on interviewer training, which covers problems that respondents might have and how to deal with them. Since usability testing suggested that the training was sufficient in this area, we had only very minor recommendations for improvement. Had we only conducted cognitive interviewing, we might have felt the need to make significant changes to the question wording to assist respondents, which could have gone into the field untested because of the very late date on which this testing was conducted. Conducting these tests concurrently allowed us to see the entire picture, from both the interviewer's and the respondent's perspectives, before recommending any changes.

Summary and Discussion

Though our experiment with concurrent testing was born out of necessity, it afforded us many benefits beyond those provided by traditional separate testing. By having the same group of researchers work on both the cognitive and usability portions of the study, we were able to gain a better understanding of how respondents react to questions (which are usually assessed during cognitive interviewing), how interviewers react to the questionnaire (which is usually assessed in usability testing), and how interviewers react to real respondent situations (usability test scripts were written by researchers that have been involved in iterative testing of these questions so that the scenarios reflected real interview situations). Through joint testing, we were able to make recommendations to improve the form, spanning from typical cognitive recommendations about question wording to typical usability recommendations about visual design and navigational instructions. We were also able to offer suggestions to improve interviewer training based on how respondents reacted to these questions in cognitive interviews and how interviewers dealt with (or failed to deal with) these issues in usability testing. These recommendations were more comprehensive than would have been available from either cognitive or usability testing conducted independently.

Each research team, with their expertise and knowledge of relevant research, contributed accordingly – with the usability lab team focusing on navigational and visual design issues and the cognitive lab team focusing on question wording and understanding. Jointly, the two teams were able to assess the magnitude of potential problems by looking at how the respondent/ interviewer interaction is likely to play out and were able to make recommendations, either for the form or for interviewer training. Key to the research presented here was the ability for experts in cognitive interviewing and usability testing to work jointly on test plans, and then form recommendations stemming from the entire set of interviews. Working together on the test plans allowed each lab to draw from experiences in the other lab to create plausible situations to study and to highlight areas that were likely to cause problems for the interviewer, the respondent or both. Working together on recommendations allowed every problem that was identified to be considered both from a respondent-comprehension perspective as well as from an interviewer- (or user-) centered perspective. Then, each research team, with their expertise and knowledge of relevant research, contributed to finding an optimal recommendation that would satisfy both interviewer and respondent needs.

Had this testing occurred separately, we argue that sub-optimal recommendations would have been generated, as the cognitive researchers would have focused mainly on the respondent, and the usability researchers would have focused predominately on the interviewer. Working together provided an efficient means of testing and generated recommendations that covered both areas of research.

Though the literature has discussed the need for both cognitive and usability testing in computer-assisted survey instruments, this is the first study to our knowledge that integrated both tests with researchers from each lab resulting in a single set of recommendations. Additionally, it is one of the few studies that applied usability testing to a paper form (see Moore, 2005 for a notable exception).

Future Directions

Although the study had a very limited time for testing, the combined resources of the cognitive and usability labs provided a good venue to demonstrate how a joint process could work. Based on this experience and the quality and quantity of data gathered, we recommend conducting cognitive and usability testing concurrently with early versions of surveys, with time to iteratively test changes in question wording, visual design (i.e., format and layout of the questionnaire), navigational strategies, and interviewer training. Testing should begin early in the process, when time is not a constraint, and continue until optimal response is achieved. Iterative testing is recommended in both cognitive and usability testing literature (Willis, 2005; Nielsen, 1993b; Bailey, 1993). We argue that this joint testing could be applied to any survey– self- or interviewer-administered and computerized, or paper. In any case, the usability portion would focus on the user of the survey (either the interviewer or the respondent) and the cognitive interviewing portion would focus on the respondents, particularly their ability to understand and respond to the questions accurately.

At the Census Bureau, another joint cognitive and usability test is currently underway. The testing of American Community Survey Internet survey is occurring early in the development cycle, with the opportunity for iterative testing. Because it is a self-administered survey, and the user is the respondent, we will be conducting single sessions that combine usability and cognitive interviewing techniques with the same respondent. This work will continue the discussion on the effects of conducting joint cognitive and usability testing.

Acknowledgements

The authors wish to acknowledge Elizabeth Murphy’s contributions to this paper and the work behind this paper. The authors also thank Theresa DeMaio, Kathleen Ashenfelter, and Yuling Pan for their comments on a previous version of this paper.

References

Bailey, G. (1993). Iterative methodology and designer training in human-computer interface design. *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, 198-205.

Bates, N. & DeMaio, T. (1989). Using cognitive research methods to improve the design of the decennial census form. *Proceedings of the U.S. Bureau of the Census Annual Research Conference,* 267-277.

Beatty, P. (2004), The Dynamics of Cognitive Interviewing, in Presser et al. (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*, Hoboken, NJ: Wiley.

Boren, T. & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication 43*, 3, 261-278

Childs, J. H. (2008). 2010 NRFU Questionnaire Development: From the 2004 Census Test to the 2008 Dress Rehearsal. *Statistical Research Division Study Series SSM 2008-05.* U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/ssm2008-05.pdf

Childs, J.H., Norris, D.R., DeMaio, T.J., Fernandez, L., Clifton, M., & Meyers, M. (2009). 2010 Nonresponse Followup enumerator questionnaire cognitive test findings and recommendations. *Statistical Research Division Research Report Series RSM2009-05*. U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rsm2009-05.pdf

Couper, M. P. (1999). The application of cognitive science to computer assisted interviewing. In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds.), *Cognition and Survey Research*. (pp. 277-300). New York: Wiley.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18,* 383-396.

DeMaio, T. J., Ed. (1983). Approaches to developing questionnaires. Statistical Policy Working Paper 10. Washington, D.C., Office of Management and Budget.

DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. Schwartz & S. Sudman (Eds.), *Answering questions: Methodology for cognitive and communicative processes in survey research* (pp. 177-196). San Francisco, CA: Jossey-Bass.

Dumas, J. S. & Redish, J. C. (1999). *A Practical Guide to Usability Testing*. 2nd ed. Intellect Books, Exeter, UK.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data.* MIT Press, Cambridge, MA.

Gerber, E. (1994). Hidden assumptions: The use of vignettes in cognitive interviewing. *Working Papers in Survey Methodology SM94/05.* Washington, D.C.: U.S. Census Bureau, Statistical Research Division. http://www.census.gov/srd/papers/pdf/sm9405.pdf

Gerber, E., Wellens, T. R., & Keeley, C. (1996). Who Lives Here? The Use of Vignettes in Household Roster Research. *Working Papers in Survey Methodology SM96/02,* Washington, D.C.: U.S. Census Bureau, Statistical Research Division. http://www.census.gov/srd/papers/pdf/erg9601.pdf

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology.* New York: Wiley.

Hansen, S.E., Couper, M.P., & Fuchs, M. (1998). Usability evaluation of the NHIS CAPI instrument. Paper Presented at the $_{53}$rd Annual Conference of the American Association for Public Opinion Research, St. Louis, MO.

Howard, D.L., & MacEachren. (1996). Interface design for geographic visualization: tools for representing reliability. *Cartography and geographic information science, 23*, 59-77.

ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. ISO/IEC.

Mayhew, D. (1999). *The Usability Engineering Lifecycle: a practitioner's handbook for user interface design*. Morgan Kaufmann, San Francisco, CA.

Moore, D. (2005). Cognitive interview usability testing of the redesigned SUIDIR form. Paper presented at the 2005 American Association for Public Opinion Research Conference, Miami, FL.

Nielsen, J. (1993a). *Usability Engineering*. Academic Press, San Diego, CA

Nielsen, J. (1993b). Iterative User-Interface Design. *Computer,26*, 32-41.

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, *41*, 385-397.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, 206-213.*

Olmsted, E. (2004). Usability Study on the Use of Handheld Devices to Collect Census Data. *Proceeding of the International Professional Communication Conference, 131-138.*

Olmsted, E. & Hourcade, J. P. (2005). NRFU Round II. Quick Report submitted to DMD August 25, 2005.

Olmsted, E., Hourcade, J. P., & Abdalla, A. (2005). NRFU 2006 Quick Report Round I. Submitted to DMD August 8, 2005.

Olmsted-Hawala, E., Murphy, E., Hawala, S., & Ashenfelter, K. (2010). Think-Aloud Protocols: A Comparison of Three Think-Aloud Protocols for use in Testing Data-Dissemination Web Sites for Usability. *Proceedings of CHI.*

Partala, T., & Kangaskorte, R. (2009). The combined walkthrough: Measuring behavioral, affective, and cognitive information in usability testing. *Journal of Usability Studies, 5*, 21-33.

Roth, R.E., & Harrower, M. (2008). Addressing map interface usability: Learning from the Lakeshore Nature Preserve interactive map. *Cartographic Perspectives, 60*, 4-24.

Romano, J. C., Olmsted-Hawala, E. L., & Murphy, E. D. (2009). A usability evaluation of iteration 1 of the new American Fact-Finder web site: conceptual design (Statistical Research Division Study Series SSM2009-05). U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/ssm2009-05.pdf.

Romano, J., Murphy, E., Olmsted-Hawala, E., & Childs, J. H. (2008). A usability evaluation of the Nonresponse Followup enumerator (NRFU) questionnaire form. *Statistical Research Division Study Series SSM2008-10*. U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/ssm2008-10.pdf.

Rubin, J. (1984). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley and Sons, Inc.

Schwede, L. (2009). What can we learn from within-site pretesting of the Census 2010 enumerator questionnaire on the Navajo reservation? Paper presentation at the Society for Applied Anthropology Annual Meetings, Santa Fe, NM.

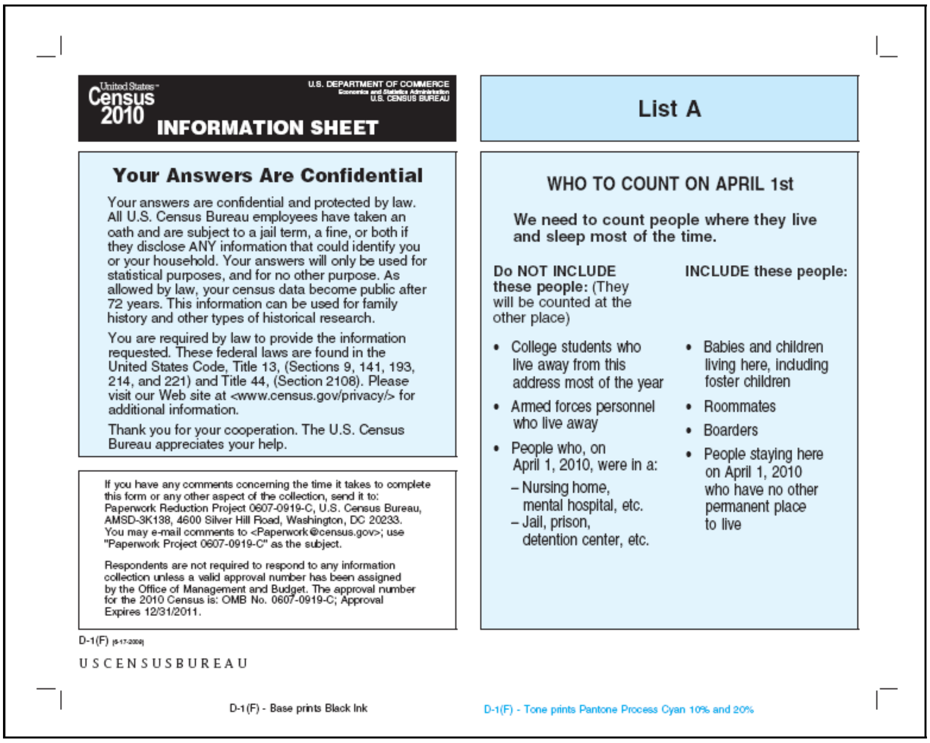
Tourangeau R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press. Cambridge United Kingdom.

Willis, G. B. (2005). Cognitive Interviewing: A tool for improving questionnaire design. London: Sage.

Figure Captions

Figure 1. Front and back of NRFU Information Sheet.

Figure 1



United States™ Census 2010 INFORMATION SHEET

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Your Answers Are Confidential

Your answers are confidential and protected by law. All U.S. Census Bureau employees have taken an oath and are subject to a jail term, a fine, or both if they disclose ANY information that could identify you or your household. Your answers will only be used for statistical purposes, and for no other purpose. As allowed by law, your census data become public after 72 years. This information can be used for family history and other types of historical research.

You are required by law to provide the information requested. These federal laws are found in the United States Code, Title 13, (Sections 9, 141, 193, 214, and 221) and Title 44, (Section 2108). Please visit our Web site at <www.census.gov/privacy/> for additional information.

Thank you for your cooperation. The U.S. Census Bureau appreciates your help.

If you have any comments concerning the time it takes to complete this form or any other aspect of the collection, send it to: Paperwork Reduction Project 0607-0919-C, U.S. Census Bureau, AMSD-3K138, 4600 Silver Hill Road, Washington, DC 20233. You may e-mail comments to <Paperwork@census.gov>; use "Paperwork Project 0607-0919-C" as the subject.

Respondents are not required to respond to any information collection unless a valid approval number has been assigned by the Office of Management and Budget. The approval number for the 2010 Census is: OMB No. 0607-0919-C; Approval Expires 12/31/2011.

D-1(F) (6-17-2008)

USCENSUSBUREAU

List A

WHO TO COUNT ON APRIL 1st

We need to count people where they live and sleep most of the time.

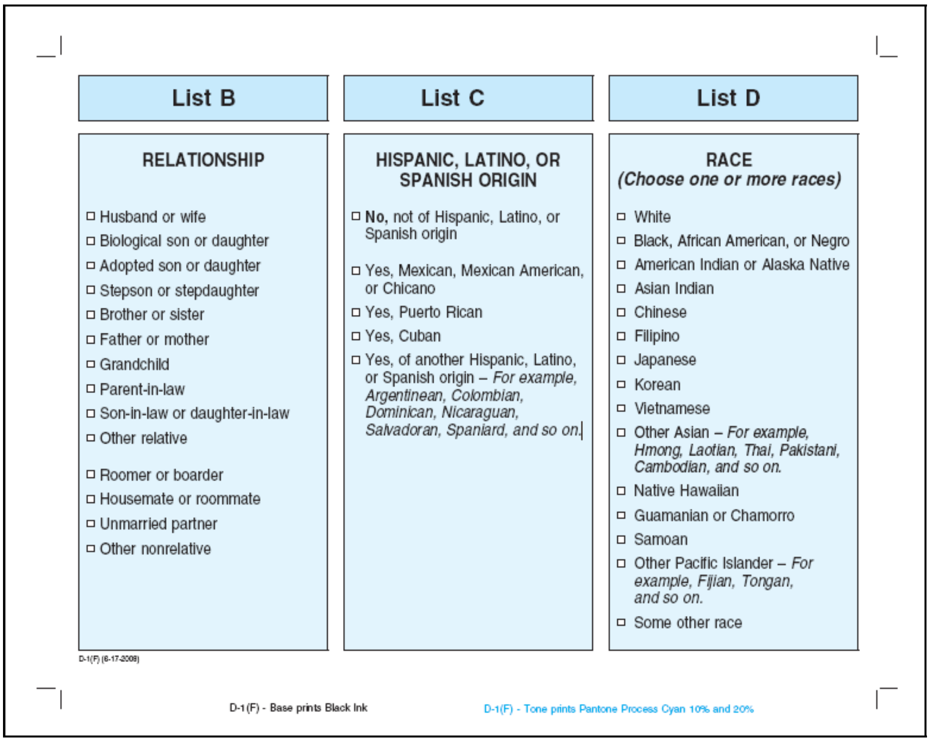
Do NOT INCLUDE these people: (They will be counted at the other place)

- College students who live away from this address most of the year
- Armed forces personnel who live away
- People who, on April 1, 2010, were in a:
  - – Nursing home, mental hospital, etc.
  - – Jail, prison, detention center, etc.

INCLUDE these people:

- Babies and children living here, including foster children
- Roommates
- Boarders
- People staying here on April 1, 2010 who have no other permanent place to live

D-1(F) - Base prints Black Ink D-1(F) - Tone prints Pantone Process Cyan 10% and 20%



List B

RELATIONSHIP

□ Husband or wife
□ Biological son or daughter
□ Adopted son or daughter
□ Stepson or stepdaughter
□ Brother or sister
□ Father or mother
□ Grandchild
□ Parent-in-law
□ Son-in-law or daughter-in-law
□ Other relative

□ Roomer or boarder
□ Housemate or roommate
□ Unmarried partner
□ Other nonrelative

List C

HISPANIC, LATINO, OR SPANISH ORIGIN

□ **No,** not of Hispanic, Latino, or Spanish origin

□ Yes, Mexican, Mexican American, or Chicano
□ Yes, Puerto Rican
□ Yes, Cuban
□ Yes, of another Hispanic, Latino, or Spanish origin – *For example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

List D

RACE
(Choose one or more races)

□ White
□ Black, African American, or Negro
□ American Indian or Alaska Native
□ Asian Indian
□ Chinese
□ Filipino
□ Japanese
□ Korean
□ Vietnamese
□ Other Asian – *For example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander – *For example, Fijian, Tongan, and so on.*
□ Some other race

D-1(F) (6-17-2008)

D-1(F) - Base prints Black Ink D-1(F) - Tone prints Pantone Process Cyan 10% and 20%

Table 1

Participant Accuracy from Usability Testing of Personal-Situation Question.

Scenario	Question #2: How is (Name) related to (Person 1)?
Nanny	68%
Foster daughter	21%
Half sister	37%

Table 2

Participant Accuracy from Usability Testing of Spanish-Origin Question.

Scenario	Question #5: Is (Name) of Hispanic, Latino or Spanish origin?
Dominican Republic	95%
Colombian	95%
Puerto Rican	100%
No, Cambodian	95%
Mexican	100%

Laying the foundation for good survey questions

Bente Hole

This paper focuses on the very first stages of a questionnaire design- and evaluation process, i.e. on what Esposito calls the observation and conceptualisation phases (Esposito 2002). These phases involve forming concepts, defining important sub-domains of the concept's meaning and finding empirical indicators for each concept or sub-domain (Hox, 1997). The paper points to some of the central literature on the subject, presents a few examples on how this part of the process has been implemented in practice and discusses advantages and drawbacks related to the different approaches. It seems clear that an iterative process is needed and that proper assessment in connection with every iteration is crucial in order to achieve good results. Further more it is considered whether techniques related to focus groups and exploratory interviews can be of value and whether a combination of a bottom up and a top down research strategy is profitable in order to fill in the gap between theory and measurement.

Testing of concepts - trust as an example

Merja Kallio-Peltoniemi, Statistics Finland

The origin of questions used in statistical surveys is often somewhat unclear. Questions are adopted e.g. on the basis that they have been used in previous surveys elsewhere. The actual phase of operationalising the research objective is omitted and the theoretical concept to be investigated is left without an elaboration. In such case one cannot be quite sure if the question measures what it is meant to measure.

The Finnish Cognitive Laboratory received the assignment from Statistics Finland's Leisure Survey, where researchers had doubts about the validity of the trust measurement. My research colleague Riitta Hanifi has tried to trace the origin of the questions measuring trust. Ms Hanifi has also tried to find an explanation for why precisely these questions are a good indicator of social trust. So far, she has not found any information or reasons that would seem good enough. It is interesting that operational definitions behind the questions are hard to find, and if found, they are blurry and not specific enough with regard to the standards set for unambiguous measures and operationalisation. Difficulties also stem from the fact that researchers are not unanimous about the definition of trust.

In this project I try to outline a way or produce a recommendation to test the concepts in general. (My work is still unfinished, it is meant to start in summer and be completed in autumn.) Exhaustive testing of concepts is something of a forgotten area in the field of pre-testing methodology and questionnaire evaluation standards. My view is that by testing concepts from the respondents' point of view we can increase our knowledge about the concepts and thus produce better survey questions. I use the concept of trust (in other people, in institutions) as an example. I will try to outline a procedure with which a researcher could ensure that the operationalisation of the concept is valid, which would also make the formulation of questions easier. This kind of an analysis can also provide an answer to the question 'what have we actually been measuring' by comparing the definitions used by respondents to the operational definition. This kind of information can be used when interpreting the statistical information.

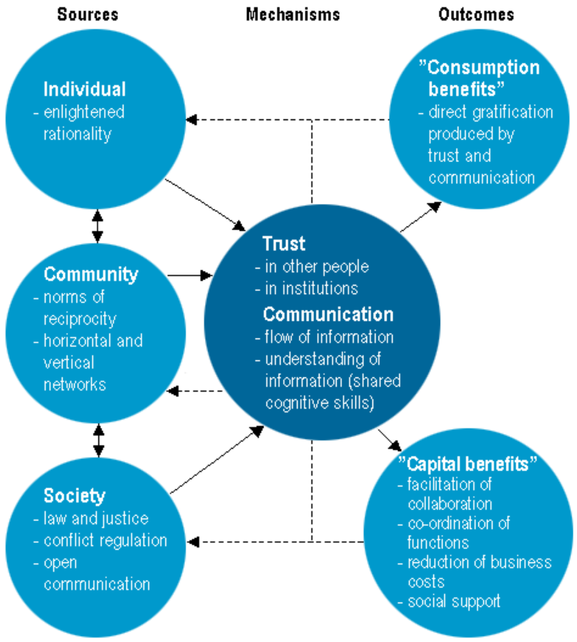
Trust as a key indicator of social capital

The concept of social capital has its roots in the notion that a proper understanding of welfare and the economic situation of society can only be achieved if the social dimension is also taken into account, i.e. society's capacity for collective action and the networks that support collective action

(Iisakka & Alanen 2006, 7). There has been a growing multidisciplinary debate on the concept of social capital, starting from 1980s and continuing into the 2000s. In spite of the interest, there is still no single, universal definition of social capital. However, it seems that an agreement is emerging according to which social capital is related, most particularly, to informal networks and norms (Halpern 2005, ref. Iisakka & Alanen 2006, 8).

Trust is often considered as a key indicator of social capital. Trust is considered either as a source or an outcome of social capital. In Finland, Petri Ruuskanen (2001) has proposed a distinction between the **sources, mechanisms and outcomes** of social capital, stressing the importance of keeping these dimensions separate in the measurement of social capital. This distinction is illustrated in Figure 1 below.

Figure 1. Sources, mechanisms and outcomes of social capital


Sources
Mechanisms
Outcomes
Individual
- enlightened rationality
Community
- norms of reciprocity
- horizontal and vertical networks
Society
- law and justice
- conflict regulation
- open communication
Trust
- in other people
- in institutions
Communication
- flow of information
- understanding of information (shared cognitive skills)
"Consumption benefits"
- direct gratification produced by trust and communication
"Capital benefits"
- facilitation of collaboration
- co-ordination of functions
- reduction of business costs
- social support

Source: Ruuskanen 2001.

The sources of social capital are considered separately at three different levels, i.e. the **individual, community and society**. The mechanisms of social capital, trust and communication facilitate the flow of information from one individual to another and make it easier for people to maintain contact with one another. According to Ruuskanen, both the sources and the outcomes of social capital are apparently context-dependent, whereas its mechanisms seem to work in the same way across different contexts. (Iisakka & Alanen 2006)

Statistics Finland's testing project of the problematic trust

In spite of the absence of a universal definition, the measurement of social capital has attracted much interest. It has been said that measurement of trust is the most difficult part when measuring the dimensions of social capital. Below is a quotation from the upcoming ESRA 2009 conference session description concerning 'Trust and Trustworthiness' (see http://surveymethodology.eu/conferences/warsaw-2009/sessions/72/) which illustrates the problem:

Citizen trust in social and political actors and institutions is currently a key area of concern for policy makers, social commentators and academic scholars around the world. Yet, despite the central position of trust in social and political theory and empirical research alike, its origins, conceptual status and socio-historical influence remain somewhat ambiguous. There is ambiguity about the meaning and validity of survey measures of trust. More importantly, there is confusion regarding the causes of varying levels of trust across individuals, time and societies and the consequences of declining trust for the effective functioning of social and political systems.

Measures of trust in the Finnish Leisure Survey

If we want to perform international comparisons of trust and social capital, we have to use already established and standardised questions (from e.g. WVS, ESS, ISSP). The questions of the Finnish Leisure Survey are formulated with the help of formulations used in international surveys such as WVS, ESS and ISSP.

Generalised trust concerns also other people than one already knows. According to Putnam (2000), generalised trust increases interest in taking care of common and shared interests and the probability of participating in voluntary work.
When constructing the 2002 Leisure Survey, researchers wanted to study diverse aspects of trust. They familiarised themselves with questions used in international surveys and came up with three statements, which also include the concept of mistrust:

Generalised trust: I can mostly be sure that other people want what is best for me.
If I am not careful, other people will take advantage of me.
People can generally be trusted.

(4-point scale from totally agrees to totally disagrees)

Kaj Ilmonen (2005, 60) criticizes the whole concept of generalised trust. He thinks it is conceptually vague and simplifies complex phenomena. According to Ilmonen, research results of generalised trust are almost impossible to interpret. Miller & Mitamura (2003) have observed that the survey questions measure differences in caution levels rather than in trust.

Informal trust is defined as trust in a particular group of people. Measures on informal trust are as follows:

Informal trust: There are only a few people in whom I can trust completely.
I trust most people living in my area.

However, it is not entirely clear how the respondents have understood the question. The statement used to measure informal trust – ‘There are only a few people in whom I can trust completely’ – proved somewhat problematic. It is difficult to know whether to interpret the statement in a positive or negative light, and the respondents have no doubt had the same difficulty when answering. Respondents may fully trust more than just a few other people. On the other hand they may not trust others at all.

The Leisure Survey included only one item on ***institutional or governmental trust*** (or confidence), which was worded as follows:

Institutional trust: A person like me does not have say in what the powers that be do.

Institutional trust or confidence had to be measured with just one question concerning confidence in government. Again, the statement – 'A person like me does not have say in what the powers that be do' – is open to different interpretations. The statement is problematic because we don't know if it actually measures trust in government or state, or trust in one's own ability to influence.

Blomqvist et al. (2005, 388) say that when measuring trust it quite often seems unclear what is actually being measured. Ahola (2000, 70) states that if the researcher is not aware of the nature of the information he or she requires, it is impossible to formulate relevant questions. Often questionnaires are designed by extracting questions from other surveys in which case the definition of their intention is given far too little consideration. However, this is the way trust is measured and international comparisons made.

About operationalisation

The principles guiding social research are sometimes forgotten in statistical social surveys. As social researchers know, the operationalisation of the concept is the most important part before formulating the questions. However, operationalisation itself is quite complex and a nuanced area in more ways than one. I think we all know what operationalisation means. You start with defining the research problem and move on to the theoretical concept you are interested in, then you undertake some research of the relevant literature, and proceed, little by little, from theoretical to concrete language and actual indicators and questions. It is, of course, important to describe the operationalisation performed, because it makes the measurement transparent. (Alkula et al. 1994)

Statistical analysis and measurements are often lacking in the analysis of so-called *process validity*, which refers to the best possible description of the all the relevant stages involving validity, that is, concept analysis, operationalisation and measurement. (Alkula et al. 1994)

When terms and concepts become established, the research techniques and operationalisation of variables become standardised. This is not necessarily a good thing, as it may also lead to the standardisation of originally erroneous, misleading and ambiguous conceptual definitions and operationalisations (Lehto 1996, 137). This might be the case with the measurement of trust, as there is so much research information which still needs clarification and interpretation and raises many questions.

About quality

In survey reports quality is often expressed in terms of sampling and response rate and questions of e.g. validity are given less attention. Methodological literature recommends that the success of operationalisation be examined with indicators describing validity and reliability (Lehto 1996, 137). However, validity is often defined rather unquantitatively and measuring it is not as simple as assumed. There are several types of validity and various ways of measuring it. Generally validity can be defined as a question of does the indicator measure what it is intended to measure. In cognitive interviews this is examined from the respondent's perspective.

Examining the success of operationalisation is, therefore, difficult to do in an unambiguous way. However, the thorough examination of concepts aims at this. If quality, as reported from the typical survey perspective, is reduced to an examination of sample, response rate and validity, a thorough testing of concepts can be used to produce broader reports of quality. Concept testing overlaps all these quality perceptions. *Substance quality* consists of conceptualisation. *Process quality* consists of the *description* of all the relevant stages involving validity (Alkula et al. 1994, 91). *Method quality* consists of good questionnaire design and operationalisation, which are possible only if good conceptualisation and pre-testing is made. This kind of analysis is needed also in *publication quality*, as it can provide an answer to the question 'what have we actually been measuring' by comparing the definitions used by respondents to the conceptual and operational definitions. This kind of information can be used when interpreting the statistical information, and it offers the user of the statistics an opportunity to assess and interpret the results. When all these aspects of quality are taken into account when publishing results or statistics, and they include the results of concept testing, we can reach the best possible process quality. This view of information includes the idea that the best information and quality do not consist of flawless 'factual information' but instead of justified and interpreted information and a transparent process.

How to test concepts?

The CASM (cognitive aspects of survey methodology) paradigm focuses on understanding the responding process. This includes also the understanding of the concepts used in the questions. Focus groups can also be used to examine the concepts used by respondents when discussing some phenomenon. Information obtained by concept mapping can thus be used when designing questions. Less has been talked about how to operationalise concepts and turn them into good questions. Thorough examination of concepts has been seen as the turf of substance researchers. Substance researchers, however, do not necessarily have experience of the cognitive factors influencing responses and, consequently, no real-life background in examining concepts from the perspective of the responding process and understanding concepts. In the following I develop a model which we intend to use in our thorough examination and testing of the concept of trust.

I am using Joop J. Hox's article *From Theoretical Concept to Survey Question* (1997) as a guideline in approaching the operationalisation of the concept of trust and outlining the general procedure of the research process of concept testing (Figure 2 illustrates the general outline of the concept testing procedure). Hox represents general strategies and specific techniques that can be used to structure the process of proceeding from theoretical concept to prototype survey question. Because I already have the questions which are meant to measure trust, I am also approaching the problem backwards: what have we actually been measuring? This kind of information can be used to formulate better questions and also in the interpretation of the statistical measures.

Hox (1997) notes that survey methodologists have paid much attention to issues of sampling and questionnaire construction, but have not dedicated as much effort to the steps that precede questionnaire construction, namely the clarification of the research objectives and the **elaboration** of the theoretical concepts that are to be translated into questions. (Hox states that the development of concepts for social research is usually placed in the context of discovery, and not in the context of verification. There are no fixed rules that limit the researcher's imagination and concepts are not rigorously verified or falsified, but they are judged by their fruitfulness for the research process.)

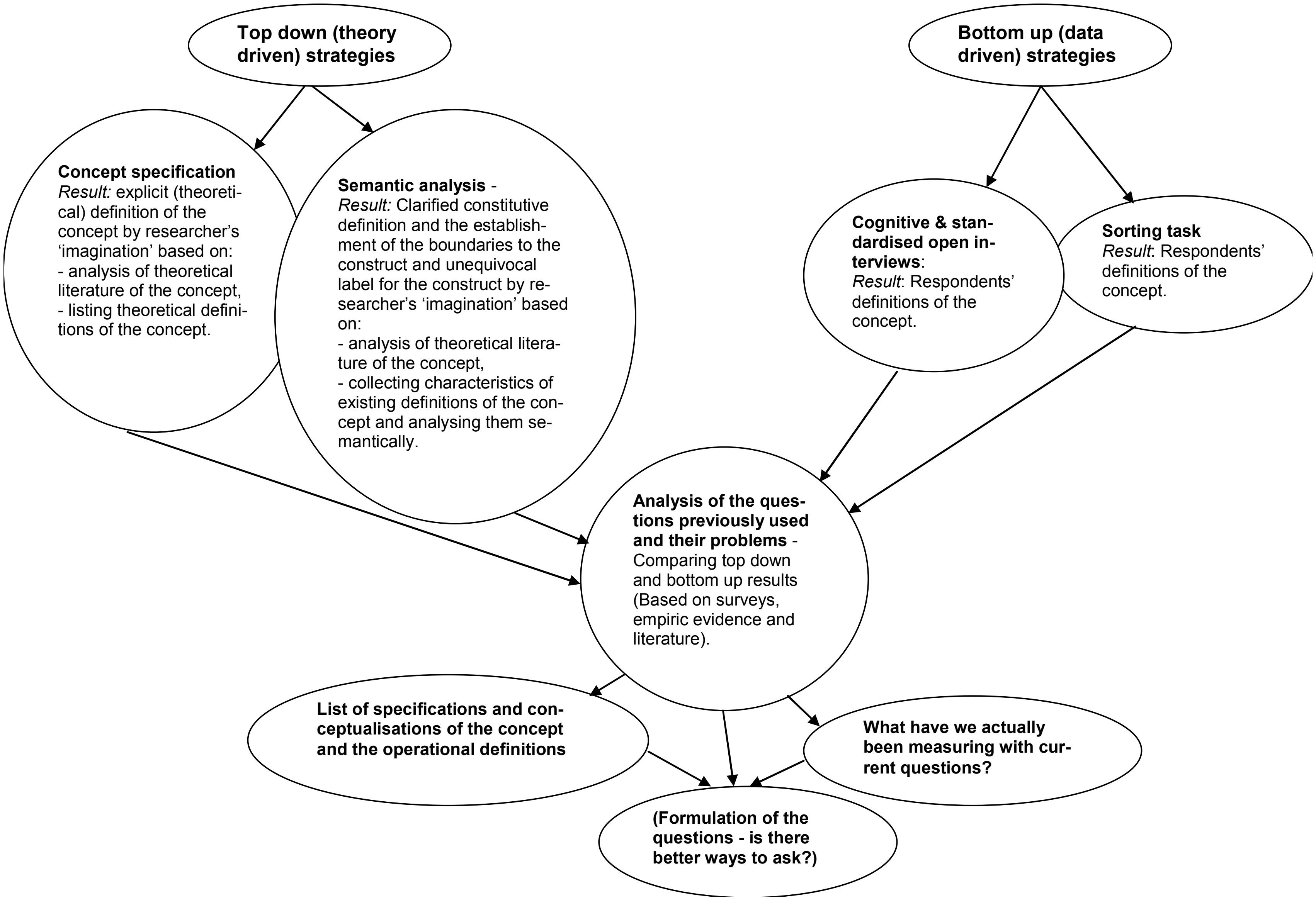
Hox clarifies the differences between the terms *concept, construct* and *variable*. The difference between a concept and a construct is small. A concept is an abstraction formed by generalisation from similar phenomena or similar attributes. A construct is a concept that is systematically defined to be used in scientific theory. A variable is a term or symbol to which values are assigned based on empirical observations, according to indisputable rules. (Hox 1997, 49.)

Following Hox, I use these terms loosely when constructing trust. Both concept and constructs must be linked to observed variables via an operational definition that specifies to which variables they are linked and how values are assigned to these variables.

Hox states that methodologists act as if there were a clear distinction between the theoretical language and the observational language. This leads to the sharp distinction between conceptualisation and operationalisation. Conceptualisation involves concept formation, which establishes the meaning of a construct by elaborating the nomological network and defining important subdomains of its meaning. Operationalisation involves the translation of a theoretical construct into an observable variable by specifying empirical indicators for the concept and its subdomains. To bridge this gap between theory and measurement, Hox represents different strategies. “Top down” or theory driven strategy starts with constructs and works towards variables and data driven or “bottom up” strategy starts with observations and works towards constructs.

In this process of researching the concept of trust I am going to adopt both these strategies in some way or another. I will also use some other techniques not necessarily mentioned by Hox.

Figure 2. Outline of the research process used in the testing of concepts


Top down (theory driven) strategies
Bottom up (data driven) strategies
Concept specification Result: explicit (theoretical) definition of the concept by researcher's 'imagination' based on: - analysis of theoretical literature of the concept, - listing theoretical definitions of the concept.
Semantic analysis - Result: Clarified constitutive definition and the establishment of the boundaries to the construct and unequivocal label for the construct by researcher's 'imagination' based on: - analysis of theoretical literature of the concept, - collecting characteristics of existing definitions of the concept and analysing them semantically.
Cognitive & standardised open interviews: Result: Respondents' definitions of the concept.
Sorting task Result: Respondents' definitions of the concept.
Analysis of the questions previously used and their problems - Comparing top down and bottom up results (Based on surveys, empiric evidence and literature).
List of specifications and conceptualisations of the concept and the operational definitions
What have we actually been measuring with current questions?
(Formulation of the questions - is there better ways to ask?)

1) Top down procedures:
a. Concept specification and
b. Semantic analysis of the concept of trust

Hox quotes Lazarsfeld (1958, 1972) and Fiske (1971) and their description of the process called *concept specification.* At first, in this procedure (according to Fiske), the theoretical context of the concept must be identified. The researcher's task here is to clarify the theoretical background and the way the concept is extended to fit a specific research problem. The concept of trust is very theory-laden. Literature about the concept of trust can be found quite easily. In my research project I will list and analyse the different definitions given to the concept of trust. The second step is to delineate the core of the construct, the unique quality to which construct refers. This means explicitly defining the core meaning of the construct and its subconstructs. This also means defining what is not included in the concept. The third step is to construct a measurement instrument. This procedure may be derived logically by using the researcher's own imagination or it may be based on the empirical results from previous research. (In my final analysis, I may construct a hypothetical measurement instrument based on all the strategies I have used.)

Hox also quotes Sartori (1984), and says that conceptualisation is mediated by language, which makes it important to study the semantic structure of our statements. Sartori states that first we have the semantic *connotation* of the construct, which means the associations it has in the minds of the users or the list of all characteristics included in the construct. Secondly, we have the semantic *denotation* of the construct, for instance the listing the set of objects to which it refers. These two definitions are the same as constitutive and operational definition, terms used by social scientists. Connotation can be problematic. There can be a confusion of meanings because the construct is associated with more than one meaning, or two constructs point to the same meaning, and researchers have not made it explicit to which one(s) they intend to point. There may also be problems with the denotation: it may be vague to which objects or referents the construct applies. There may also be terminological problems: we may have chosen a label for the construct that leads us to refer to the wrong characteristics or referents.

To proceed with the semantic analysis, we must at first clarify the constitutive definition of the construct. This means assessing the *defining* (essential) characteristics of a construct, as opposed to the *accompanying* (variable) characteristics. The constitutive definition should be adequate and parsimonious. (This means collecting characteristics of existing connotative definitions, e.g. dictionary, previous research, and abstract a common core from these.) Checkpoints can be homonyms and synonyms. The next step is determining the empirical referents of the construct. This is not the same as providing an operational definition, because this may include empirical referents for which we do not have a measurement operation (yet). This means the establishment of boundaries and the decision which empirical referents belong to the construct (and which do not). If the boundaries are difficult to establish, it usually means deficiency in the constitutive definition. The third step is to make sure that the verbal label for the construct is understood unequivocally. One must consider the level of abstractness of the chosen term. One may do a substitution test: if in constitutive definition, a word can be substituted by another word with a gain in clarity or precision, the first word is being misused. Semantic analysis doesn't directly lead to survey questions, but it helps to disentangle different meanings and to recognize ambiguity in our constructs.

2) Data Driven Approaches (Bottom up)
a. Symbolic interactionism - Pre(and post)-test(s interviews)
b. Concept mapping

a. In colloquial every-day language, trust is a commonly known concept. Often concepts used in social sciences are extracted from every-day language. To make it more useful in scientific theory, researchers have to attach a more precise meaning to such a concept. Still, if we want to describe and explain the experiences and behaviours of individuals in society, we should understand how these concepts are used and understood in social life. Symbolic interactionism is based on Blumer's (1969) thoughts and it views the social world as a world of interacting individuals who are constantly negotiating a shared meaning of the interaction situation. This type of qualitative research focuses on the processes of interaction and interpretation. This can be a sort of field research but it also comes close to the method of open and cognitive interviews. With cognitive interviews we can investigate the understanding of previously used survey questions. With thematic interviews we can investigate more thoroughly the understanding of the concept of trust, types of trust and related concepts. Those can then be analyzed with the aim of reducing the set of concepts (which the respondents used) to a smaller set of key concepts, which are then integrated in an overall theory by establishing their relationships. In this project I plan to do a minimum of 15 interviews with standardised open questions and cognitive interviews. I will use spontaneous probes in addition to standardised probes. I will also analyze the data by sorting out the conceptualisations associated with the construct and investigate the relation of these respondent conceptualizations with definitions in social capital theory, empirical research and the questions previously used.
The interview protocol is in Appendix 1.

b. Concept mapping is a faster and cheaper technique to research respondents' thoughts on the concept than interviews. We can, for instance, use focus groups in generating statements related to the concept. Then the statement list is structured by researchers. After this the respondents are asked, for instance, to pair up statements or to group them into different clusters of mutually similar statements. There are a variety of different sorting tasks; one is to ask a respondent to split statements to natural groups and to explain how they differ from each other, and so on. Hox proposes that concept statements can be translated into survey questions quite easily and a concept map can provide a hypothesis to the empirical structure of the questionnaire.

Conclusion about the use of the testing procedures

In this testing project of trust I am going to use the above-described top down procedures as well as the bottom up procedures. We are not going to use focus groups because of the busy schedule and lack of resources over the summer. Due to scheduling reasons, we will also do the cognitive interviews and more open explorative interviews at the same time. I will also list the survey questions previously used in measuring trust from international and national surveys. The questions will be analysed with the help of the results from the top down and bottom up analysis and the problems of the previously used questions will be listed.

How to perform the interviews?

The interview protocol that we are going to use is quite normal for the Finnish Cognitive Laboratory testing. We use concurrent and retrospective think-alouds. In this testing we are going to first use a general think-aloud encouragement and then some verbal probes. An example from the cognitive protocol is given in Figure 3 below.

Figure 3. An example from the cognitive interviewing protocol

I can mostly be sure that other people want what is best for me.

General think-aloud encouragement:

– What did you think when you heard the question?

Verbal probes:

– What did you think the question means in this context? What did you think the question means, precisely?
– Who did you think were the ‘other people’ mentioned in the question?
– What does the term ‘mostly’ mean to you in this context?
– What does the expression‘want what is best for me’ mean to you in this context?
– How did you arrive at your answer (and that response alternative)?
– Was it easy or hard to answer the question? Why?

The thematic explorative interview is a bit more difficult; one cannot ask the respondent to define the concept of trust straightforwardly. Figure 4 below gives an example of the explorative interviewing protocol. This interview will be made straight after the cognitive testing. In addition, we are going to prime the respondents to think about their own trust-related behaviour e.g. by asking if they would leave their jacket/bag unattended in restaurant/school, etc, when they have to leave the room for a short time to visit another room/the toilet, etc. We are going to invent a few hypothetical situations and discuss them with the respondents.

Figure 4. An example from the thematic explorative interviewing protocol

I have now asked you several questions intended to measure trust. Now we could discuss more informally the thoughts that came to your mind when I was asking the questions.

In general, what kinds of thoughts did you think when I asked the questions?

Do you think you are a trusting person? What things are you trusting about and what not?

What kinds of characteristics do think trustfulness includes? Do you consider trustfulness a bad or a good characteristic?

Do you think that trust is a personal trait or do you think it depends on external things?

Do you have anything else in mind about the questions you were asked regarding trust? Why do you think these kinds of studies are made and these kinds of questions asked?

Analysis of the data from conceptual testing

The final analysis of concept testing is done piece by piece in a kind of content analytic way by classifying and categorizing the data. Then the conceptualisation of trust is compiled into a table. As Hox (1997) states, the various approaches do not directly produce survey questions. Instead, we end up with list of specifications and conceptualisations. After this the iterative process of testing questions and devising better measures should continue.

It is also important to answer the question ‘what is actually being measured when we measure trust’ with the analysis of the data obtained with these strategies. This kind of information can be used in interpretation of the statistical measures.

Evaluation of concept testing - Challenges

Concept testing includes certain challenges which should be taken into account. Concept testing is a time-consuming and costly process and requires much work by the researchers. The interviews require special training or must be performed by the researchers. It must be considered how the information obtained from concept testing can be translated into practice and how it can be included in standardised measures and statistical standards. It is important to remember that the results from concept testing can be used in analysing and interpreting statistical information. Conceptual testing could lead to the finding that the current questions are not valid. However, altering question formulations is not simple. Changes interrupt time series and comparability is lost. In such situations the question formulations must be reconsidered and they should e.g. be used alongside old formulations for a while. Concept testing is also useless if the results for 'what have we actually been measuring' do not reach the users of the statistics. The challenge is how the information from concept testing is made available and known to the users of the statistics? This also involves the perspective of process validity, i.e. how to make the entire statistics production process more transparent and thus easier for the users to interpret and assess.

References:

Ahola, Anja (2000) Surveykysymysten tarkoitus ja vastausten tulkittavuus *(The Meaning of survey questions and interpretability of the answers*). In *The Welfare Review 1/2000*. [In Finnish only]. Statistics Finland, Helsinki.

Alkula, Tapani; Pöntinen, Seppo & Ylöstalo, Pekka (1994) Sosiaalitutkimuksen kvantitatiiviset menetelmät (*Quantitative methods in social research)* [In Finnish Only]. WSOY, Porvoo.

Blomqvist, Kirsimarja; Seppänen, Risto & Sundqvist, Sanna (2005) Organisaatioiden välisen luottamuksen mittaaminen - analyyttinen katsaus vuosien 1990-2003 empiiriseen tutkimukseen. *(Measuring trust between organisations - analytical review of empirical studies in 1990-2003).* [In Finnish only] In Pertti Jokivuori (ed.). Sosiaalisen pääoman kentät *(Fields of social capital).* Minerva Kustannus Oy, Jyväskylä, pp 378-392.

Hox, Joop J. (1997) From Theoretical Concept to Survey Question. In Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin (eds.): *Survey Measurement and Process Quality*. John Wiley & Sons, Inc., New York.

Iisakka, Laura & Alanen, Aku (2006) Social capital in Finland: domestic and international background. In Iisakka (ed.): *Social Capital in Finland – Statistical Review.* Statistics Finland, Helsinki, pp. 23-32.

Ilmonen, Kaj (2005) Luottamuksen operationalisoinnista *(Operationalisation of trust)* [In Finnish only]. In Pertti Jokivuori (ed.). Fields of Social Capital *(Sosiaalisen pääoman kentät)* Minerva Kustannus Oy, Jyväskylä, pp. 45-68.

Lehto, Anna-Maija (1996) Työolot tutkimuskohteena: työolotutkimusten sisällöllistä ja menetelmällistä arviointia yhteiskuntatieteen ja naistutkimuksen näkökulmasta *(Working conditions as a research subject: a conceptual and methodological evaluation of Quality of Work Life surveys from the social and feminist research perspectives)*. [In Finnish only]. Research Reports 222, Statistics Finland, Helsinki.

Miller, Alan S. & Mitamura, Tomoko (2003) Are Surveys on Trust Trustworthy? In Social Psychology Quarterly, Vol. 66, No. 1, pp 62-70.

Putnam, Robert D. (2000) Bowling Alone: The Collapse and Revival of American Community. New York, Simon & Schuster.

Ruuskanen, Petri (2001) Sosiaalinen pääoma – käsitteet, suuntaukset ja mekanismit *(Social capital - concepts, trends and mechanisms)* [in Finnish only]. VATT Studies 81. Government Institute for Economic Research, Helsinki.

ABSTRACT

QUEST MEETING 2009

Using split ballot testing to evaluate findings of how to best measure health knowledge questions

Carol Cosenza
Center for Survey Research
University of Massachusetts Boston

There are many formats a researcher can use to measure health knowledge. Offering respondents answer choices in a closed ended format has potential pitfalls. A respondent could simply guess and have a relatively good chance of choosing the correct answer. Whatever options the researcher includes provides the respondent with some clues and the potential to rule out answers that are wrong. For example, if the answer options to a question are “1-5, 6-10, 11-20, 20 or more”, the respondent knows that the answer is probably a relatively small number and has at least a 25% chance of just guessing the correct answer. However, respondents who are asked the question in an open-ended format are given no boundaries or clues and therefore have the ability to answer whatever they want.

This paper will examine data collected from a series of split ballot experiments with questions that measure respondents’ perception of health risk (such as being diagnosed with breast cancer or dying from colorectal cancer). We will compare and contrast two formats for asking knowledge questions. Half of the respondents will be asked a question in an open-ended format (for example: Out of every 100 people, about how many will get colon cancer?). The other half of the respondents will be asked the question in a closed ended format (Out of every 100 people, about how many will get colon cancer? Please choose the number that you think is closest to the correct answer: 2, 6, 14, 24, or 43). The results be compared with respect to how many provide a “correct” answer and the rates at which respondents provide any answer at all. The implications for how to measure knowledge will be discussed.

Using Multiple Methods to Understand Ethnicity Response

Lyn Kaye: Statistics New Zealand

This paper emphasises the influence of question design in shaping answers to survey questions and the advantages of using multiple methods to research and evaluate changes in data. Using an example from New Zealand, it presents the findings from a programme of research initiated to understand why 11% of the New Zealand population now elect to describe themselves as 'New Zealanders' in response to a question on their ethnicity in New Zealand's Census of Population and Dwellings. A growing number of respondents are opting to write in this description under the category of 'other', rather than select from the traditional ethnic groups named in the response list. Results are presented from several research projects undertaken to date, including inter-censal matching, focus group testing and cognitive interviewing which aimed to explain this change in response and test alternative questions. Those findings show that question wording and response order can interact with other factors to contribute to shifts in data. Finally, some conclusions about the ways in which multiple research methods can inform and enrich our understanding of respondent behaviour are discussed.

1. Introduction

The wording, order and visual design of survey questions are known to have significant impacts on survey response. However these influences are sometimes overlooked in the assessment and interpretation of survey results. Using Statistics New Zealand's ethnicity question as an example, this paper discusses the way in which multiple research methods, and in particular qualitative research methods, such as cognitive interviewing, are useful to fully understand the patterns we see in survey data and the ways that question wording may influence those results.

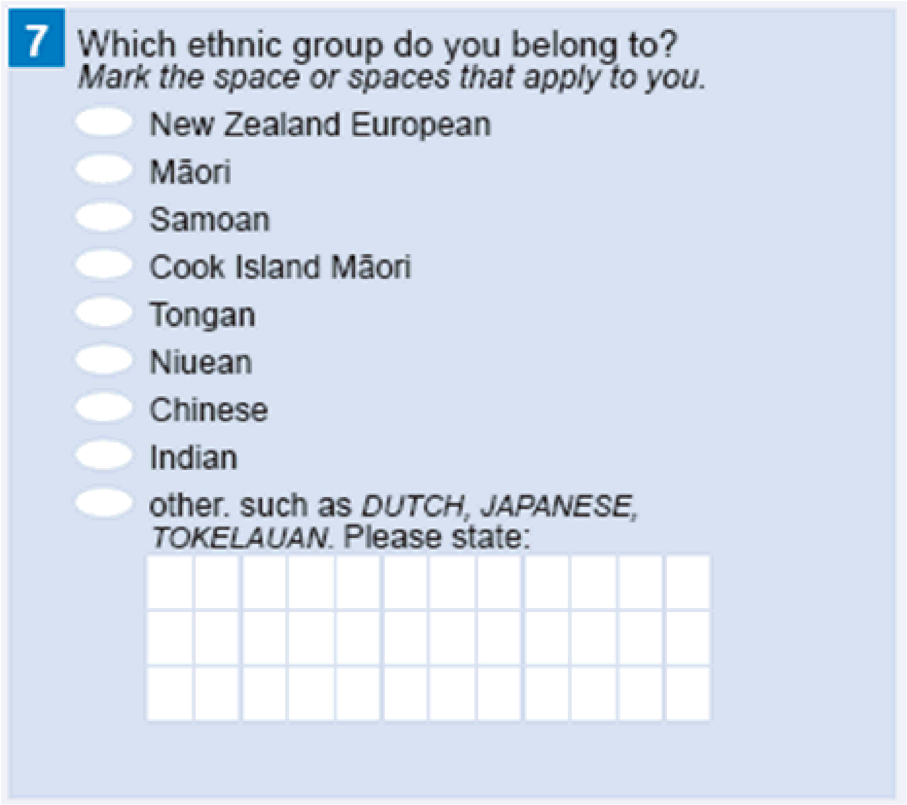
New Zealand's Census of Population and Dwellings has included a question to determine the ethnic composition of the population since the middle of the 19th century. This ethnicity data is collected and used widely for a variety of purposes. As a basic socio-economic descriptor, it is used extensively by government, non-government, and individual researchers. It is a mandatory variable in New Zealand's five-yearly census and in combination with other characteristics of the respondent, ethnicity data is used to monitor the well-being of ethnic groups and to inform research and policy development generally (Statistics New Zealand, 2003).

The ethnicity question itself has developed and evolved across time. The current version of the question is shown in Figure 1. Eight ethnicities are listed as response options along with write-in space for 'other' ethnicities. These eight options have varied across the years, but have consisted of a mix of the most predominant ethnic groups and some smaller groups indigenous to the pacific region.

However, in recent years there has been a growing proportion of the population who select the 'other' category and write in "New Zealander", rather than select from any of the response options listed.   In 2001, for example, 2.5% wrote in 'New Zealander' and by 2006 this figure had grown to 11%. This increase has brought with it a growing call for Statistics NZ to provide a specific 'New Zealander' tickbox to acknowledge and accommodate changing perceptions of ethnicity. Those with this view argue that 'New Zealander' is a legitimate response, reflecting an evolving ethnic identity within New Zealand. This issue has generated public debate and attracted media interest nationally. However, others are concerned that this will dilute the utility of ethnicity statistics and reduce the accuracy and relevance of those statistics as an information source. Data users are particularly concerned about the impact on time-series data and their ability to make comparisons across time.

To respond to these concerns, Statistics New Zealand initiated a research programme to identify and understand the issues impacting on ethnicity responses. The aims of this research programme were to gain a better understanding of the reasons why respondents elect to describe themselves as a 'New Zealander' and to evaluate the likely impacts on data if the question were changed, or if additional questions were included.

To answer these questions, multiple research methods were used and an integrated programme of research was initiated. This programme included inter-censal matching, focus groups, and cognitive interviewing. This paper briefly discusses results from each of these methods in the following sections. Explanations for these findings are explored further in the discussion section. Final conclusions and further thoughts on the ways that multiple methods were able to enhance our knowledge of this issue are reserved for the final section of this paper.



7 Which ethnic group do you belong to?
Mark the space or spaces that apply to you.

○ New Zealand European
○ Māori
○ Samoan
○ Cook Island Māori
○ Tongan
○ Niuean
○ Chinese
○ Indian
○ other. such as *DUTCH, JAPANESE, TOKELAUAN*. Please state:

Figure 1: NZ Ethnicity Question (2006)

2. Inter-censal Consistency Study

To understand the shift in ethnicity statistics between 2001 and 2006, and to identify where the increase in those describing themselves as ‘New Zealanders’ was coming from, an inter-censal linking exercise was undertaken. A full report of this study can be found elsewhere (Statistics New Zealand, 2009), and key points are only summarised here.

2.1 Method

A sample of respondents describing themselves as ‘New Zealanders’ in Census 2006 were linked back to a matched sample from Census 2001, using probabilistic matching techniques. Those samples were linked on three key variables - sex, date of birth and geographic location. The final dataset was made up of 72.6% of the total number of people describing themselves as ‘New Zealander’ in 2006. .

2.2 Results

Results from this study showed that most of the growth in 'New Zealander' responses between 2001 and 2006 came from respondents who had previously responded as 'NZ European' in the 2001 Census. This group accounted for 92% of those describing themselves as 'New Zealander' in 2006, a figure that is moderately higher than the proportion of ‘NZ Europeans’ in the general population. Although a small proportion of these respondents ticked both ‘New Zealander’ and ‘NZ European’, the large majority moved to a description of 'New Zealander' only.

However, this study further showed that 8% of respondents who provided a ‘New Zealander’ response in 2006 came from other ethnic groups, including Maori, Pacific and Asian. These flows resulted in a loss to the overall population counts for these groups of between 0.9 - 2.0%

3. Focus groups

Focus groups were conducted to explore the concept of ethnicity and ways in which respondents think about this topic. Through the dynamics of group discussion, key themes and issues relating to the ‘New Zealander’ response could be examined. This part of the research was undertaken by an independent research organization and detailed results can be found in the full report (UMR, 2009).

3.1 Method

Ten focus groups were conducted between 18-24 March. Volunteers, recruited over the phone, were allocated to groups which were defined by ethnicity and the number of generations they had been living in New Zealand. The duration of each group was between one and a half to two hours. There were between 4-6 participants of mixed age, gender and income in each group.

A broad question schedule was prepared. As part of that schedule, respondents were asked to write down their understanding of the term ‘ethnicity’ and various ethnic labels before joining together to discuss those concepts collectively as a group.

3.2 Results

Focus group testing revealed a range of views on this topic. Participants conceived of ethnicity as a rich concept that had many facets. However participants often gave different weighting to different facets. Those most frequently mentioned in association with this topic were ancestry and race.

Many participants recognised that ethnicity was subjective, flexible and based on respondents' own self-perceptions. However, as discussion progressed, participants were more likely to mention objective and stable markers of ethnicity, such as race and ancestry. This revealed the complex and sometimes contradictory ways that respondents thought about this topic.

In focus group testing, most participants did not support a 'New Zealander' category as an ethnic description and saw this as a term which was more appropriate as a description of a person's national identity rather than their ethnicity.

However, this research also identified a number of key respondent 'types' who were likely to describe themselves as 'New Zealanders'. These respondents could be divided into three broad groups:

- those who believed that NZ had developed its own unique culture
- those who did not feel that the description 'NZ European' was relevant to them, as they felt no affiliation with Europe.
- those who felt that ethnicity statistics were used for discriminatory purposes. This group was divided between respondents who perceived the question as socially divisive and preferred to view the population as one, and others who felt it was used to distribute public funding unfairly between groups.

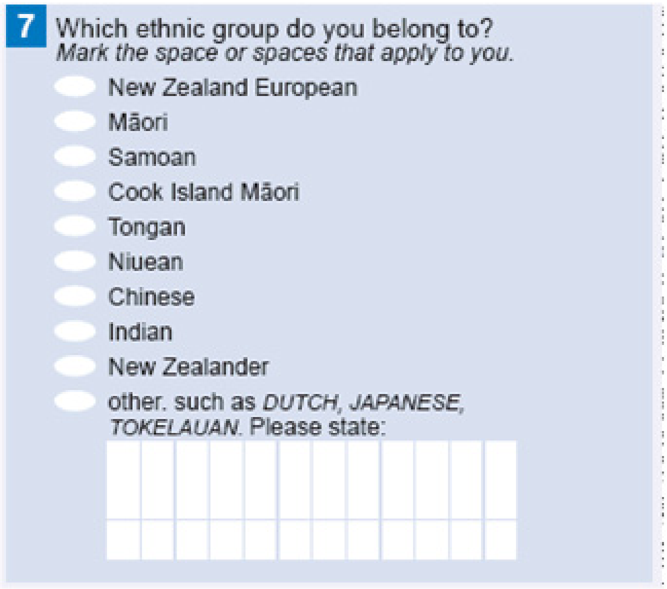
However, these were not discrete groups and there was some overlap between them.

4. Cognitive Interviewing

To examine the ways that an explicit 'New Zealander' response option would influence answers, we undertook cognitive interviewing. Cognitive interviewing was a way to explore, understand and explain the ways that people answered questions about their ethnicity. This involved one-on-one interviews with respondents as they filled out a questionnaire and included retrospective probing to discover how respondents understood key terms and how they had decided which answer to select.

4.1 Method

Almost 100 cognitive interviews were conducted across the duration of the project (October 2008 - May 2009). Participants for this research were volunteers who were recruited through a variety of methods, including personal networks. They were given a gift voucher to the value of $20 as a thank you for their participation. To compare how respondents answered a version of the ethnicity question which had been adapted to include a 'New Zealander' response, with the way they answered the question without that option, respondents were asked to complete two question sets. In one set, the ethnicity question included a 'New Zealander' response (Figure 2) and in the other set, the question did not. In every other way, the two questions were identical. To minimise the influence of the question completed first on the question completed second, question sets was administered in reverse order for half of the interviews. However, because this was qualitative research, impacts of order could not be ruled out.

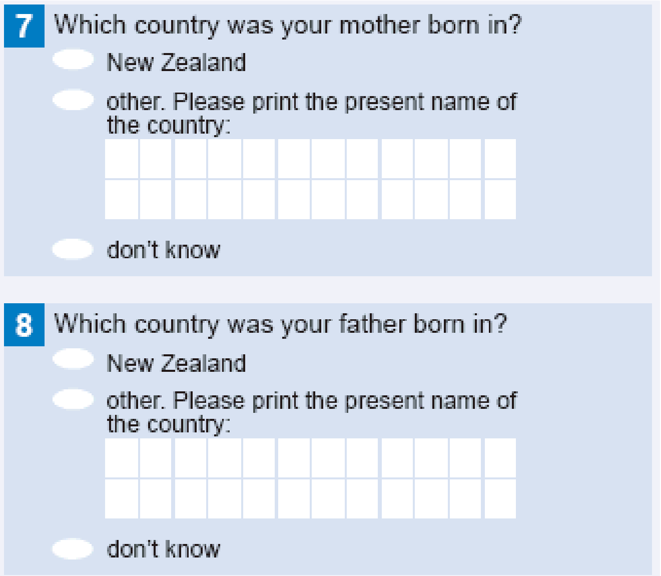


7 Which ethnic group do you belong to?
Mark the space or spaces that apply to you.

○ New Zealand European
○ Māori
○ Samoan
○ Cook Island Māori
○ Tongan
○ Niuean
○ Chinese
○ Indian
○ New Zealander
○ other, such as *DUTCH, JAPANESE, TOKELAUAN*. Please state:

Figure 2: Altered version of the Census 2006 ethnicity question

Cognitive interviewing also examined the theory that respondents may be inclined to write in 'New Zealander' as a way to identify their generational links to the country. To explore this idea, half of the respondents in this sample were also asked to complete two additional questions asking them to name their parents' birthplace. The aim of these additional questions was to provide an alternative way that respondents could identify their generational attachment to New Zealand, and to examine whether this reduced respondents' inclination to tick 'New Zealander' as an ethnicity. These extra questions were placed immediately before the ethnicity question in the altered version of the questionnaire (with a 'New Zealander' tickbox). These questions are referred to as the 'generational attachment' questions (see Figure 3).



7 Which country was your mother born in?

○ New Zealand
○ other. Please print the present name of the country:
○ don't know

8 Which country was your father born in?

○ New Zealand
○ other. Please print the present name of the country:
○ don't know

Figure 3. Additional questions to measure 'generational attachment'

After all of the questions were completed, interviewers asked follow-up questions to further explore respondents' perceptions of the ethnicity question. A protocol was used for this purpose. This protocol included a script of the core probes to be used during the interviews and acted as a guide for interviewers. However, the actual question line remained flexible so that interviewers had the freedom to respond to unique issues that arose. Interviews were recorded and reports were subsequently written to document the findings from each interview. Key themes were identified from an analysis of those reports.

4.2 Results

In cognitive interviews respondents generally had a good understanding of the intent of the ethnicity question and typically found it relatively straight-forward and easy to answer. Respondents seemed to recognise the ethnicity question as one they had seen on other forms and surveys. They often answered quite quickly and didn't usually spend a great deal of time thinking about the question before selecting their answer. However, respondents sometimes gave different answers to each of the two ethnicity questions they completed, depending on the response list presented to them.

4.2.1 Comprehension

Concepts of ethnicity were described in ways which closely aligned with the focus group testing and there were a range of views on this topic. Respondents sometimes expressed their discomfort with distinctions based on ethnicity and felt that the term 'New Zealander' was an inclusive term that cut across ethnic distinctions. However, others questioned the legitimacy of this response as an ethnic description. These respondents held the view that ethnicity statistics were only valuable if they were able to make distinctions between various groups of the population.

Interpretations of the term 'New Zealander' tended to vary considerably. Some respondents used this description as a way to indicate that they could trace their ancestry back several generations within New Zealand. However, participants with parents who had been born elsewhere, but who had themselves lived in New Zealand all of their lives, also used this description to show their attachment and belonging to the country. In addition, immigrants who had been born elsewhere but had lived in the country for many years also liked to describe themselves as 'New Zealanders', reflecting their sense of immersion into the culture.

Some respondents expressed a dislike and a disassociation with the term 'NZ European', believing this to be a poor description of their ethnicity. This view was associated with a belief that the New Zealand population was now sufficiently different to other populations that they warranted recognition as a separate ethnic group. One respondent's interpretation of the term 'New Zealander' provides an example. This comment alludes to a unique and distinctive culture within New Zealand, and the influence that culture can have in shaping identity.

"Being born here, having had your family be here for more than two generations, you like L&P and pies and gumboots and pineapple lumps ... that makes you a New Zealander. So it's just all of the things that we stand for."

Although a notable number of respondents expressed some discomfort with ethnic distinctions, this discomfort was often motivated by different things. Some respondents disliked the way that those distinctions were used to marginalize minority groups, and preferred to use more inclusive terms. In contrast others disliked the way that ethnicity statistics were used to allocate state funding to minority groups, and perceived this to be unfair and exclusive. For these respondents the 'New Zealander' label appeared to be a way of 'opting out' or protesting against the ways that ethnicity data was used.

Whether or not respondents agreed with the idea that 'New Zealander' was a valid description of ethnicity, it was interesting that many tended to prefer the 'New Zealander' label when they they thought that they had to make a choice. For these respondents a sense of national identity appeared to be stronger than their sense of ethnic identity, as the following quote illustrates:

"NZ European… felt that if I selected that, it would be like saying I wasn't a New Zealander, which I am. I would rather say I'm not a NZ European than not a New Zealander."

4.2.2 Reading and responding to the ethnicity question

When respondents answered the ethnicity question that did not have an explicit 'New Zealander' response option, they were normally satisfied to select one of the options listed. It was unusual for respondents to consciously seek out a description that was different to those listed. Even respondents who held a strong preference to be described as a 'New Zealander' occasionally selected existing options, rather than write in that response.

In both versions of the question, respondents often skimmed over the response categories and once they had identified an adequate answer, they would commonly fail to read and consider the remaining options. When answering the version of the question which included a 'New Zealander' response this tendency to skip to the next question immediately after making their selection meant that some respondents, who subsequently declared their preference for 'New Zealander', did not mark that option when it was available to them.

Many respondents also failed to realise that more than one answer could be given. This meant that even if respondents read all of the options available in the list, they were often unaware that they could select more than one response and therefore incorrectly felt that they needed to make a choice between the 'New Zealander' option and another suitable description of their ethnicity. In this situation respondents typically stated a preference for the 'New Zealander' option.

In summary, the design of the ethnicity question had a significant influence on final answers. Respondents often read the question poorly, and failed to realise that they could tick more than one response. These respondents also failed to read and consider a 'New Zealander' category appearing last in the response list. However, those that noticed and read the 'New Zealander' option felt confused because they believed that they needed to make a choice between that option and another suitable description of their ethnicity.

4.2.3 Generational attachment questions

Evaluation of the generational attachment questions (see Figure 3) was exploratory only and numbers were too small to draw any firm conclusions. However, our results did not reveal any distinct difference between those who got the additional questions on parents' birthplace and those that did not. Our observations during interviews indicated that there was no perceptible impact on respondents' answers to the ethnicity question and respondents' own self reports suggested that they were not influenced by the extra questions.

5.0 Discussion

Various versions of the New Zealand ethnicity question have been tested over many censuses. The current question is relatively short, uses simple language structure and is seemingly straight-forward. In general it works reasonably well. However, this research showed that the addition of a 'New Zealander' tickbox introduces a complexity that confuses response in a variety of ways. We observed a number of measurement problems in testing this question, some of which related specifically to the question structure, and others which related more directly to respondents' own motivations and interpretations of the question's purpose. These effects combined in complex and contradictory ways to influence respondents answers.

During general discussion, respondents sometimes expressed their dislike of ethnic distinctions and a general discomfort with the ethnicity question. However, in practice most respondents were willing to select a response category from the list, regardless of which version they were completing. Interestingly, even those respondents who stated a preference for the description 'New Zealander' usually chose to tick a listed option, rather than write in 'New Zealander' when it was not listed. This is an important finding, as it shows that most respondents are willing to cooperate with the researcher by providing an answer that meets the researcher's expectations. This illustrates the principle of 'cooperativeness', a central assumption that underpins communication (Grice, 1975) and replicates documented research which shows that 'other' responses are commonly underreported (eg. Schuman and Presser 1981).

Schwarz and Hippler (1992) apply Grice's theories of communication and the cooperativeness principle to explain that underreporting. They suggest that response categories play a powerful part in shaping respondents' answers because they define the information that researchers are most interested in. Therefore respondents may discard or revise information that is not listed in the response options because they perceive it to be irrelevant to the purpose of the enquiry.

This also suggests that respondents electing to write in 'New Zealander' when this category does not appear on the list, may be doing this as a way to protest against the researcher's intentions and visibly 'opt out' of the process. This observation helps to remind us that respondents are not passive recipients of survey questions and that a survey interview, even in a written format, is part of a social exchange in which the respondent takes an active part (Clark & Schober, 1992).

However, the more common tendency to cooperate with the researcher becomes problematic when a 'New Zealander' response is offered to respondents. Even those who are normally happy to describe their ethnicity in more specific terms, and may question the legitimacy of a 'New Zealander' response as a description of ethnicity, nevertheless tick the 'New Zealander' option when it is included in the list. For most respondents 'New Zealander' is a description of their national identity and it seems accurate that they should tick this option. Respondents consult the response categories to understand the question and to make some assumptions about the researcher's intended meaning. When response options differ from the assumptions held by the respondent, they subsequently adjust their answer to fit the new information (Sudman, Bradburn & Schwarz, 1996). Therefore respondents reading the 'New Zealander' option assume that national identity is a dimension of interest to the researcher and adjust their answers to conform to that expectation.

A complimentary theory, offered by cognitive psychology, suggests that respondents adjust their answers to fit the response categories as part of the 'editing' or 'formatting' phase of the question and answer process. This is one of four cognitive steps the respondent engages in to provide an answer (eg. Tourangeau, Rips & Ransinski, 2000).

Both theories would suggest that adding a ‘New Zealander’ tickbox to the ethnicity question can change and influence a respondents’ answer, by attracting response, not only from respondents who actively wish to describe themselves as a ‘New Zealander’, but from respondents who would normally be happy to select a more specific description.

However, whatever preference respondents had, many did not see the ‘New Zealander’ response appearing last on the list and therefore didn’t endorse this option when it was available. This is because respondents commonly selected the first option available to them and failed to read and consider the remaining options (a ‘primacy effect’). This finding is not surprising when many respondents belong to a single ethnic group and do not expect to mark more than one response. However, this meant that those who might otherwise have selected a ‘New Zealander’ category often did not see that option before selecting another suitable answer.

Primacy effects in the marking of self-complete questionnaires are well recorded (Dillman, 2002). Such behaviours have been explained by Krosnick and Alwyn (1987) as examples of ‘satisficing’. This theory suggests that respondents will put in just enough effort to select an answer that seems appropriate before moving to the next question.

Theories of satisficing also help explain why, despite being a multiple response question, many respondents incorrectly believed that they needed to make a choice between the ‘New Zealander’ response and another description of their ethnicity. Respondents may not have read or understood the instruction asking them to ‘mark the space or spaces which apply to you’, as this instruction appeared as a separate sentence immediately after the question. Respondents who were inclined towards satisficing behaviour may have read only the main question and skipped immediately to the response categories, without reading the additional instruction.

One strategy to reduce the impact of these problems would be to incorporate the instruction at the beginning of the question stem, as was done in the 1996 version of the ethnicity question, so that it read “Tick as many circles as you need to show which ethnic group(s) you belong to”. This may help to ensure that respondents understand that multiple responses are acceptable. However, this is unlikely to completely remove the influences of satisfiicing altogether.

Satisficing behaviours pose a serious barrier to the introduction of a ‘New Zealander’ option within the ethnicity question. Electing to place this option earlier in the response list may encourage greater and more consistent endorsement of the 'New Zealander' option amongst those that prefer this description because it appears first in the list. However, it may subsequently lead to reduced levels of response for other options which appear later in the list.

Regardless of the measurement problems associated with this question, an important finding from this research was that respondents expressing a preference for ‘New Zealander’ were not a homogenous group. Both focus groups and cognitive interviewing confirmed that the motivations for answering ‘New Zealander’ can be different for different people and endorsements of this term are ambiguous within the context of ethnicity. This means that statistics derived from a question which includes a ‘New Zealander’ category would be difficult to interpret.

Overall, results highlight the need for researchers to think carefully about question categories, and the ways that they intend to use the information they collect. Those requirements should directly shape and guide the decisions relating to questions, so that questions will best meet the data need. Because the primary aim of the ethnicity question is to identify the diversity of the New Zealand population, an explicit 'New Zealander' response may only counteract that objective.

5.1 Generational attachment questions

The results from the testing of the generational attachment questions were largely inconclusive. Questions on parents' birthplace appeared to do little to either reduce or increase the preference to mark 'New Zealander', and there was no evidence in this research that such questions had any perceptible impact on ethnicity response.

However, other research has shown that questions which precede another can have order effects which may impact on answers in both directions (see for example, Schuman & Presser, 1981) and this may have masked the true impact of these questions. For example, the generational attachment questions may have acted as a kind of 'primer' by reminding respondents of their parents' birthplace and making it more probable that they would include that aspect of their identity in their answer to the ethnicity question. In other words, this may have produced an 'assimilation effect' (Schwarz, Strack & Mai, 1991), where respondents align their answers to be consistent with their earlier answers. Note, however, that this effect would work in opposite ways for different respondents. For example, an assimilation affect is likely to discourage a 'New Zealander' response amongst first generation New Zealanders (those with parents born elsewhere), but may encourage a 'New Zealander' response amongst second generation New Zealanders (those with parents born in New Zealand).

An alternative explanation is that the additional questions could create a 'subtraction' or 'contrast effect' (Schuman & Presser, 1981). This is explained by social discourse theory as the 'relevance' rule (Grice, 1975). Respondents feel obliged to communicate relevant information and not to repeat information that they have already 'given'. This principle further implies that respondents should add 'new' information to remain relevant and informative. Therefore, by asking questions which give respondents the opportunity to identify their generational attachment to New Zealand first, respondents may then feel more inclined to give answers to the ethnicity question which tap other dimensions of their identity. Like contrast effects, assimilation effects would also work differently for those with parents born in New Zealand and those with parents born elsewhere. To feel confident about the impacts of generational attachment questions on respondents' ethnic responses, additional testing would be required.

6.0 Benefits of multiple methods

This programme of research reinforced the value that multiple evaluation methods can have in building our knowledge of an issue. Findings from one research method can complement and support findings from another. For example, the results from cognitive interviewing confirmed the results from focus group testing and both methods identified a number of key motivations to explain why respondents opt to describe themselves as New Zealanders.

However, findings from each research method were also able to extend and expand the findings from the other methods, and each bought a unique contribution to our understanding of the ethnicity issue. Inter-censal matching, for example, allowed us to identify which ethnic groups were most vulnerable to data shifts from the introduction of a 'New Zealander' category. The focus group testing helped develop our understanding of the shift in ethnicity statistics by identifying the social and political motivations associated with a 'New Zealander' response. It also allowed researchers to explore those issues in depth and consequently enhanced and enriched our knowledge of the respondent's context. Finally, the cognitive interviewing component of this research demonstrated that other factors were equally likely to impact on respondents answers and demonstrated how the design of the questions can interact with the respondents' intentions to influence the final result.

However, as a qualitative methodology, cognitive interviewing was unable to draw firm conclusions about order impacts and the effect of the generational attachment questions. For this particular research issue other methods, such as split sample testing, would have been more useful and would also have helped quantify the risk to ethnicity data if a 'New Zealander' category were introduced. This conclusion further reinforces the value of using multiple methods to evaluate an issue.

7.0 Conclusion

Overall, these findings from testing a 'New Zealander' tickbox within the ethnicity question illustrated many of the problems we face as questionnaire designers. Results reinforce the knowledge that response categories and question wording can have a powerful influence on answers. The tendency for respondents to satisfice and adapt their answers to fit the listed response categories showed that it would be difficult to separate question effects from 'true' shifts in respondents' perceptions of their ethnicity when examining the data. It also showed that a 'New Zealander' category has the potential to convert response away from important groups which are of high public policy interest, such as Maori, Pacific peoples and Asian. Such results show that question design is not a trivial aspect of the measurement process and highlights the importance of developing and testing response options which are appropriate, mutually exclusive and free of order impacts to obtain an accurate measure.

Together the various methods used in this research were able to provide an improved understanding of ethnicity response for Statistics New Zealand. The combined findings from these studies strengthened our knowledge of the issue and gave Statistics New Zealand greater confidence in making decisions about the ethnicity question for Census 2011.

8.0 References

Clark H H & Schober M F (1992). Asking Questions and Influencing Answers. In J M Tanur (Ed), Questions about Questions (pp 15-48). New York: Russell Sage Foundation.

Dillman, D. A. (2002). Mail and Internet Surveys: The Tailored Design Method: With New Internet, Visual, and Mixed-Mode Guide. New York: Wiley.

Grice, H P (1975). Logic and Conversation. In P Cole & J L Morgan (Eds), Syntax and Semantics 3: Speech Acts. New York: Academic Press.

Krosnick, J & Alwin, D F (1987). Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement, Public Opinion Quarterly, 51.

Schuman H & Presser S (1981). Questions and Answers in Attitude Surveys. New York: Academic Press.

Schwarz N. & Hippler (1992). Response Alternatives: The impact of their choice and ordering. In P Biemer, R Groves, N Mathiowetz & S Sudman (Eds). Measurement Error in Surveys (pp 41-56). Chichester: Wiley.

Schwarz N F; Strack F & Mai H-P (1991). Assimilation and contrast effects in part -whole question sequences: a conversational logic analysis. Public Opinion Quarterly, 55, 3-23.

Sudman S, Bradburn N & Schwarz N (1996). Thinking about answers: The Application of Cognitive Processes to Survey Methodology. Josey-Bass: San Francisco.

Statistics New Zealand (2003). Preliminary Views on Content: 2006 Census of Populations and Dwellings Catalogue 23.001. April.

Statistics New Zealand. (2009). Final report of a review of the official ethnicity statistical standard 2009. Wellington

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

UMR Research Limited. (2009). Public Attitudes and Understanding of Ethnic Identity: A Qualitative Report, UMR Auckland Available from www.stats.govt.nz.

Standardization at Statistics Sweden
– experiences and challenges

Gunilla Davidsson & Andreas Persson

Since 2007 Statistic sweden is re-organizing from an organization based on statistical products to a process-oriented organization. Last year all methodologists were centralized to a process department. A project group developed a model of the statistics production process. It includes seven different processes of which all have a number of subprocesses. It also included a general process for evaluation and feedback. The goal is that every process and subprocess should be standardized. That is, to have common set of tools instead of different tools for different surveys, and to have a common set of best practices.

The reasons for standardization, as Statistic Sweden views it, are, among other things, to increase the overall quality by raising all surveys to the level of best practice. Another reason is to be in control of the production process and not rely too much on particular individuals. Third reason; it is easier and cheaper to support and develop a chosen few electronical tools than many. The support of electronical tools used to be, and still is in many cases, a huge cost for Statistics Sweden. Therefore introduction of common tools is a big incentive. All the tools, checklists, standards etc. are going to be available in process-support tool. This is going to be a work station in which survey managers can design and document their survey. All together these are quite big changes for our organization. The question is, what does this mean to the questionnaire designers?

The re-organization has influenced the questionnaire-design group in many different ways. First, our position in the organization has changed. We used to be a ”virtual group” with members from different places in the organization. Now we are a unit – Unit for cognitive methods – like the other methodologists. Second, we have expanded, from 9 to 12 co-workers in the last year, and more people will join the unit in the near future. Third, we have become an explicit part of the statistics production process (the model described above includes subprocesses such as ”Build collection instrument” and ”Test collection instrument”), which meant we had to standardize our work as well.

To standardize how we develop and test questionnaires, two projects were conducted. The overall goals of the projects were to develop detailed guidelines and checklists for all methods/procedures we use (expert reviews, cognitive interviews, focus groups, in-depth interviews, usability studies, debriefings, question writing and questionnaire layout and design). For each method we wrote general guidelines. Each guideline included five sections: Introduction, preperations, conducting the method, analysis and reporting the results. For the ”preperations” and ”conducting the method” we also developed detailed checklists.

They include a row for every decision in the work process with boxes to check. There is also a reference back to the general guidelines if one wants to read more about what to consider when making a certain decision.

So, we developed guidelines and checklists for all our methods/procedures as a consequence of the standardization process Statistic Sweden is going through. However, we have many remaining challenges such as how to work in this new position and which our role is in the new organization. Since testing questionnaire now is an explicit part in our process model, the amount of questionnaires that need to be tested is going to increase substantially. In addition, Statistic Sweden is trying to be certified to the ISO standard and this puts further demands on the testing of questionnaires. One challenge is how our group of questionnaire designers should handle the increased workload. Probably we need to delegate some parts of today's work but the question is which parts and to whom? One solution could perhaps be to delegate the review and design of questionnaire layout to the group of questionnaire constructers who build the questionnaires. However, it is still uncertain how this would work.

Another way to meet the increased amount of questionnaires needing to be tested is, of course, to become more efficient. One way to become more efficient is to take better care of results from, for example, previous cognitive interviews. We have more than 200 test reports in our archives. They are rarely used because of limited access to the information in them. However, making previous test results more accessible has many advantages. We might not need to test the same or similar questions over and over again. We will also have important information at hand in those situations where it is not possible to conduct new cognitive interviews due to lack of time and/or money. In addition, it would enable us to get an overview and evalaute both own work and more general issues concerning questionnaire design. This is the reason why we, in 2008 inspired by Q-bank, started a project to organize our findings from cognitive interviewing in a database. The idea is to allow searching based on question characteristics, question topics, problem areas and free text. So far we have a coding scheme and a plan concerning the functionality. Unfortunately, building the database has been postponed until 2010. Thus, to delegate some parts of our work and to become more efficient in other parts, are some ways to meet the increased demands of testing.

Another challenge is that even though we have developed standards and guidelines, some parts of our work is not standardized enough. Concerning cognitive interviews, we have just a few explicit procedures on how to choose what to focus on in a questionnaire, how to choose cognitive technique depending on the situation or how to analyze the interviews. This might not be a problem in itself, all this have worked quite well so far. However, it might become a problem in this new situation, since inexperienced people are joining our group while experienced people are retiring. It could lead to variations in the results and problems in describing the methods to other groups such as survey managers testing their survey for the first time. It might also lead to problems in developing cognitive interviewing within our group due to the lack of grounding and difficulties in communicating our ideas about the test to the interviewers, who are not belonging to our group but who do a majority of all cognitive interviews. To avoid such problems, we are trying to structure how to work with cognitive interviews in the future. We are considering starting using coding scheme and software support tools for our analyses. Today we primarily use restrospective delayed verbal probing. But we want to examine other cognitive techniques (such as think aloud, card sorting, vignettes etc) to make them natural parts of our toolbox as well. This, however, is a part of our future work.

To summarize, Statistics Sweden is being re-organized from an organization based on statistical products to a process-oriented organization. This is influencing the role of the questionnaire designers in many ways and will continue to do so in the next couple of years. We have now become an explicit part of the production which will result in more questionnaires to test. We also had to standardize the methods we are using. Many challenges still remain but we a trying to structure our work and become more efficient to meet all forthcoming demands.

Developing Standard Questions of Surveys
Statistics Canada's Experience

Paul Kelly, Statistics Canada

Introduction

In 2006, Statistics Canada initiated a program entitled the New Household Survey Strategy. This program was to review the current way that Statistics Canada conducts its many social surveys with a view to better integrate many of those surveys into one survey program.

One part of the New Household Survey Strategy was an initiative for greater standardization of questionnaire modules across Statistics Canada's household surveys. Most of these surveys collect data for socio-demographic and economic concepts, however very few standard approaches are used for collecting these data. The benefits and importance of standardization can be measured in time and cost savings, increased data quality and reduced response burden for Statistics Canada's household survey respondents.

While all Statistics Canada household surveys were in scope for this initiative, the largest surveys were key to the development of the standard questions. These included:

- The Census of Population - centennially
- The General Social Survey (GSS) - annually
- The Labour Force Survey (LFS) – monthly
- The Canadian Community Health Survey (CCHS) – annually
- The Survey of Household Spending (SHS) – annually
- The Travel Survey of Residents of Canada (TSRC) – annually

Project Objectives

The objective of the project was to develop standardized questionnaire modules for key cross-cutting socio-demographic variables being collected, regardless of the mode of data collection. The scope of the project included standard concepts and definitions, wording of questions, interviewer instructions, data editing approaches, output groupings, response categories, and comparability across surveys which would support analysis and data quality.

First, project team members identified concepts and variables of interest common to the current household surveys. A standardization process was developed to reword each concept and variable of interest. The steps of this process were:

- evaluation of existing questionnaires and outputs
- expert consultation (data users, subject matter experts, analysts, interviewers), including international standards
- design of questions
- evaluation of questions including testing for multi-mode approaches
- approvals by Statistics Canada management (steering committees)
- specification development and testing
- implementation
- maintenance

The steps towards final implementation included a rigorous procedure involving subject matter experts, external committees and senior steering committees. In order to have a full review of the standards, early presentations were given to subject-matter committees and lateral consultation proved to be a valuable exercise. A fundamental component of this work was to involve external data users in the consultation and development process. A timetable including the above steps for each concept was created and a gradual implementation plan for each survey area was agreed upon, although to date not all surveys have implemented the new standard questions.

Why Standardize Questions?

The main premise of the initiative was that using standard questions across all household surveys at Statistics Canada could:

- save money in the existing household survey program
- address response burden and decreasing response rates
- address changes in the population (cell phone use, new technologies, etc.)
- increase coherence and comparability within the household survey program

The development and maintenance of standardized questionnaire modules for social surveys could improve the cost-effectiveness of household surveys. With on-the-shelf questions and accompanying Blaise code, Statistics Canada could become more flexible in its' data collection approaches and be able to respond to clients more quickly. Collection activities could be better tailored to respondent needs and in turn allow Statistics Canada to offer lower-cost options to clients. Respondent burden could be reduced directly for respondents participating in more than one survey, and indirectly as interviewers become accustomed to the wording and sequencing of standard questions.

This initiative could promote greater integration and harmonisation in Statistics Canada's collection approaches, particularly in the computer assisted interviewing (CAI) and multi-mode environments. As social survey content becomes more harmonized this will improve data coherence and can simplify field staff training and data collection.

Pre-programmed, approved and tested BLAISE modules on the shelf are an attractive incentive, a way of reducing time and cost. Having pre-tested questionnaire modules electronically in a way that they can be easily integrated into surveys under development will ensure that the most appropriate questions are used and improve data quality. Timeliness is an obvious factor, but the benefits also stretch to improved interpretability and coherence of disseminated products and output. Added harmonisation across many of Statistics Canada's social surveys, including the Census, are of increasing interest to clients and researchers who want to be able to combine data from various sources. Standard code sets could also be created for the variables, leading to consistency of outputs for Public Use Microdata Files as well as master files for data users and researchers.

The notion of standardization did not end with standardized questionnaire modules programmed in BLAISE. It carried through to processing and documentation to support dissemination. The need to have some flexibility was recognized. For example, the standards provided choices from a standard set of questions providing limited detail versus more extensive detail.

The success of this project continues to depend critically on the approval and buy-in needed at all stages and at many levels throughout Statistics Canada as well as externally. The biggest impediment to implementation is the real concern that new approaches will impact historical continuity. Implementation requires the engagement of survey managers and support from senior management during and after the transition. It also requires rigorous testing and evaluation to determine quality and how the change could affect historical data series.

A process to ensure the use of standard modules and control mechanisms was developed and continues to be implemented. Communication of decisions regarding the standard questions was of utmost importance. This communication had to occur across a wide audience including management committees, subject matter areas and data collection areas. Flexibility was also critical as the one size fits all will not always apply. Consistent support from senior management on decisions related to the standards was and continues to be essential.

A transition plan was developed and communicated so as to reduce the impact that making a change may have on time series, and to inform data users and clients of possible changes to expect in Statistics Canada's ongoing household surveys. This plan had to allow for existing data collection, processing and dissemination systems accommodation and transition to the standardized modules. Parallel implementation was an important option for ongoing surveys to have in moving to the standard questions. Agreement and support from the subject matter divisions was and is a key to any success of the standard questions initiative.

Variables Identified for Standardization

By researching Statistics Canada's current household surveys and by thinking forward to data needs of future surveys, the following variables were identified as good candidates for standardization:

- Household composition and relationships
- Age and sex
- Marital status
- Education
- Income*
- Aboriginal status
- Immigration and citizenship
- Ethnic origin*
- Language
- Religion
- Fertility
- Labour market activities
- Unpaid household activities
- Volunteering and civic participation*
- Health and activity limitations*
- Mobility and migration*
- Place of residence one year ago*
- Dwelling and household characteristics

As the process of developing questions and consulting with subject matter experts began, it became clear that some of the variables that were initially identified as candidates for standardization were in fact, not so easy to standardize. This could be because the variables were not common across the many household surveys or because the variable was too complex to standardize in a few simple questions or the subject matter experts in the different survey program areas could not agree on standard questions for that variable. These variables are identified in the list above with an asterix (*). It was decided, relatively early on in the process for some, to remove these variables from the standard questions initiative.

Some Examples

Listed below are some of the variables identified for standardization along with the questions used for that variable in some of Statistics Canada's household surveys.

MARITAL STATUS

- Census question (self-complete)

  MARITAL STATUS
  - Never legally married (single)
  - Legally married (and not separated)
  - Separated, but still legally married
  - Divorced
  - Widowed

  There is a separate question on the Census for common law relationships.

- GSS and LFS question (CATI)

  What is your marital status? Are you:
  … married
  … living common-law
  … widowed
  … separated
  … divorced
  … single, never married

- SHS question (CATI or CAPI)

  What is your marital status? Is it:
  … married
  … common-law
  … never married (single)
  … other (separated, divorced or widowed)

There are three slightly different ways on four different surveys of asking for the same marital status information.

LANGUAGE

- Census question (self-complete)

  What is the language you first learned at home in childhood and still understand?
  - English
  - French
  - Other, specify

- CCHS question (CATI)

  What is the language that you first learned at home in childhood and can still understand?
  … English
  … French
  … Arabic
  … Chinese
  … Cree
  … several more languages are listed as response categories

- GSS question (CATI)

  What language did you first speak in childhood?
  … English
  … French
  … Italian
  … Chinese
  … German
  … several more languages are listed as response categories

Again, there are three slightly different ways of asking for the same information.

INCOME

- Census question (self-complete)

  During the year ending December 31, 2005, did you receive any income from the sources listed below?
    - Paid Employment
    - Self-Employment
    - Income from government
    - Other Income
    - Total Income

  Definitions and examples are provided for each of these sources of income. This question is followed up with a question on the amount of income from each source.

- SHS question (CATI or CAPI)

  What is your best estimate of your total income, before taxes and deductions, from all sources in 200X?

- GSS question (CATI)

  What is your best estimate of your total personal income, before deductions, from all sources during the past 12 months?

- LFS question (CATI)

  Including tips and commissions, what is your yearly wage or salary, before taxes and other deductions?

- CCHS question (CATI)

  What is your best estimate of your total income, before taxes and deductions, from all sources during the past 12 months?

The same piece of information on income is collected in five different ways in five different surveys.

IMMIGRATION

- Census question (self-complete) and LFS question (CATI)

  Are you now or have you ever been a landed immigrant?

  In what year did you first become a landed immigrant?

- SHS question (CATI or CAPI)

  Are you now or have you ever been a landed immigrant to Canada?

  In what year did you first immigrate to Canada?

- GSS question (CATI)

  Are you now or have you ever been a landed immigrant in Canada?

  In what year did you get your landed immigrant status?

- CCHS question (CATI)

  Were you born a Canadian citizen?

  In what year did you first come to Canada to live?

Once again, the same piece of information is collected in four slightly different ways in five different surveys.

Many more examples could be provided. These four examples illustrate the subtle differences that can occur across surveys that are attempting to collect the same piece of information.

Question Testing and Evaluation

Before submitting the standard questions for approval of all subject matter divisions and senior management, the questions developed were tested and evaluated by questionnaire design experts, survey interviewers and respondents. This evaluation was done:

- to ensure the questions were universally understood
- to ensure the appropriate response categories were used
- to ensure respondents could answer the questions and were willing to answer the questions
- to research the respondent friendliness and the interviewer friendliness of the questions.

The testing was also conducted to provide insights into the complexity and burden associated with alternative question sets, which would hopefully help Statistics Canada control respondent fatigue. This research paid particular attention to cultural differences in how questions are interpreted.

Question testing was done using two methodologies: focus groups and cognitive interviews.

Focus groups were conducted with Statistics Canada's experienced interview staff. Five focus groups were held in four regional data collection centres (Winnipeg, Sturgeon Falls, Sherbrooke and Halifax). At these focus groups, interviewers provided comments and recommendations for improving the suggested standard questions.

Cognitive interviewing of respondents took place in three phases. Between phases, changes were made to the questions based on the findings of cognitive interviews conducted in the previous phase. This allowed for fine tuning of the questions. During each phase, some interviews were conducted using a self-complete version of the standard questions while other interviews tested the interviewer assisted version of the questions. Cognitive interviews took place in rural and urban areas. The test respondents were purposively sampled so as to have a mix of ages, genders, incomes and education levels, religions, immigrants and non-immigrants, complex and simple family types as well as a mix of dwelling types.

Phase one of cognitive interviewing took place in February and March of 2007. In total, sixty-six interviews were conducted in six locations (Vancouver, Winnipeg, North Bay, Ottawa, Montreal and Sherbrooke).

Phase two of cognitive interviewing took place in May and June of 2007. Thirty-four interviews were conducted in four locations (Edmonton, Red Deer, Moncton and Halifax.

The focus groups and cognitive interviews provided a lot of information on the standard questions that were developed and tested. Many recommendations were brought forward and improvements made to the questions that were first developed as standards. There are far too many findings to list in this paper. A few examples are provided below.

LANDED IMMIGRANT STATUS questions

A few different versions of the immigration and citizenship questions were tested.

The major findings of the test were:

- these types of questions can be difficult for recent immigrants, especially those answering questions in their second or even third language
- the landed immigrant status questions need to flow with other questions on immigration (country of birth and year of landing)
- the terminology permanent resident and landed immigrant are not well understood. These terms seem to have different meanings for different organizations, including different levels of government. For this reason, the definition of landed immigrant status should be included in the question itself. Rather than asking for the respondent’s landed immigrant status, ask if they have been granted the right to live in Canada permanently.

INCOME questions

As is often the case when testing questions related to income, respondents in the standard question testing voiced concerns about the sensitive nature of this information. A transition statement explaining the purpose of income type questions to respondents helps ease their sensitivity. In the early versions of the standard questions there was no standard transition statement included with the recommended wording of the income questions. A transition statement was added for subsequent phases of testing.

Two versions of the income question were tested. One version asking for the overall income amount. Another version asking for all the types of income and amounts for each type. During testing it was discovered that the detailed information was difficult to provide in an interviewer administered survey. In order to provide the detail, a respondent would need to take time to review the various types of income and decide which types apply to them and then provide the amounts for those types. A self-complete questionnaire allows respondents to do this while an interviewer assisted questionnaire probably does not. A recommendation was made to keep the detailed income question for self-complete questionnaires but to eliminate it for interviewer assisted questionnaires.

HEALTH AND ACTIVITY LIMITATIONS questions

Before any testing of the questions related to health and activity limitations took place, subject matter experts surmised that this topic could not easily be standardized because the topic was too context sensitive. Asking questions about general health or about activity limitations in a health survey, where these questions are surrounded by detailed health questions, can often produce different results than asking the same questions in a survey of another topic. An attempt was made to standardize the questions and they were tested in the first phase of cognitive interviewing. However, based on the advice of subject matter experts and partly on the results of the phase one testing, a decision was made to drop this topic from the standard questions initiative.

Conclusions

The move to a more standardized way of collecting socio-demographic information for household surveys has many potential benefits. Improved timeliness and data quality in terms of coherence and interpretability are the most substantial benefits. Secondary benefits range from more consistency across surveys for researchers to potential reductions in response burden to possible cost savings in terms of specifications, testing and programming in subject matter divisions.

However, any attempt to standardize will only go as far as the relevant surveys will take it. The benefits can only be realized if senior management and survey managers buy into the initiative and actually adopt the standard questions.

To date, Statistics Canada has invested in the creation of standard questions for some socio-demographic variables. Questions have been developed, tested and approved. The future will tell if the standards are implemented and maintained and if the benefits can be realized.

Health Insurance Measurement: A Synthesis of Cognitive Testing Findings
October 14, 2009
Joanne Pascale

1. OVERVIEW

Over the past two decades, health researchers have been grappling with the issue of how to best measure health insurance. These efforts have included literature reviews, comparative studies, taxonomies of differing methodologies and estimates across surveys, cognitive testing, and split-ballot field experiments. This paper focuses on one aspect of these efforts – cognitive testing – whose aim is to better understand how the questions are being understood and answered from the respondent's perspective, and to identify features of the questions and questionnaire as a whole that may be associated with misreporting. The purpose of this paper is to first assemble all known findings from cognitive testing[1] and, second, to synthesize those findings across tests and take some first steps toward recommendations on best practices for health insurance questionnaire design.

2. METHODS

Because most cognitive testing studies are undertaken to inform questionnaire design for production data collection, results are often generated simply in terms of recommendations. That is, written reports or memoranda are not always produced, and findings are published very infrequently. With this in mind, an effort to gather both published and "grey literature" (unpublished reports such as conference proceedings, internal reports and memoranda) was undertaken, beginning in late 1998. First a series of phone calls was made to relevant staff at key federal agencies and research firms that carry out national surveys on health insurance. At the state level, staff were contacted at the State Health Access Data Assistance Center (SHADAC), which provides technical assistance and grants to states to carry out health-related research. Through their workshop with state representatives, efforts were made to gather and share relevant findings on health insurance measurement. Finally, a roundtable was organized at the 2002 American Association for Public Opinion Research for the same purpose. In addition to soliciting research findings from other agencies, the Census Bureau conducted a number of internal studies during these years.

Altogether these efforts resulted in a collection of eleven reports from various agencies (the National Center for Health Statistics, Westat, the University of Massachusetts Survey Research Center, and the Census Bureau) which assessed nine different questionnaire designs, including the Current Population Survey (CPS), American Community Survey (ACS), and the National Health Interview Survey (NHIS). Table 1 displays the surveys for which reports were obtained,

[1] This is intended to be a "living document". While substantial effort has been made to gather existing reports and contact researchers who may have conducted relevant research, it is quite possible that important studies and findings are missing from this analysis. Readers who are familiar with any relevant reports are encouraged to contact the author. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Address correspondence to Joanne Pascale; e-mail: joanne.pascale@census.gov.

along with the authors and dates of those studies (all of which are available on request).

The table is organized by “design type.” Among the questionnaires for which reports were available, they generally fell into one of three categories in terms of basic questionnaire structure. One approach (labeled “Status-Type” in Table 1) starts by asking a global question on status (covered or not), and if the respondent is covered, a followup question is asked to determine the type of coverage. Another approach (labeled “Type-by-Type”) uses a series of roughly eight questions, each asking if anyone is covered by a particular type of coverage. In addition to these common designs, some experimental work has taken place over the past decade examining new design approaches. The University of Massachusetts conducted a focus group and developed a short set of questions aimed at deriving only health insurance status (not type), and then conducted cognitive testing and a small pretest. The result was a design made up of just three questions: one global question on status, and two followup questions to capture in-scope coverage that may have been missed by those who said “no” to the global question. As a follow up to that work, at the Census Bureau an experimental design was developed which begins with a global question on status but then asks a question on general source (through employment, direct purchase, government or some other source). Tailored questions on each of these major sources are then asked in order to gather the necessary detail (e.g.: policyholder versus dependent, Medicare versus Medicaid). This was first tested in the fall of 2003. It was later revised and modified based on additional research, and this modified version was tested in the spring of 2008. Since the UMASS design is somewhat similar in structure to the three experimental designs developed by the Census Bureau, all four are grouped together in the table. Complete question wording for all eleven surveys is displayed in Appendix A, and details of the methodology employed in each test is presented in Appendix B.

Table 1: Questionnaire Design Types and Cognitive Testing Studies

Status-Type		Type-by-Type		Experimental	
Surveys	**Studies**	**Surveys**	**Studies**	**Surveys**	**Studies**
Behavioral Risk Factor Surveillance System (BRFSS)	Beatty and Schechter, 1998	Medical Expenditure Panel Survey (MEPS)	Westat, 1994; Kerwin, Cantor and Sheridan, 1995	University of Massachusetts (UMASS)	Roman, Hauser and Lischko, 2002
National Health Interview Survey (NHIS)	Beatty et al, 2002	Current Population Survey (CPS): Medicaid & SCHIP Questions	Loomis, 2000	Experimental Design Round 1 (EXP1)	Pascale, 2001
American Community Survey (ACS)	Pascale, 2005	CPS (full questionnaire)	Pascale, 2006	Experimental Design Round 2 (EXP2)	Pascale, 2003
		State and Local Area Integrated Telephone Survey (SLAITS)	Willson, 2005	Experimental Design Round 3 (EXP3)	Pascale, 2009

3. STATUS-TYPE DESIGN

3.1 Global Question on Health Insurance Coverage Status

3.1.1 Question Wording

Altogether seven different questionnaire versions started with some type of global question on coverage status (BRFSS, ACS, NHIS, UMASS, EXP1, EXP2, EXP3):

BRFSS: Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?

ACS: Is this person CURRENTLY covered by any type of health insurance? Include insurance obtained through a job or purchased directly from the insurance company, and government health insurance such as Medicare, Medicaid, VA and military programs.

NHIS [Are you/Is anyone] covered by any kind of health insurance or some other kind of health care plan? READ IF NECESSARY: Include health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills.

UMASS:Do you currently have any kind of health insurance coverage at all?

EXP1: [Are you/Is NAME] now covered by any kind of health insurance plan, HMO or government assistance health plan such as Medicare or Medicaid?

EXP2: Are you covered by any type of health insurance? [Post-test recommendation: Do you have any type of health coverage or health plan?]
EXP3: Do you have any type of health plan or health coverage?

While each of these seven surveys employed a different set of followup questions, they all began with a basic question asking about coverage status, so there is little chance that results from the global question would be contaminated by the variation in followup questions. Thus results for this particular question will be discussed as a group.

3.1.2 Results

The BRFSS reported that the question successfully captured a variety of health insurance situations, and other tests noted that the question generally worked well. However, there were a number of problems detected, the most serious of which was underreporting of certain types of plans.

Under-reporting: The most fundamental problem, which emerged across designs, was that respondents with public or "atypical" coverage said "no" to this question. In many cases the reason for this misreporting was that respondents only considered something to be "insurance" if they were paying for it in some way. Following are the specific scenarios:

- Participants in the ACS and the NHIS said "no" to this question but turned out to have veterans' coverage. In the NHIS case, this coverage was discovered in the followup question on type of coverage. In the ACS case, however, the respondent changed his "no" answer to a "yes" after hearing the "include" statement (see below for discussion).
- In the ACS three respondents with coverage said "no" and all three had public coverage of some type (VA and Freecare). They all said the reason for initially answering "no" was that "insurance," to them, meant something they pay for.
- In the UMASS test a respondent on Medicaid initially answered "no" but later reported the Medicaid in the followup question on public coverage. He said he felt he lost his insurance when he lost his job, and that he does not consider Medicaid to be "insurance."
- In the BRFSS a college student, who has access to acute care through the university, did not report it.

There were also some problems with question comprehension that led to underreporting. In EXP1 two respondents answered "no" but should have said "yes." One respondent misunderstood the question as asking if anyone was eligible for Medicare or Medicaid. Another said that no one was covered by an HMO, Medicare, Medicaid or any government plan, but that they have a PPO. The respondent seems to have focused on the particular examples listed and missed the key phrase "any kind of health insurance plan."

Though there is quite a range in the questions, in terms of how much they elaborate on what is meant by "health insurance," findings suggest little association between elaboration and reduced underreporting. In particular, two of the designs (NHIS and ACS) include a specific statement instructing respondents on what "counts" as insurance. In the NHIS case, this appears as a "read if necessary" statement. The NHIS report does not provide direct evidence on respondents' reaction to the “read if necessary” statement in the global question, but in general cautions against using this type of material because it relies on the interviewer’s judgement regarding what is “necessary.” The authors suggest incorporating the material into the question itself, or developing a follow-up question for respondents who say “no” to the initial question. The ACS design does just that; the “include” statement is incorporated as part of the question text. In the ACS case the testing protocol did explicitly probe respondents on the statement itself (separate from the yes/no part of the question), and there was not strong evidence that this “include” statement helped or hurt respondents’ understanding of the question. While one respondent with veterans’ coverage did change his answer after hearing the statement, two other respondents with public coverage (“Freecare”) failed to report it in spite of the statement. In EXP1 there is a slightly abbreviated version of the “include” statement embedded within the question text. However, in one case in that test, the respondent said “yes” because she had private coverage but then was confused by the mention of Medicare and Medicaid.

Avoiding all underreporting is probably impossible, but findings suggest it may help if the global question avoids or downplays the word *insurance* in order to capture those with free coverage that they therefore don’t consider “insurance.” Indeed, the EXP3 test – which asks about “health plan or health coverage” found no evidence of comprehension problems or underreporting.

Double-barreled Question: Some respondents in EXP1 ultimately answered the question correctly but were thrown off by the question structure and seemed to be unsure if it was a yes/no question on coverage, or a question on type of coverage. One said “I have employer-sponsored coverage – is that it?” Another said “It threw me off at the end...First you say ‘Are you covered by any health insurance plan’ and I thought ‘yes’ and then you said ‘Medicare and Medicaid’ and it threw me off.”

“Pre-Reporting”: It was noted in the BRFSS, ACS, EXP1 and EXP3 studies that the global question invited many respondents to volunteer more detail than was asked for, such as their plan type. This caused confusion in the BRFSS followup question on type of coverage; respondents did not understand the point of the followup question since they had already reported type of coverage in the first question.

Other Household Members: There seemed to be some confusion over which household members were “in scope” in some cases. In the NHIS several participants did not include their own coverage when first asked, even though the question refers to “anyone in the household.” And in EXP2, which asks about the respondent by name, the opposite problem was detected; respondents were confused over whether the question referred to the respondent only, or to all household members. Finally, in EXP3 some respondents said they had only a limited knowledge of other household members’ coverage status.

3.1.3 Summary

There is fairly compelling evidence that it is unlikely that one global question can successfully capture coverage status. Specifying types of coverage that are in-scope within the question does not necessarily aid in reporting, and a statement that includes descriptions of what types of coverage "count" may only serve to introduce interviewer variability (if not all interviewers read the statement) and may even confuse respondents. That is, it appears that including several examples of plan types can cause respondents to misinterpret the question as asking which of several types of coverage they have, rather than whether they have any coverage at all. Research findings to date suggest the EXP3 version is the least problematic option for asking about coverage status. Followup questions should ask explicitly about types of coverage commonly dismissed or not considered "insurance," such as Medicare, Medicaid, veterans' coverage, state-specific coverage, and possibly university-based coverage (if this type of coverage is in-scope for the survey).

3.2 Single Followup Question on Type of Coverage

3.2.1 Question Wording

Three of the surveys discussed above – BRFSS, NHIS and ACS:

BRFSS: How do you obtain the health care coverage you use to pay for MOST of your medical care? Is it through: [PLEASE READ] □ your employer □ someone else's employer □ a plan that you or another family member buys on your own □ Medicare □ Medicaid/state name □ another federal program such as the military, CHAMPUS or the VA or □ some other source □ none

NHIS: What kind of health insurance or health care coverage [do you/does NAME] have? INCLUDE those that pay for only one type of service (nursing home care, accidents or dental care), exclude private plans that only provide extra cash while hospitalized. □ Private health insurance plan from employer or workplace □ Private health insurance plan purchase directly □ Private health insurance plan through a state or local government program or community program □ Medicare □ Medi-gap □ Medicaid □ CHIP (Children's Health Insurance Program) □ Military health care/VA □ Tricare/CHAMPUS/CHAMP-VA □ Indian Health Service □ State-sponsored health plan □ Other government program □ Single service plan (eg: dental, vision, prescriptions) □ No coverage of any type

ACS: [Self-administered]: What type of health insurance does this person have? Mark (X) all that apply. □ Insurance through a current or former employer or union (of this person or another family member) □ Insurance purchased directly from the insurance company (by this person or another family member) □ Medicare, for persons 65 years old and older, or persons with certain disabilities □ Medicaid, Medical Assistance, or any kind of government-assistance plan for low-income children and families □ TRICARE, CHAMPUS or other military care □ CHAMPVA or VA □ Indian Health Service □ Supplemental plans that cover one type of care (e.g.: dental, accident, nursing home care plans) □ Other/specify

. [CATI]: What type of health insurance does this person have? Is it...insurance through a current or former employer or union (of this person or another family member)? ...insurance purchased directly from the insurance company (by this person or another family member)? ...Medicare, for persons 65 years old and older, or persons with certain disabilities ...Medicaid, Medical Assistance, or any kind of government-assistance plan for low-income children and families ...TRICARE, CHAMPUS or other military care ...CHAMPVA or VA...Indian Health Service...Supplemental plans that cover one type of care (e.g.: dental, accident, nursing home care plans)...Other/specify NOTE: Each response category used a discreet set of yes/no response categories.

. [CAPI] Next I am going to show you a list of health insurance categories [show Flashcard]. What type of health insurance does this person have? You may choose more than one.

The question stem varies across the three surveys, as does the list of plan types in the set of response categories, ranging from 8 (in the BRFSS) up to 14 (in the NHIS). The mode in which production survey are conducted also varies across surveys. The BRFSS is a telephone-administered survey, ACS is multi-mode (self-administered questionnaire (SAQ), computer-assisted telephone interview (CATI) and face-to-face (CAPI) administration), and NHIS is face-to-face. Each study attempted to mimic production conditions in the lab; however, in the BRFSS case about half were conducted face-to-face and half were conducted over the telephone with follow-up face-to-face probing. In the ACS SAQ mode the question text was presented on a mocked-up printed form and respondents were probed as they filled out the questionnaire. In the ACS CATI version discrete yes/no questions for each plan type were asked, and in the ACS CAPI version a flashcard was used. The NHIS also used a set of flashcards.

3.2.2 Results

Stem Question: Findings from all of the reports focused on the response categories and only the BRFSS commented explicitly on the stem question. The report noted that the term “most of your medical care” did not work with respondents. Those with only one source of coverage were confused by the question, and those with multiple sources didn’t know which source paid most of the costs.

Underreporting: The “include” statement in the NHIS followup question seems to have been ineffective in at least some cases, as the question repeatedly failed to capture single-service plans, even though it explicitly prompted respondents to include these types of plans.

Length: The BRFSS report noted that the question is very long, and that respondents did not grasp the stem part of the question unless the response categories were also read. Note that the BRFSS is telephone-administered and all response categories are read out to the respondent in one continuous question. However, even in the NHIS and ACS SAQ and CAPI modes, where respondents were provided with a flashcard listing a set of plan types and were free to read over the categories themselves, the length and complexity of the list seemed to be an issue. In the ACS respondents either didn’t read or didn’t understand the parenthetical phrase (“of this person or another family member”) in the first two response categories and so failed to report dependent coverage. In other cases respondents with public coverage misreported it as private coverage because they failed to read down the entire list and instead stopped at a response category higher up in the list that seemed like it might fit their situation.

Miscategorization of plan type:

- Job-based plans misreported: There were several instances of respondents misreporting job-based coverage as some other source, and there was a host of different reasons for the misreporting. In the ACS SAQ and CAPI modes, several respondents who were dependents on someone else’s job-based plan reported their source as “other” because they did not grasp the meaning of the first response category. Another respondent who was very uncertain about his mother’s plan ultimately reported it as directly-purchased, because he was confident she had some kind of private coverage, though it is more likely that the coverage was job-based (she had worked for the federal government for 25 years). A respondent with a policy through a professional organization reported it as job-based because she gets a group rate. She assumed the direct-purchase category was strictly for individual buyers who pay on the "open market" (versus a group rate available to employers, trade associations, etc.). And in the ACS CATI mode, a respondent mentioned her military-based coverage at the job-based question. In the NHIS and both the ACS SAQ and CAPI modes, respondents did not initially take in the term "former employer" and were unsure where to report this sort of coverage. In the NHIS test, several participants misclassified insurance received through a government employer as being "government plans."

- Public assistance and Medicare were misreported as directly-purchased and job-based plans in the ACS SAQ and CAPI modes because respondents hadn't read down the entire list so they incorrectly chose a category higher up on the list.
- Medicare vs. Medicaid: respondents confused these two programs in the ACS SAQ and CAPI modes, even when reporting for themselves. When reporting for other household members, sometimes the respondent knew it was government assistance of some kind, but did not know or could not recall whether it was Medicare or Medicaid. The BRFSS also found that Medicaid respondents were sometimes confused about which program they were on, but Medicare recipients did not seem to confuse their coverage with Medicaid. In the NHIS respondents on government assistance plans had difficulty classifying their coverage and the authors note that this could stem from the fact that there were two flashcards – one (Card C) with national-level plan types, and another (Card D, not available) with state-specific plan types.

Plan names versus plan types: In the ACS SAQ and CAPI modes some respondents misunderstood the task and scanned the list for their particular plan *name* (eg: BC/BS), rather than their plan *type*.

Reporting for other household members: In the ACS SAQ and CAPI modes some respondents knew other household members were covered but didn't know the specific plan type. As discussed above, sometimes this was uncertainty over Medicare and Medicaid. In another case the respondent knew her mother was covered but said the source could be either her former employer (since she's retired), direct purchase or the military (because she'd been in the military).

Military and VA Coverage: In spite of VA being explicitly listed as a response category, several respondents in the ACS SAQ and CAPI modes who had VA coverage did not report it, saying they rarely used it. And in the NHIS at least three participants with military coverage had a difficult time finding a category that applied to them. The authors don't know why this was the case but suggest rewording as "any military health care plan (including VA)."

Supplemental Coverage: In the ACS SAQ and CAPI modes this response category failed to pick up coverage that several different respondents had, mainly because they had in mind only comprehensive type plans (in spite of the word "supplemental" in the response category). We found no evidence that respondents confuse supplemental with comprehensive coverage. The one respondent who did have supplemental-only coverage did not consider it comprehensive, and many other respondents with both comprehensive and supplemental plans explicitly said they did not consider their supplemental coverage to be "health insurance."

3.2.3 Summary

None of these designs appeared to fare particularly well, as several instances of plan type misreporting were observed in each case. In some cases respondents were confused about which of two categories described their plan:

- job-based versus direct-purchase
- job-based versus military
- job-based versus government (if the job was government-related)
- public versus private
- Medicare versus Medicaid

In other cases the description of the plan type did not adequately convey the meaning:

- dependents failed to report their coverage at the job-based question
- respondents with coverage through a former employer had trouble finding an appropriate category
- respondents with VA or other military coverage had difficulty
- respondents reporting for other household members did not have enough knowledge of the plan details to select a category
- respondents with supplemental coverage consistently failed to report it
- some respondents were scanning the list for their plan *name* (e.g.: Blue Cross) and had trouble with the concept of plan *type* (e.g.: job-based)

Finally, some respondents lost track of which household members the questions were asking about.

The length of the response category list was clearly a factor that led to misreporting, but a bigger issue may be the nature of the list itself. The individual response categories were problematic in many cases, and some may be too detailed and technical for respondents to grasp, particularly when they have only limited knowledge about the coverage of other household members. In some cases respondents did not explicitly express confusion, but misreported plans nonetheless (which was only revealed during later probing). In other cases respondents were confused and could not find an appropriate category and so made their best guess. In a sense the respondents who may have doubts but simply provide an answer without expressing that doubt or confusion are worrisome because the interviewer has no indication at all that the respondent may be misunderstanding the question and thus cannot intervene with probes.

Few reports explicitly commented on the stem question itself. However, it is possible to make recommendations by process of elimination. The BRFSS noted problems with "most of your medical care" and the "include" statement in the NHIS seemed unproductive, for reasons noted above. Neither the NHIS nor the ACS reported any problems with the simpler versions ("What kind of health insurance or health care coverage do you have?" and "What type of health insurance does this person have?" respectively). Combining these findings with those above on the global question on status (specifically, to keep the question short and to avoid the word "insurance"), a recommended stem question would be: What type of health coverage or health plan do you have? However, as noted above, none of the lists of plan types – even the shorter lists – was without serious problems. In general the length of the lists and the complexity of individual response categories was problematic across tests. One way to address this would be to decompose the reporting task into several discreet, more manageable parts by splitting the list into "tiers." The first could be a question on general source of coverage (e.g.: employment, government), and later tiers could gather details within each of those categories. Each individual question, then, would be shorter and more narrow in scope and the respondent would be processing less irrelevant information. All the experimental designs (discussed below) take this approach.

4. TYPE-BY-TYPE DESIGN

Four reports on this design were obtained: one on the Medical Expenditure Panel Survey (MEPS), two on the Current Population Survey (CPS), and one on the NCHS State and Local Area Integrated Telephone Survey (SLAITS). Note that the 2000 CPS report was not a comprehensive test on the entire series of questions; it was a test conducted soon after the introduction of the State Children's Health Insurance Program (SCHIP) and it was targeted at just the Medicaid and SCHIP questions. A question on SCHIP was subsequently embedded within the CPS questionnaire and it is that series of questions that was tested in the 2005 round. Note also that the MEPS cognitive testing reports focused on a set of questions on managed care features of health plans, not the basic set of questions that determined status and source of coverage. Thus the written reports contain observations only about the managed care questions, not health insurance questions per se. However, in verbal communications, Westat staff made observations about the basic set of questions on health insurance and these observations are noted below where relevant. (Both the full written reports on the managed care questions and the personal telephone communications are cited in the references).

4.1.1 Question Wording

Complete question wording, including followups, probes and interviewer instructions, is displayed in Appendix A for all four of these surveys. To facilitate the discussion on findings, abbreviated versions are displayed in Figure 1. Note that for the CPS 2000 testing round, due to prior concerns regarding respondent confusion between public and private programs, two different sequences of questions were tested – both the standard sequence employed in the CPS (listing private and then public plans) and an alternative sequence (listing public and then private plans). Note also that in addition to variation in question wording for any given item, other survey design features differed. The MEPS and CPS were asked at the household level (i.e.:

"Was anyone in this household covered by...") but the SLAITS was asked at the person level ("Is CHILD covered by..."). Furthermore, the reference period varied across surveys. MEPS used a variable time frame (1-6 months), CPS used the past 12 months, and SLAITS asked about current coverage status.

Figure 1: Overview of Questions in Type-by-Type Designs

MEPS

1. Medicare
2. Medicaid
3. Other public program
4. Military
5. Job-based
6. Group or association (church, club, professional, business, retirement association, union)
7. School
8. Direct purchase from an insurance company
9. Someone who does not live here
10. Purchased from some other place not listed

CPS 2000 (Probing on Medicaid and SCHIP items only)

A. Standard Order

1. Job- or union-based
2. Direct purchase from insurance company
3. Someone who does not live here
4. Medicare
5. **Medicaid**
6. **SCHIP**
7. Military
8. Other

B. Alternative Order

1. Medicare
2. **Medicaid**
3. **SCHIP**
4. Job- or union-based
5. Direct purchase from insurance company
6. Someone who does not live here
7. Military
8. Other

CPS 2005 (full series)

Same as CPS 2000, Standard Order

SLAITS

1. Job- or union-based or direct purchase from insurance company
2. Medicaid
3. SCHIP
4. Indian Health Service
5. Military
6. Other

4.1.2 Results

Questionnaire Structure

Across all four tests, the questionnaire structure seemed to pose major comprehension and reporting problems for some respondents. The general issue seemed to be a combination of two intertwined factors. One was that several of the individual questions on plan type were problematic for various reasons (described in more detail below). The other problem was that respondents were "blind" to the fact that the discrete questions were actually part of a series of yes/no questions on type of coverage. That is, often the intent of each question, and the series as a whole, only became clear to respondents after all (or most) of the questions had been asked. As a result, sometimes respondents reported a plan too early in the series, where it didn't quite "fit", because they didn't know that more questions were coming, and that a later question was more appropriate to their type of coverage. A related problem was that sometimes individual questions contained too much detail which confused respondents and made them question their own judgement as to whether their plan "belonged" in that category, and so they failed to report it at all.

In sum, it seems that in some cases the problem was not respondent knowledge, but that the design of the series did not adequately tap that knowledge. The MEPS researchers noted that respondents generally knew whether there were covered or not, but that deriving status of coverage from individual questions on source was problematic (UMASS researchers noted this as well). Similarly, the SLAITS author reported that respondents usually knew whether their child was covered by private or public insurance, but the intent of the questions on particular source of coverage was not clear and hence some respondents were led to misreport what they did know about their child's coverage. This could have serious consequences for measuring the uninsured since under these designs there is no single question to determine status. Rather, anyone not reported as being covered in one of the individual questions on type of plan is generally considered uninsured.[2]

Listed here are specific examples of the general problems discussed above, and their consequences:

"Pre-reporting"

- As was observed in reports on the global status question, the CPS 2005 noted that respondents had a tendency to "pre-report" – that is, to report all plans for all household members within the first few questions in the series – whether those questions were the most appropriate to the type of coverage or not. Part of the problem may well be that the respondent is blind to the list of questions to come. One respondent reported that he was covered by a job-based plan and when asked who was the policyholder he said he was. He then reported that his mother was also a policyholder. Only later in the interview was it revealed that he was referring to his mother's Medicare "policy" – and that she was not the policyholder of a job-based plan that covered him. Another respondent, a retiree, reported his job-based plan at the first question in the series, and at the second question (on directly-purchased plans) asked "how would you consider Medicare?" And a third

[2] While most surveys using this type of design now include a "verification" question to explicitly ask those not covered in the main series whether they have coverage that may have been missed, this adds burden and introduces another source of reporting error.

respondent reported his Medicaid plan at the question on directly-purchased coverage because he was thinking of insurance he "got on his own," that is, not through an employer. This type of "pre-reporting" may result in respondents misreporting their public coverage as private insurance because they are trying to "fit" the coverage they do have into these questions early in the series.

- Similarly, in the CPS 2000 test one respondent reported her SCHIP plan at the question on coverage through "someone outside the household". Her rationale was that, after hearing the first two questions on coverage through an employer and direct purchase, she knew that the coverage was not through her job or anything she bought, but rather through an "outside" source. Again, consequences in this case would be that public coverage gets misreported as private.
- The CPS 2005 test also noted that some respondents who volunteered the health insurance situation of the entire household up front were annoyed with later questions because they felt they were redundant. These respondents said they stopped paying attention once they felt the questions were no longer relevant. This kind of respondent fatigue could lead to underreporting if questions late in the series are indeed relevant but respondents are no longer attending to the questions.

Double-barrelled Questions

- The SLAITS report noted that the first question (on private coverage through an employer, union or through direct purchase) was perceived as "double-barreled" – that is, asking two questions at the same time: (1) is your child covered? (yes or no) and (2) is that coverage through (a) an employer (b) a union or © purchased directly? For respondents with children on public coverage, they were reluctant to say "no" since they knew their child was covered, but then they were left to choose which of the perceived response categories best "fit" the type of coverage, often feeling that the last option ("obtained directly from an insurance company") might be the correct choice. This thought process led them to (incorrectly) say "yes" to the question on private coverage, even though their child was actually covered by a public plan. When interviewers read the scripted probes, and/or continued on with the next two questions (on Medicaid and SCHIP) respondents usually (but not always) understood the intent of the series better. Some then double-reported their public coverage while some, after hearing the SCHIP question, reported it at the appropriate questions and were prompted to go back and change their "yes" answer to the first question on private coverage to a "no".
- The CPS 2000 report noted a similar problem with the structure of the Medicaid question, which asks a yes/no question about Medicaid or state-specific names for Medicaid. Some respondents interpreted this not as a yes/no question but as asking "Which of these programs are you covered by?" The author, however, did not note how respondents ultimately answered, given this confusion.

Questions on Private Plans

The CPS 2005 report noted several problems with the type-by-type approach. There were several miscellaneous problems with the first three questions on private coverage. At the job-based item, some respondents overlooked the term “former employer” and failed to report retiree coverage, and some were uncertain where to report coverage obtained through a family member who is now deceased (this occurred at both the job-based question and the item on coverage through someone outside the household). There was also general confusion in cases where the categories of coverage in the questions were not mutually exclusive, such as job-based and military coverage, and job-based and coverage from someone living outside the household. On the over-reporting side, some respondents included out-of-scope plans such as worker’s compensation, vision and dental plans. And at the question on directly-purchased coverage, some respondents asked if they should include other forms of insurance, such as auto or life insurance.

Questions on Public Programs

Both CPS tests as well as the SLAITS (and UMASS) reports noted a great deal of confusion over the questions on public coverage. First, there were fundamental misconceptions about the public programs themselves. Both the CPS 2005 and the UMASS report observed some respondents who didn’t consider Medicaid, Medicare or VA coverage to be “insurance” since they didn’t pay for it in some way.

Another problem was confusion between programs. Some of the confusion stemmed from the similar-sounding names (Medicaid and Medicare) and some from inherent blurry lines between the programs themselves (Medicaid and SCHIP). Both CPS tests observed some respondents who confused Medicare and Medicaid, and some even thought they were one and the same program. Regarding Medicaid and SCHIP, in the CPS 2000 test two respondents reported their SCHIP at the Medicaid question because Medicaid was the program they had applied for (even though their children were subsequently enrolled in SCHIP). Another respondent ultimately reported correctly but said she was tempted to report the child’s SCHIP at the Medicaid question, partly because the insurance card itself said both Medicaid and SCHIP on it. Both the CPS and the SLAITS authors noted that sequencing of the questions on various public programs could be an issue, speculating that respondents may say “yes” to the first question that sounds somewhat similar to the plan they have.

In many other cases the reporting problems stemmed not so much from these basic misconceptions and blurred lines between programs, but from the state-specific program names embedded within the questions. The CPS 2000 report noted several respondents who were on Medicaid but used alternative terms for the program, such as “medical card,” “access card,” and “medical assistance.” These terms, rather than Medicaid, were more familiar to many respondents. One respondent, who recognized the state-specific term “Health Choice,” said this was the specific name for the more generic “medical assistance,” (versus “Health Choice” being the specific name for the more generic “Medicaid.”) Another respondent said she is covered by “medical assistance” and that Medicaid is “supplemental medical insurance.” For these respondents, the term “Medicaid” seemed to actually be a hindrance to their understanding of the question.

In other cases, respondents were thrown off by all the state-specific program names and acronyms, and some interpreted the Medicaid question more narrowly than it was intended. More specifically, in SLAITS respondents were sometimes unsure whether the question was asking about "straight" Medicaid, or some more specific program, such as those geared toward women and children. Similarly, in CPS 2000, a respondent thought the state-specific program name embedded within the Medicaid question was referring to a particular asthma program (which some of her children had qualified for) and hence she thought the question applied only to some of her children and failed to report Medicaid for the rest of her children.

In the SLAITS, some respondents thought the Medicaid and SCHIP questions were asking the same thing, and reported their coverage twice. Others (correctly) thought the questions were trying to get at which public program their child was covered by, but got confused by the various state-specific program names. Specifically, for respondents who knew their child was on Medicaid or SCHIP, if the specific program name that the respondent used for the coverage was not embedded within the question, respondents lost confidence in giving a "yes" answer, and some said "no", thus underreporting. Similar problems were noted in the CPS 2000 report. One respondent who knew she had Medicaid did not report it because the question read: "Was anyone in this household covered by Medicaid or Health Choice?" The respondent focused on Health Choice but did not recognize it as Medicaid and thus said "no" to the question, even though she was covered by Medicaid. Thus, in some cases, use of state-specific names can actually backfire because respondents are waiting to hear the program name they are familiar with from among the other program names, and if it's not part of the list they don't report the coverage.

In that vein, the UMASS researchers were concerned that respondents on Medicaid who were enrolled in commercial HMOs contracted through Medicaid would fail to report that coverage in questions on public insurance. Thus they conducted research and identified four specific HMOs which covered about 90% of all Medicaid HMO recipients and added a question containing these HMO names. Next they conducted a pretest with these names embedded in the questions. The sample (n=74) was comprised of names drawn from Medicaid administrative records. The response rate was very low (only 17 respondents) but among those 17, four respondents answered "no" to the initial standard Medicaid question and when asked the followup question with HMO names embedded, four respondents said "yes."

12-Month Reference Period

Only the CPS 2005, which employs a 12-month reference period, commented on reporting problems associated with the time frame. There were three main patterns of response to questions about the 12-month time period. Some respondents reported thinking of the past 12 months and exhibited careful attentiveness toward the time frame. Other respondents, however, seemed to pay no attention to the reference period and said they were simply thinking of "what they have now". The third type of respondent said they thought of neither the 12 month period stated in the question nor their current situation per se, but rather, the circumstances that defined the current spell of the insurance. One respondent said she was thinking of "the time that I myself was eligible, and only because of my pregnancy." Another thought of the past 3 years because she started her current job 3 years ago, and the health insurance situation has not changed in those 3 years.

While these findings suggest respondents are not necessarily attending to the reference period specified in the question, the consequences for underreporting among those currently insured were fairly benign in the test design. That is, respondents who have coverage at the time of the interview generally do not fail to report that coverage and since the reference period encompasses the date of the interview it is inconsequential whether the respondent is thinking of today, the past month, past 12 months, or past 12 years. However, consequences for the actual CPS could be more serious since there is a gap between when the survey is conducted and the calendar year reference period targeted in the questions. That is, if respondents are covered at the time of the interview but lacked coverage throughout the previous calendar year, they may mistakenly report that coverage. Furthermore, if those currently *uninsured* employ the same patterns of response observed in the testing, there could well be consequences for underreporting among those who think in terms of only their current situation or spell of uninsurance if those respondents had been covered at some earlier point in the reference period. For example, someone who is uninsured in March may well fail to report coverage they had during the previous calendar year since they are focused on only their current situation. Similarly, for those who think in terms of spells, if a respondent had job-based coverage for years but then lost her job (and coverage) in February, she may have this particular spell of uninsurance in mind when asked the questions on coverage, and hence fail to report her job-based coverage from January.

Household Composition and Shared Coverage

As with the reference period, among the type-by-type reports only the CPS 2005 report made observations about household composition (though note that the issue of respondents generally losing track of which household members the questions referred to came up in reports on the global status questions above). The CPS 2005 employs a set of "household-level" questions ("Was anyone in the household covered by [plan type]?") and this is in contrast to a person-level approach ("Was [NAME] covered by [plan type]?"). Respondents did seem to have difficulty with the household-level approach, particularly in relatively large, complex or non-traditional households. One respondent thought the questions were asking only about adults, not children. Another lived with her mother, father, and three teenage nephews. In general this respondent had difficulty remembering that the whole series was asking about everyone in the household, not just herself, and at one particular question she forgot to report her nephew's coverage because she "wasn't thinking of him." Another respondent living with his parents, his son and his brother neglected to report his brother's coverage even though he'd mentioned his brother in previous questions. A third respondent forgot to include her mother in one question, and her nephew in another, even though they were both covered. Household size did appear to be a factor in these cases; it was only in households with four or more people that respondents forgot about certain members.

Other respondents had difficulty reporting for other household members because they had only a vague understanding of the person's coverage and did not know the particular plan type. One, who lived with his partner and her 21-year-old son, said he had only a general idea of their coverage but did not feel confident reporting a specific plan type for them. Another said he was not sure whether his mother was covered by his father's military health insurance and did not know the name of that military coverage. And finally one respondent, who lived with several non-related housemates, could not provide details on his housemates' health insurance situation.

Overall, the reporting problems observed seem to stem from a combination of the sheer number of household members for whom a respondent is asked to report, and the respondent's familiarity with the details of those peoples' coverage. Somewhat surprisingly, neither of these problems was necessarily associated with the "closeness" of the relationship between the respondent and the household member for whom he or she was reporting. Some respondents forgot about distant relatives and others forgot about their mother and brother. In terms of lack of knowledge, somewhat predictably, respondents often did not know details of housemates' coverage, but some respondents had difficulty reporting for their mother and live-in partner. Perhaps this is not surprising, given that health insurance eligibility tends to revolve around the nuclear family – that is, husband, wife and children under 21. Therefore, a respondent may be in a fairly good position to report on their spouse's and children's insurance, because they all share the same coverage, but other combinations of reporter-reportee are perhaps more prone to error. Adult children reporting for their parents and siblings, for example, are not likely to share the same coverage and hence may not be very familiar with the plan type. Indeed, few errors or issues were observed for respondents who were policyholders reporting on their own policies and individuals covered under that policy.

4.1.3 Summary

Many of the findings from reports on the type-by-type approach reflect the same issues turned up in the status-type testing. Problems observed in both sets of testing are:

- some types of public coverage don't "count" to the respondent
- public plans sometimes get misreported as private plans due to a tendency to pre-report
- respondents have only limited knowledge of other household members' plan type
- respondents sometimes lose track of which household members are being asked about
- categories of coverage are not mutually exclusive (e.g.: job-based and military; job-based and coverage from someone outside the household; job-based and government, if the employer is a government agency)
- respondents have trouble reporting coverage from a former employer
- the "directly-purchased" category is unclear, especially when it involves a trade or professional association
- respondents confuse Medicaid and Medicare

In addition to these issues, the type-by-type reports noted:

- the questionnaire structure caused major reporting and comprehension problems

- respondents covered by public plans don’t necessarily use the same terms for that coverage as those specified in the questions
- state-specific names for public programs can sometimes backfire, leading the respondent to think the question is asking a more narrow question, or generally causing the respondent to lose confidence in reporting their public coverage
- respondents confuse Medicaid and SCHIP
- respondents do not necessarily attend to the 12-month reference period
- respondents who share the same type of coverage with another household member may be better able to report that coverage than for a household member with different coverage

In sum, there seem to be significant problems with individual questions on particular plan types, and these problems are compounded by the fact that in the type-by-type design respondents are blind to the structure of the series and have a tendency to try and “fit” their coverage in the earliest question that may be appropriate. However, many of these same problems were observed in the status-type design where respondents do have a better sense of the entire list of plan type options, either through a self-administered questionnaire, a flashcard or a respondent reading down a list of plan types.

Both the status-type and the type-by-type design exhibited overlapping problems stemming from the general disconnect between respondents’ understanding of their coverage and the terms used in the questionnaire. Thus recommendations for the type-by-type design are similar in structure to those for the status-type design. Specifically, a more promising design may be one that “decomposes” the reporting tasks into more manageable parts – starting off with a global question on status and, for those covered, moving to a question on general source of coverage (employment, government, etc.) and then moving on to questions tailored to each of these paths to collect the detail.

The type-by-type reports also brought to light issues associated with two dimensions not noted in the status-type designs: reporting for other household members and time. On the latter issue, findings on the calendar year reference period were fairly conclusive and suggest that the phrase “At any time during ...” is inadequate to generate reports of past coverage. There were several issues surrounding reporting for other household members: respondents sometimes weren’t sure which household members to report for, some respondents had only limited knowledge about other household members’ coverage, and some seemed to have more of a challenge when other household members’ coverage was different from their own. Thus another design choice in the spirit of decomposition would be to first ask about the respondent’s own coverage, and later ask about other household members, one at a time. That is, once the respondent goes through the series for him/herself, their own coverage could serve as a kind of anchor around which other household members’ coverage could be asked about. For example, if a respondent reports that she is on Medicaid a followup question could ask “Who else in this household is also covered by Medicaid.” Many (though not all) of these recommendations are embodied in the three experimental designs. EXP3, in particular, explores a different way of asking about retrospective coverage over the past 12 months.

5. EXPERIMENTAL DESIGNS

5.1 General Structure of the Questionnaires

All four of the experimental designs (UMASS, EXP1, EXP2 and EXP3) were structured in a similar way, to some extent. Specifically, the first question asked about coverage status (though EXP2 and EXP3 included an age screener and for respondents 65+ first a question on Medicare was asked). For respondents who said “no,” followup questions were asked on plans types typically underreported, and a verification question was asked to be sure those not reported as covered really were uninsured. In EXP1, EXP2 and EXP3, for respondents who reported some kind of coverage, a series of followup questions was asked to determine the type of coverage. First a question on general source of coverage (employment, government, etc.) was asked. Those in the employment path were asked policyholder and dependent questions, and those in the government path were asked type of government coverage. Once the specific plan type was identified, a question was asked to determine whether other household members were also covered by that same plan. In terms of reference period, UMASS, EXP1 and EXP3 all began with a current reference period, and EXP2 asked about coverage “at any time during the last 12 months.” Once a specific plan type was identified, EXP1 and EXP3 both included questions on monthly coverage in the past 4 months and the past 15-17 months, respectively.

One other point should be made about the general structure of the questionnaires. The EXP1 design was meant to be as “user-friendly” as possible and so started the series by asking respondents what the “easiest way” would be for them to report health insurance. The specific question wording was:

> Next I'd like to ask you about health insurance coverage over the past four months for all the people living in this household. What's the easiest way for you to tell me about that?
> READ IF NECESSARY: Would it be easiest to go person-by-person, one policy at a time, or some other way?
> □ Person-by-person
> □ One policy at a time
> □ My policy only (DK about other hh members)
> □ Some other way/doesn't matter
> □ Everyone is uninsured

Perhaps not surprisingly, this question was a tremendous failure. Respondents generally did not know what to make of the stem question, and when response categories were read they still were not quite sure how to answer. Most respondents generally let the interviewer decide because they seemed to be confused about what the question was getting at. Ultimately about a third of respondents went down the “person-by-person” path and two thirds went down the “one policy at a time” path.

5.2 Followup Questions on Underreported Plan Types

The initial question on coverage status was discussed above in Section 3.1

5.2.1 Question Wording

Three different questionnaire versions included at least one follow up question to capture coverage that may have been missed in the global question on status:

UMass: Do you currently have any health insurance coverage through government programs such as Medicare, Medicaid or MassHealth?

EXP2: 1. Are you covered by Medicaid, [Medicaid and SCHIP state names], Medical Assistance, or any other kind of government assistance program that helps pay for health care?
2. Do you get a medical card from the state health or welfare department so you can go to the doctor or hospital?

EXP3: 1. Are you covered by Medicaid, Medical Assistance, S-CHIP, or any other kind of government assistance program that helps pay for health care?
2. [If not already asked] Are you covered by Medicare?
3. Are you covered by
DC: DC Healthy Families, DC Healthcare Alliance, the State Child Health Plan or Medical Charities?
MD: Health Choice or the Maryland Children's Health Program?
VA: FAMIS Plus?

5.2.2 Results

In cognitive testing of the UMASS version, findings showed that even on a very small set of subjects, coverage not previously reported was captured. A subsequent pretest with a larger sample (n=33), which was partially seeded with respondents whose coverage status was known, was shown to correctly capture coverage that was not reported in the earlier global question on coverage. In EXP2 and EXP3, there was no evidence of respondent comprehension problems, but none of the items served to capture people who were initially reported as uninsured but were actually covered by some type of public program. A number of factors could explain this in part: the relatively small number of respondents tested in the cognitive lab, the low prevalence of respondents covered by a public plan, and/or the strong "brand name" associated with the particular Massachusetts name for Medicaid (MassHealth). In EXP2, when probed about the state-specific Medicaid and S-CHIP names, none of the respondents was familiar with any of these plans and, incidentally, none had heard of the generic "S-CHIP." A later (and larger) pretest of EXP3, however, revealed that some respondents did report coverage only in response to followup questions which included state-specific program names (Pascale, 2009).

5.2.3 Summary

The evidence appears rather mixed with regard to the utility of state-specific program names (see Sections 4.1.2 and 5.6 for a more thorough discussion of program names). The bottom line seems to be that sometimes the state-specific name aids in reporting for some respondents, but often respondents don't recognize these names and they do not aid in reporting. Further evidence (discussed below) also reveals that providing too many state-specific names can confuse respondents and possibly even induce underreporting. These findings are likely related to the extreme variability of naming conventions at the state-level, and the extent to which the program names are in flux over time. Given the demonstrated underreporting of public plan coverage, however, prompting correct reporting of public plan coverage is not an issue that should be easily dismissed. Therefore the EXP3 version may offer the best approach so far – asking about Medicare, Medicaid, S-CHIP and any other government plan in a fairly generic way, and only if no coverage is reported, asking a followup question which supplies state-specific program names.

5.3 Verification Question on Uninsured

All four versions included a question to verify that those for whom coverage had not been reported were, in fact, uninsured:

UMass: So you currently do *not* have *any* health insurance coverage at all. Is that correct?
EXP1: I have recorded that NAME(s) does not have health insurance coverage of any kind. Is that correct?
EXP2: OK, I have recorded that NAME was not covered by any kind of health insurance at any time during the past 12 months. Is that correct?
EXP3: OK, I have recorded that you are not covered by any kind of health plan or health coverage. Is that correct?

Results from all four versions seemed to indicate this question posed no comprehension problems, and that it worked as intended. Some respondents simply confirmed that the person in question was, in fact, uninsured. Other respondents said they just did not know about the health insurance status of other household members, and were reluctant to report anything. Given the lack of problems detected, and the proper functioning of this item, no further testing is recommended.

5.4 Question on General Source of Coverage

5.4.1 Question Wording

The question on general source of coverage varied slightly across versions:

EXP1: How [do you/did NAME] get the health coverage – [is/was] it through an employer or union, through the government, purchased on your own, or some other way?
EXP2: Is that [card/coverage] provided through a job or union, through the government, is it purchased directly from the insurance company, or is it obtained some other way?
EXP3: Is that coverage provided through an employer or union, the government, or some other way?

5.4.2 Results

Results from all three tests showed that most respondents had no difficulty with the basic categories (job-based, government, etc.) and most offered reasonable paraphrases of what each of the general sources of coverage meant. Indeed many respondents found it a very straightforward, easy question to answer.

One problem detected in all tests, however, had to do with respondents who held a job with some kind of government agency. Some had difficulty deciding between the "government" and the "job" categories. In EXP1 two respondents noted that the "government" category made them think of job-based coverage through a government employer. One offered the following paraphrase: “If you’re a government employee, or on a subsidized government insurance program like Medicaid.” Another said the response categories “started me thinking about my husband’s employment (he was a government worker) and I wondered if that would be a government-sponsored plan.” In EXP2 and EXP3 many respondents who were government employees (federal as well as state and county workers) expressed some doubt about which category was “correct,” and even when the interviewer note was read it did not seem to clear up the confusion.

There were a few other miscellaneous problems detected in EXP1 as well. One respondent said she has “medical assistance” but didn’t understand the question, and so did not choose “government.” Another (a student) had university-based coverage but said her university is not really her “employer” and that she didn’t really purchase it “on her own.” In EXP3 the response categories were further simplified to just employer, government and other, and followup questions were modified to capture direct-purchase plans and other miscellaneous plans such as school-based coverage. In EXP3 respondents with coverage through a parent’s or spouse’s employer (versus their own employer) had trouble choosing between “employer” and “other” (the “other” being their parent or spouse), since it was not *their own* job that supplied the coverage. Respondents with retiree coverage also had trouble choosing between “employer” and “other” since they were no longer working.

5.4.3 Summary

It is promising that this general approach seemed to be effective for respondents in many different situations. The main problem detected is that when the general source of coverage is cut down to just a few major categories, some unavoidable overlap and hence ambiguity is introduced (eg: (job/government for government workers; job/other for retired enrollees; job/other for dependents on someone else’s job-based plan). However, this kind of ambiguity can be handled in followup questions that can accommodate any number of choices a respondent might make if they have difficulty choosing the “right” category.

For example, the problem that government workers were having in choosing between “job” and “government” was detected early on in the EXP2 test (it had arisen in EXP1 as well). Thus a new question was introduced during mid-testing of EXP2 which was a followup for respondents who chose the “government” category: “Is that coverage was related to a *job* with the government?” Those who say yes – that their “government” coverage is actually related to a job -- are then routed down the job-based path. Only a few respondents received this new item in EXP2 but no problems were detected with it. The question was then re-tested on all respondents in EXP3 and the only issue that arose had to do with former government employees. Some were thrown off by the present tense “is” (the coverage through a job with the government) since they were no longer working at that job. This could be addressed by asking “Is or was that coverage related...”

Regarding other overlaps and ambiguities, similar paths were developed in EXP3. For example, those who chose “other” were asked an initial followup question to determine if the coverage was from a parent/spouse, direct purchase, school, or some other way. Further refinements were made mid-testing in the EXP3 version and a pretest was conducted later and no problems were detected with these refinements.

5.5 Questions on Policyholders and Dependents

5.5.1 Question Wording

The wording of items identifying policyholders and dependents varied somewhat across versions:

EXP1 (person-by-person path):

5. [Do you/Did NAME] get the coverage through your/his/her own employer/union or through someone else's [employer/union]?
6. Who is that? [Whose employer or union provides your health coverage]?
7. [Does your/Did NAME's] policy cover anyone else in this household?

EXP1 (policy-by-policy path):

14. Who is the policyholder for that plan? (In whose name is the policy written)?
15. In addition to [POLICYHOLDER NAME], is anyone else in this household covered by his/her/your policy?

EXP2:
15. Who has the job that provides this insurance policy?
16. Who purchases the insurance policy?
17. Is anyone else (within this household) covered by [your plan/the person not living here]?
18. Who else is covered (by your plan/the person not living here]?

EXP3:
15. And who is the policyholder?
23. Is anyone else within this household also covered by [if job-based or directly-purchased fill: your/policyholder's/else fill name of plan (e.g.: that Medicaid, Medicare, VA] plan?
24. Who?

5.5.2 Results

In EXP1 respondents who reported job-based coverage were asked to identify the policyholder and dependents and no problems were detected with any of these questions. In EXP2 and EXP3, however, the question "Who was the policyholder" surprisingly caused problems for a number of respondents. Many provided the name of an entity (either an employer or a health plan), rather than the name of a family member. Some of these respondents gave reasonable definitions of the term policyholder (even though they provided an incorrect response), but many gave poor definitions ("the one who carries my insurance," "the person in charge of everything...I don't know if it's an individual or a company"). These misunderstandings did not seem to be related to the age or education of the respondent. The EXP2 instrument was thus modified to be more direct. For the job-based path the new question read: "Who had the job that provided the insurance policy?" and for the direct-purchase path the question read: "Who purchases the insurance policy?" The few respondents who received these new versions seemed to have no difficulty.

The items on dependents in EXP2 and EXP3 were unproblematic and respondents reported a range of situations – in some cases the policy covered all other household members, in some only the children were covered (the spouse had his/her own coverage) and in some cases only the respondent was covered.

5.5.3 Summary

While the term "policyholder" seemed to be problematic for some respondents, in most automated instruments only a household roster will appear (possibly with one response category for "someone outside the household") so if respondents do offer an entity rather than an individual, interviewers will be prompted to probe for an individual given their response category choices. No evidence on the alternative questions ("Who had the job...Who purchased the policy...") but these were tested an a relatively small number of respondents.

5.6 Question on Type of Government/Other Plan

5.6.1 Question Wording

All three experimental versions asked a similar basic question of respondents who reported government coverage – asking what type of plan it is, and including Medicare, Medicaid, SCHIP, VA/military plans, and "other" as response categories. All versions also embedded state specific program names in the response categories, and EXP2 also embedded these names into the question stem itself. And finally, EXP2 also had an intervening open-text question on plan name which was asked between the item on general source of coverage (job, government, etc.) and the specific type of government/other coverage.

5.6.2 Results

For the most part respondents in EXP1 seemed to put themselves in the correct category, or they offered sufficient detail to indicate that the coverage was a public assistance plan of some sort. The reporting task was perhaps not as straightforward as it could have been because response categories were not necessarily read. Respondents, then, offered a range of names for their coverage. Some reported generic-type names (Medicaid, medical assistance), one said "it's part of social services - an HMO called Free State", and one said her children were covered by "the health department." Another respondent said "TANF, social service, AFDC, whatever they call themselves now" but did not recognize any of the state-specific names listed. Other respondents provided state-specific names with no prompting or hesitation (Charter Health Plan, Americaid, Medallion, etc.), but in many cases these state-specific names were not the same as those names embedded in the question. When probed about whether they considered these to be Medicaid plans, one said "don't know" but all the others said "yes." And perhaps inevitably, some respondents confused Medicaid and Medicare.

In EXP2, again respondents were asked both an open-text question on the plan name and then the question on type of coverage, with the response categories (including state-specific names) embedded in the stem. For many on Medicare and Medicaid, the reporting was straightforward and unproblematic. They reported simply Medicaid or Medicare at the open-text question on plan name, and then some offered more detail in the next question on plan type. For example, those on Medicare sometimes said Part A or B, and some Medicaid enrollees offered the commercial provider as well (e.g. Advantage Health Plan).

As was observed in other reports, the most substantial problems stemmed from the use of state-specific program names. One particularly interesting case had to do with the Medicaid/SCHIP distinction. In the open text question, the respondent reported his coverage as FAMIS. He said he had heard of Medicaid and Medallion but those programs were for people who earn less than his family does, offering "your health plan...there's a scale, based on wages." As we later learned, FAMIS is the state-specific name for SCHIP in Virginia; it was simply lacking from the questionnaire. And though the respondent clearly understood the system and knew he was not on Medicaid, he had not heard of SCHIP per se, only the actual state-specific name of SCHIP in his state. Thus, at the item on what type of government coverage, he was unsure which category to choose because he'd never heard of SCHIP and he did not see FAMIS on the list. So essentially he was a perfectly educated respondent but the questionnaire was flawed and failed to tap into his knowledge.

Another respondent had been enrolled in two different public programs during the previous year - a "Mothers and Babies" program which the hospital helped enroll her in when she gave birth (the program only lasted a few months), and "regular Medicaid" when that program expired. Because the "Mothers and Babies" program name was not on the list, though, she chose "Maryland Children's Health Program" (the state-specific name for SCHIP) because it seemed close. However, given her narrative, it seems unlikely that her child was ever on SCHIP, but rather a short term pregnancy-related program, and then Medicaid. In this case the disconnect between the program names embedded in the question and the respondent's own understanding of her coverage led to a false positive on SCHIP, though no under reporting of Medicaid.

Finally, in the EXP2 the respondents who had reported "other" as their general source gave reasonable answers when asked about the coverage here. One said it was her fiance's coverage and she knew no details, another said it was university-based coverage. A third (discussed below) ultimately reported Medicare.

EXP3 employed a more stream-lined approach and no comprehension or reporting problems were detected.

5.6.3 Summary

Taken as a whole these findings indicate that the terms respondents use for their public plans cover a wide range, and it suggests that crafting the "correct" list of state-specific names could be an elusive and possibly unattainable goal. Clearly some respondents think of public coverage in traditional terms (regular Medicaid, social services, regardless of "whatever they are called now"), some know the generic Medicaid name as well as the commercial provider, and some know only the specific name of the program they signed up for. However, if the name the respondent uses for the program is not listed in the question, it's unclear whether they would report correctly.

Due to all the uncertainty over whether the questions provide the right cues for reporting of public coverage, a followup question may be necessary for those respondents who simply don't find "their" coverage on the list of response categories. Indeed in mid-testing of EXP2 another change was made in cases where respondent did not choose Medicare, Medicaid or SCHIP. The question simply asked whether the coverage was some type of medical assistance plan. No problems were detected with this approach. It was since refined in EXP3 and no problems were detected in the pretest.

5.7 Public Coverage "Dependents"

Though obviously public coverage is not administered in terms of policyholders and dependents, in many households multiple members are covered by the same type of public coverage. To capitalize on this, in the EXP1/Policy Path, EXP2 and EXP3 designs, once the respondent reports that one person is covered by a public plan, a followup question is asked to determine who else within the household is also covered by that plan. This question was shown to be unproblematic in EXP1 where in most cases everyone in the household was indeed covered by the same plan. No serious problems were detected in EXP2, though one respondent asked "about my friends or just my family?" This prompted yet another slight change to the questionnaire mid-testing. Rather than simply ask "Who else is covered" a screener was added: "Was anyone else (within this household) also covered by [plan name]?" and if yes the next item collected the names. This version was tested in EXP3 and no problems were detected.

5.8 Medicare

In EXP2 several respondents who had Medicare also had coverage through other plans (QMB and job-based coverage). All these respondents reported their supplemental plan first, and one respondent ultimately never did report her Medicare. Because she had mentioned having Medicare while giving details of her payment arrangements for her supplemental job-based coverage, though, the interviewer proceeded to ask about the Medicare plan. When the interviewer asked whether she considered it Medicare, Medicaid, SCHIP or something else, she said "other" and explained that she got Medicare when she turned 65. That is, this respondent interpreted the question as asking why she was eligible, not what plan type she had. In sum these findings seem to bolster those above regarding respondents' dismissal of some types of coverage, in this case many respondents seemed to discount their Medicare coverage. For this reason a change was made to the questionnaire mid-test. Rather than the series beginning with the simple global question, an age and disability screener was developed and those who screened in were asked the Medicare question. Those who said "no" to the Medicare question were then routed to the global question on any kind of coverage.

In EXP3 this same approach was taken and no problems were detected, other than the known problem that some respondents confuse Medicare and Medicaid. Adding a definition to the Medicare question is advisable.

5.9 Military Coverage

Following the lead of the CPS, this question was originally worded in EXP2 with several military program names embedded:

> Was this plan through TRICARE, CHAMP VA, VA, military health care, or the Indian Health Service?

The question seemed to create unnecessary cognitive burden for respondents, and they struggled to understand the question and determine if their coverage fit the situation. Following are some examples:

- I've heard of those terms - VA, TRICARE - I think that's military but I haven't heard of the others. This kind of threw me because it didn't include anything I was truly familiar with.
- I've never heard of any of them; I didn't understand.
- I was trying to think really hard what my plan is actually under. Trying to remember which plan I have...what it's called. Trying to see if I matched any of the ones...
- Thinking about whether or not I was going to say "Optimum Choice" [the private plan]
- I was trying to figure out which one to answer...none of these was familiar...did have some doubts about my answer, and mostly was confused about what they were.

One respondent was confused to the point of misreporting. The following are the interviewer's notes:

- Her first reaction was "I don't understand those," but she then went on to explain why she thought her plan must be TriCare. She was convinced that TRICARE referred to the tri-county area plan that she had heard about through her work, through the book she used to sign up for her health insurance. She definitely had doubts here, and did not understand what the other options were at all.

Because of the burden and confusion caused by this question, it was modified mid-testing into two parts. The first question was a much simpler and direct way of determining whether there was any connection between the plan and the military or Indian Health Service. If yes, a follow-up question was added to identify the particular type of military or Indian Health Service coverage.

- Part 1: Was this plan related to the military or the Indian Health Service in any way?
- Part 2: What plan are you covered by? Is it TRICARE CHAMPVA, VA, military health care, the Indian Health Service, or something else?

Several respondents were asked this simplified version of the question, and none exhibited any comprehension difficulties. It was later refined somewhat in EXP3 as:

Is that plan related to military service in any way?

No evidence of comprehension or reporting problems was detected.

5.10 Additional Plans

In EXP1, EXP2 and EXP3 there was a catch-all question to determine if there were any plans in addition to those already reported. It did not appear to create any difficulties for respondents in any of the tests. Note that in EXP2, at this point in the instrument, respondents had the opportunity to report multiple plans covering any one person. For example, in the question on general source of coverage respondents were permitted to choose more than one category, and details on each plan type would have been asked in sequence before getting to this catch-all "additional plans" question. Similarly, for government plans, respondents were asked what type of government plan and they were permitted to choose more than one category (Medicare, Medicaid, SCHIP, etc.). Again, if they did choose multiple categories, each plan type would have been asked about in sequence. Also, any individual could have been reported as a dependent on multiple household members' plans, if questions were asked about those policyholders prior to questions about the dependent.

Among the households with individuals who ended up reported as covered by multiple plans, half reported those additional plans through one of the means described above, and the other half reported them at the catch-all question. There was a range of plan types reported at the catch-all item (secondary plans for teenagers, Medicare, Medicaid). In EXP3 there was a slightly different structure, and at the questions on general source of coverage and type of government plan respondents were allowed to choose only one option. The question on any additional plans did not appear to pose any problems for these respondents, though some did mention single-service plans. Though this question did include a probe telling respondents to exclude single-service plans, later iterations of EXP3 embedded this instruction into the question itself, and no problems were detected.

5.11 Reference Period

In EXP1, which employed a 4-month reference period, respondents were simply asked what month the coverage began and ended, and no problems were detected. In EXP3, a different approach was taken. Respondents with current coverage were asked at the time of the interview (March thru May) whether that coverage began before or after January 1 of the previous calendar year (thus covering a time span between 15 and 17 months). If after, a question determined the start month. Followup questions then determined whether the coverage was continuous from the reported start date, and thus monthly coverage was derived. No comprehension or reporting problems were detected with this approach, in either the cognitive testing or in the later pretest of EXP3.

6. SUMMARY OF REPORTING ISSUES

Below is a summary of the major reporting problems detected.

General Source of Coverage. The main problem here had to do with overlap between categories (e.g.: employees of government agencies who were uncertain whether to choose the job-based or the government category). Interviewer instructions did not always help, and these types of instructions are questionable because it requires the respondent to express some doubt and the interviewer to notice this and read the instruction at the appropriate time. Therefore, the remedy seems to be to develop follow up questions that accommodate respondents' doubts. For example, all respondents who choose "government" as their general source of coverage should be asked if the coverage is related to a job of some kind, and if so, they should be rerouted down the job-based path.

Type of Government Plan. This item, while straightforward for some, caused difficulty for others and the reasons seem to mainly stem from the wide variety of names that public programs are known by. While more effort could be expended on a state-by-state basis to offer a more current list of program names to respondents, there is no guarantee that such a list would ever be comprehensive. Program names seem to be constantly in flux, in part as a response to legislation and in part as a result of states making arrangements with commercial providers. Furthermore, there is no real evidence that such a comprehensive current list would help in all cases. Many respondents simply refer to the programs by their more generic names (Medicaid, medical assistance, etc.), and as was shown in other tests above, sometimes offering some names but not the respondent's particular name can lead to underreporting.

Therefore, in terms of crafting the actual question and deciding on which names to embed, the best approach may be to ask questions using generic terms (e.g.: Medicaid) and follow up on an "as needed" basis with questions using state-specific names only if the answer is "no" to the more generic question. Questionnaire designers should also work closely with state program administrators to develop a best attempt at a comprehensive list of program names.

Medicare. Also consistent with earlier findings, respondents on Medicare did not always seem to have this coverage in their minds, and some were distracted by secondary/supplemental plans they had in addition to their Medicare. Given the extremely high correspondence between age and Medicare coverage, it is recommended to ask age-eligible (and disabled) respondents directly whether they are covered by Medicare first (and include a definition of Medicare), and then ask about additional plans.

Military. Presenting respondents with a rather lengthy list of what may be perceived as arcane program names seems only to confuse them and may lead to misreporting. A more effective method was to split the question in two and first ask a simple question about coverage "related to military service" and if yes, ask a followup question to determine particular type.

7. CONCLUSIONS

There are a number of cognitive demands on respondents when answering these types of questions, the most basic of which is for them to determine how their conception of their coverage maps on to the terms for that coverage used in the questionnaire. All reports examined here suggest that the status quo designs hinder accurate reporting by generally imposing the categories of interest for analysis without regard to how respondents think about their coverage. Encouraging findings, though, were that many respondents do seem to have fairly good knowledge of their own coverage, and that a modified design was demonstrated to evoke few reporting problems. This suggests that the specific categories of analytic interest can be derived from a set of questions that decomposes the reporting task and presents respondents with questions in terms they can understand.

More specifically, the earliest available tests (BRFSS, NHIS and MEPS) all suggested that an improved design would be one that asked a global question first, followed by a question general source of coverage. The EXP1, EXP2 and EXP3 versions followed this recommendation and fleshed out – in an iterative fashion – the particular wording, response categories, and routing path that each response would follow. By the end of testing of EXP3 there were no more major problems detected with this new design.

Two particular design features deserve more comment – the reference period and the household- versus person-level design. Regarding the latter, previous empirical research has shown that asking questions at a person-level results in more reporting of overall coverage (Blumberg et al, 2004; Hess et al, 2001), possibly because respondents are prompted with each individual household member's name (versus a more generic "anyone in this household"). However, research also shows that in large households (four or more people) the person-level design results in fewer reports of Medicaid (Pascale, 2000), possibly due to an exceptionally tedious, burdensome questionnaire. Blumberg et al also found that larger administration times were related to higher rates of uninsurance and suggested that "respondent fatigue may contribute to higher uninsurance rates" (Blumberg et al, 2004). In response to these findings, the EXP2 and EXP3 designs employed a kind of hybrid household-person approach. It begins by asking about the respondent's own coverage, and any time a respondent reports a plan type, a followup question is asked to determine whether other household members are also covered. After the entire series is finished for the respondent, the series is repeated for the next person on the roster, but to avoid respondent fatigue, the information previously reported is harnessed. Specifically, if the second household member was reported to have job-based coverage (e.g.: as a dependent on the respondent's plan), a question is first asked to determine whether the second person has any coverage in addition to that dependent coverage. If not then no more questions are asked about that person. Ultimately the design has the advantage of prompting the respondent with individual names of all household members, but in many cases the length of the questionnaire is drastically reduced since many household members share the same type of coverage.

Regarding the 12 month reference period as specified in EXP2, given that some respondents clearly do not attend to the reference period but rather think about their current status and spell, there seems to be no guarantee that analysts can interpret respondents' answers as if they apply to the entire 12 month period. The approach explored in EXP3 appeared to be effective and could prove to be a promising new method of asking about retrospective coverage. Further testing on accuracy of reporting (e.g.: through administrative records) would shed light on whether this approach does indeed offer improvements over the status quo.

8. FUTURE RESEARCH

Given the iterative nature of all four experimental versions, and the final pretest of EXP3 (in which no major flaws were detected), the final EXP3 design may represent a "best practices" approach at this time. It is important to note, however, that cognitive testing is designed to identify *sources* of comprehension and misreporting problems, but not their *prevalence*. Furthermore, because the lab is a somewhat artificial setting, it often means there is interaction between an interviewer and respondent in the midst of the question series. Thus the role that a production interviewer may play in repairing any misunderstandings is simply not detected in the cognitive lab. However, the range of problems in the 11 reports gathered in this paper, and the reasons for respondents' misinterpretation and confusion, suggest that misreporting in a production setting could be non-trivial. Thus a large scale pretest or field test – including a validation component – is recommended to determine whether this misreporting manifests often enough to seriously and systematically affect the estimates.

REFERENCES

Beatty, Paul, and Schechter, S. (1998), Questionnaire evaluation and testing in support of the Behavioral Risk Factor Surveillance System (BRFSS), 1992-98, Office of Research and Methodology, NCHS, Working paper series, No. 26, p. 12-17

Beatty, Paul, Wilson, B., Miller, K., Calvillo, A., Whitaker, K. (2002), NHIS Insurance Module and additional questions on Work, Income, and Health Care Utilization, Unpublished report, April, 2002

Blumberg, Steve J., Osborn L., Luke J.V., et al (2004). "Estimating the Prevalence of Uninsured Children: An Evaluation of Data from the National Survey of Children with Special Health Care Needs, 2001." National Center for Health Statistics. Vital Health Stat 2(136). Hyattsville, MD.

Cantor, David, personal communication, August 30, 2000

Hess, Jennifer, Moore, J. Pascale, J. Rothgeb, J. and Keeley, C. (2001), "The Effects of Person-level vs. Household-level Questionnaire Design on Survey Estimates and Data Quality." Public Opinion Quarterly, Winter 2001, 65:574-584.

Kerwin, Jeffrey, Cantor, D., Sheridan, S. (1995) "Results of Rounds 3 and 4 of Managed Care Cognitive Interviews for the Household Portion of NMES."

Loomis, Laura (2000). "Report on Cognitive Interview Research Results for Questions on Welfare Reform Benefits and Government Health Insurance for the March 2001 Income Supplement to the CPS," Center for Survey Methods Research, Statistical Research Division U.S. Census Bureau, July 25.

Pascale, Joanne. "Findings from a Pretest of a New Approach to Measuring Health Insurance in the Current Population Survey." Paper prepared for the Federal Committee on Statistical Methodology Research Conference, November 2-4, 2009.

Pascale, Joanne (2006). "Measuring Health Insurance in the U.S." Proceedings from the Annual Meeting of the American Association for Public Opinion Research, Montreal, Quebec, Canada

Pascale, Joanne (2005). "American Community Survey: Cognitive Testing of Health Insurance Questions." Unpublished Census Bureau Report, March, 2005.

Pascale, Joanne (2003), "Questionnaire Design Experimental Research Survey (QDERS) 2004: Cognitive Testing Results on Health Insurance Questions." Unpublished Census Bureau Report, November 5, 2003.

Pascale, Joanne (2001), “Summary of Cognitive Testing of Experimental Questions on Health Insurance in the SIPP Methods Panel: Wave 1, Replicate 2.” Unpublished Census Bureau report, January, 2001.

Pascale, Joanne (2000). “Alternative Questionnaire Design Strategies for Measuring Medicaid Participation.” Paper presented at the American Public Health Association, Boston.

Roman, Anthony M., Hauser, A., Lischko, A. (2002), "Measurement of the Insured Population: the Massachusetts Experience." Paper presented at the 2002 Annual Meetings of the American Association for Public Opinion Research. St. Pete's Beach, Fla., May, 2002.

Willson, Stephanie (2005), “Cognitive Interviewing Evaluation of the National Immunization Survey Insurance Module: Results of Fieldwork and Laboratory Interviews” Unpublished report.

APPENDIX A: INSTRUMENTS USED IN COGNITIVE TESTING

A. BRFSS (1995)

1. Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?
□ Yes => 2
□ No

2. How do you obtain the health care coverage you use to pay for <u>MOST</u> of your medical care? Is it through: [PLEASE READ]
□ your employer
□ someone else's employer
□ a plan that you or another family member buys on your own
□ Medicare
□ Medicaid/state name
□ another federal program such as the military, CHAMPUS or the VA or
□ some other source
□ none

B. NHIS (2002)

1. [Are you/Is anyone] covered by any kind of health insurance or some other kind of health care plan? READ IF NECESSARY: Include health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills.

□ Yes => 2
□ No

2. What kind of health insurance or health care coverage [do you/does NAME] have? INCLUDE those that pay for only one type of service (nursing home care, accidents or dental care), exclude private plans that only provide extra cash while hospitalized.

□ Private health insurance plan from employer or workplace
□ Private health insurance plan purchase directly
□ Private health insurance plan through a state or local government program or community program
□ Medicare
□ Medi-gap
□ Medicaid
□ CHIP (Children’s Health Insurance Program)
□ Military health care/VA
□ Tricare/CHAMPUS/CHAMP-VA
□ Indian Health Service
□ State-sponsored health plan
□ Other government program
□ Single service plan (eg: dental, vision, prescriptions)
□ No coverage of any type

C. ACS (2005)

1. Is this person CURRENTLY covered by any type of health insurance? Include insurance obtained through a job or purchased directly from the insurance company, and government health insurance such as Medicare, Medicaid, VA and military programs.

2a. [Self-administered]: What type of health insurance does this person have?
Mark (X) all that apply.
□ Insurance through a current or former employer or union (of this person or another family member)
□ Insurance purchased directly from the insurance company (by this person or another family member)
□ Medicare, for persons 65 years old and older, or persons with certain disabilities
□ Medicaid, Medical Assistance, or any kind of government-assistance plan for low-income children and families
□ TRICARE, CHAMPUS or other military care
□ CHAMPVA or VA
□ Indian Health Service
□ Supplemental plans that cover one type of care (e.g.: dental, accident, nursing home care plans)
□ Other/specify____

2b. [CATI]: What type of health insurance does this person have? Is it...
...insurance through a current or former employer or union (of this person or another family member)?
...insurance purchased directly from the insurance company (by this person or another family member)?
...Medicare, for persons 65 years old and older, or persons with certain disabilities
...Medicaid, Medical Assistance, or any kind of government-assistance plan for low-income children and families
...TRICARE, CHAMPUS or other military care
...CHAMPVA or VA
...Indian Health Service
...Supplemental plans that cover one type of care (e.g.: dental, accident, nursing home care plans)
...Other/specify____
NOTE: Each response category used a discreet set of yes/no response categories.

2c. [CAPI] Next I am going to show you a list of health insurance categories [show Flashcard]. What type of health insurance does this person have? You may choose more than one.

FLASHCARD

What type of health insurance does this person have? You may choose more than one.
1. Insurance through a current or former employer or union (of this person or another family member)
2. Insurance purchased directly from the insurance company (by this person or another family member)
3. Medicare, for persons 65 years old and older, or persons with certain disabilities
4. Medicaid, Medical Assistance, or any kind of government-assistance plan for low- income children and families
5. TRICARE, CHAMPUS or other military care
6. CHAMPVA or VA
7. Indian Health Service
8. Supplemental plans that cover one type of care (e.g.: dental, accident, nursing home care plans)
9. Other

D. MEPS (November 1994-March 1995)

1. Medicare is a Social Security health insurance program for disabled person and for persons 65 years old or over. At any time during the last 4 months [have you/has anyone in your family] been covered by Medicare? [Yes/No]
   1a. Who is that?
2. [Medicaid/Medical Assistance] is a state program for low income persons or for persons on public assistance. Sometimes persons with very large medical bills and persons in nursing homes are also covered by [Medicaid/Medical Assistance]. At any time during the last 4 months [have you/has anyone in your family] been covered by [Medicaid/Medical Assistance]? [Yes/No]
   2a. Who is that?
3. At any time during the last 4 months [have you/has anyone in your family] been covered by any other public program that pays for medical care [for example [STATE PROGRAM NAME], a public program that pays for prescribed medicine]? [Yes/No]
   3a. Who is that?
4. CHAMPUS is a civilian health benefits program that provides coverage to the dependents of active duty and retired service members. CHAMPVA provides health benefits coverage to dependents and survivors of disabled veterans. At any time during the last 4 months [have you/has anyone in your family] been covered by CHAMPUS or CHAMP-VA? [Yes/No]
   4a. Who is that?

INTRODUCTION: These next questions are about (other) health insurance [you/your family] may have. [SHOW CARD]. Health insurance includes any of the kinds of insurance that are listed on this card, this is [REVIEW CARD].

5. At any time during the last 4 months [were you/was anyone in your family] been covered by health insurance that [you/they] got from [your/their] job? [Yes/No]
   5a. Who is covered by [PLAN]?
   5b. Who is the policyholder or main insured person?
6. At any time during the last 4 months [were you/was anyone in your family] been covered by health insurance that [you/they] got from a group or association such as a church or a club; or a professional, business or retirement association such as a union? [Yes/No]
   6a. Who is covered by [PLAN]?
   6b. Who is the policyholder or main insured person?
7. At any time during the last 4 months [were you/was anyone in your family] been covered by health insurance that [you/they] got from a school where [you/they] are a student? Please do not include policies that pay only for injuries caused by accidents? [Yes/No]
   7a. Who is covered by [PLAN]?
   7b. Who is the policyholder or main insured person?
8. At any time during the last 4 months [were you/was anyone in your family] been covered by health insurance that [you/they] purchased directly from an insurance company? [Yes/No]
   8a. Who is covered by [PLAN]?
   8b. Who is the policyholder or main insured person?
9. At any time during the last 4 months [were you/was anyone in your family] been covered by the health insurance of someone who does not live here? [Yes/No]
   9a. Who is covered by [PLAN]?
   9b. Who is the policyholder or main insured person?
10. At any time during the last 4 months did [you/anyone in your family] purchase health insurance from some other place we have not listed? [Yes/No]
    10a. Who is covered by [PLAN]?
    10b. Who is the policyholder or main insured person?

E. CPS 2000 (SCHIP question testing round)

1. At any time in 1999 [were you/was anyone in this household] covered by Medicaid/STATE NAME? READ IF NECESSARY: Medicaid/STATE NAME is the government assistance program that pays for health care.

2. [Asked if children not covered by Medicaid] In [STATE], the [STATE SCHIP NAME] program (also) helps families get health insurance for CHILDREN. (Just to be sure,) were CHILDREN'S NAMES covered by that program?

F. CPS 2005

1. These next questions are about health insurance coverage during the past 12 months. The questions apply to ALL persons of ALL ages. At any time during the past 12 months, (were you/was anyone in this household) covered by a health insurance plan provided through (their/your) current or former employer or union? [Yes/No]
(MILITARY HEALTH INSURANCE WILL BE COVERED LATER IN ANOTHER QUESTION.)
1a. Who in this household were policyholders?
PROBE: Anyone else?
1a1. In addition to (you/name), who else in this household was covered by (name's/your) plan?
PROBE: Anyone else?

2. At any time during the past 12 months, (were you/was anyone in this household) covered by a health insurance plan that (you/they) PURCHASED DIRECTLY FROM AN INSURANCE COMPANY, that is, not related to current or past employment? [Yes/No]
2a. Who in this household were policyholders?
PROBE: Anyone else?
2a1. In addition to (you/name), who else in this household was covered by (name's/your) plan?
PROBE: Anyone else?

3. At any time during the past 12 months, (were you/was anyone in this household) covered by the health plan of someone who does not live in this household? [Yes/No]
3a. Who was that?
PROBE: Anyone else?

4. At any time during the past 12 months, (were you/was anyone in this household) covered by Medicare? [Yes/No]
READ IF NECESSARY: Medicare is the health insurance for persons 65 years old and over or persons with disabilities.
4a. Who was that?
PROBE: Anyone else?

5. At any time during the past 12 months, (were you/was anyone in this household) covered by Medicaid/(fill state name)? [Yes/No]
READ IF NECESSARY: Medicaid/ (fill state name) is the government assistance program that pays for health care.
5a. Who was that?
PROBE: Anyone else?

6. In (state), the (fill state name) program (also) helps families get health insurance for CHILDREN. (Just to be sure,) Were any of the children in this household covered by that program? [Yes/No]
READ IF NECESSARY: (fill state CHIP pgm name) is the name of (state)'s CHIP program. It is the same as the Children's Health Insurance Program, which helps pay for children's health care.
6a. Who was that?
PROBE: Anyone else?

7. At any time during the past 12 months, (were you/was anyone in this household) covered by TRICARE, CHAMPUS, CHAMPVA, VA, military health care, or Indian Health Service? [Yes/No]
NOTE: "CHAMPVA" IS THE CIVILIAN HEALTH AND MEDICAL PROGRAM OF THE DEPARTMENT OF VETERAN'S AFFAIRS.
7a. Who was that?
PROBE: Anyone else?

8. Other than the plans I have already talked about, at any time during the past 12 months, was anyone in this household covered by a health insurance plan such as the (fill state name) plan or any other type of plan? [Yes/No]
8a. Who was that?
PROBE: Anyone else?

G. SLAITS (2005)

1. At this time is [CHILD] covered by health insurance that is provided through an employer or union or obtained directly from an insurance company?
READ ONLY IF NECESSARY: These plans may be provided in part or fully by a current employer, a former employer, a union, or a professional organization, or purchased directly by an individual.
IF ONLY PLAN NAME OFFERED, PROBE (READ IF NECESSARY): Is this insurance provided through an employer or union or obtained directly from an insurance company? Do not include dental, vision, school or accident insurance.
IF NECESSARY, TO HELP THE RESPONDENT DETERMINE WHAT KIND OF INSURANCE THEY HAVE, PROBE (READ IF NECESSARY): Did you get that insurance through an employer? Does it help pay for both doctor visits and hospital stays?

2. At this time is [CHILD] covered by Medicaid, a health insurance program for persons with certain income levels and persons with disabilities? [In this state, the program is sometimes called [STATE PROGRAM].

3. At this time is [CHILD] covered by the State Children's Health Insurance Program or S-CHIP? In this state, the program is sometimes called [STATE SCHIP NAME].

4. At this time is [CHILD] covered by the Indian Health Service?

5. At this time is [CHILD] covered by military health care, TRICARE, CHAMPUS or CHAMP-VA?

6. At this time is [CHILD] covered by any other kind of health insurance or health care plan that pays for services obtained from hospitals, doctors and other health professionals?

[note: all questions present yes/no answer categories]

H. UMASS (2002)

1. Do you currently have any kind of health insurance coverage at all?

   □ Yes
   □ No => 2

2. Do you currently have any health insurance coverage through government programs such as Medicare, Medicaid or MassHealth?

   □ Yes
   □ No => 2

3. So you currently do *not* have *any* health insurance coverage at all. Is that correct?

   □ Yes
   □ No

I. EXP1 (2002):

1. Next I'd like to ask you about health insurance coverage over the past four months for all the people living in this household. What's the easiest way for you to tell me about that?
READ IF NECESSARY: Would it be easiest to go person-by-person, one policy at a time, or some other way?
□ Person-by-person => 2
□ One policy at a time => 9
□ My policy only (dk about other hh members) => 2
□ Some other way/doesn't matter => 2
□ Everyone is uninsured => [verification question]

Person-by-Person Flow

2. [Are you/Is NAME] now covered by any kind of health insurance plan, HMO or government assistance health plan such as Medicare or Medicaid?
READ IF NECESSARY: Do not include supplemental or specialty plans that provide only one type of service, such as vision, dental, cancer or nursing home coverage.
□ Yes => 4
□ No => 3

3. [Were you/Was NAME] covered by any kind of health insurance plan at any time between MONTH1 1st and today?
□ Yes => 4
□ No => [verification question]

4. ASK OR VERIFY:
How [do you/did NAME] get the health coverage – [is/was] it through an employer or union, through the government, purchased on your own, or some other way?
□ Employer/union => 5
□ Government => 8
□ Purchased on own => 7
□ Some other way => [never selected]

5. ASK OR VERIFY:
[Do you/Did NAME] get the coverage through your/his/her own employer/union or through someone else's [employer/union]?
□ Own employer/union => 7
□ Someone else's employer/union => 6

6. Who is that? [Whose employer or union provides your health coverage]?
□ [fill name of hh member]
□ Someone outside household
=> 7

7. [Does your/Did NAME's] policy cover anyone else in this household?
□ Yes => Who?
□ No
=> 9

8. What type of government plan is/was it?

□ Medicare, for people 65 years old and older and people with certain disabilities
□ Medicaid, Medical Assistance [or fill state name], a government assistance plan for people in need
□ [fill state SCHIP program names]
□ A military health plan, such as TRICARE, CHAMPUS, VA or military health care;
□ Indian Health Service
□ Something else
=> 9

9. [Intervening questions on when coverage started and stopped (not shown)]

10. At any time between MONTH1 1st and today, [were you/was NAME] ALSO covered by any OTHER kind of health insurance plan (in addition to the plans you just told me about)?

□ Yes => [loop back to 4]
□ No => next person

Policy-by-Policy Flow

11. ASK OR VERIFY:

[Are you/Is anyone in this household] now covered by any kind of health insurance plan, HMO or government assistance health plan, such as Medicare or Medicaid?

READ IF NECESSARY: Do not include supplemental or specialty plans that provide only one type of service, such as vision, dental, cancer or nursing home coverage.

□ Yes => 13
□ No => 12

12. [Were you/Was anyone in this household] covered by any kind of health insurance plan at any time between MONTH1 1st and today?

□ Yes => 13
□ No => [verification question]

13. ASK OR VERIFY:

What type of plan is that – is it obtained through an employer or union, through the government, purchased on your own, or in some other way?

□ Employer/union => 14
□ Government => 16
□ Purchased on own => 14
□ Some other way => [never selected]

14. ASK OR VERIFY:

Who is the policyholder for that plan? (In whose name is the policy written)?

□ [fill hh member]
□ Someone outside hh
=> 15

15. In addition to [POLICYHOLDER NAME], is anyone else in this household covered by his/her/your policy?

□ Yes => Who?
□ No =>
=> 18

16. What type of government plan is it?
□ Medicare, for people 65 years old and older and people with certain disabilities
□ Medicaid, Medical Assistance [or fill state name], a government assistance plan for people in need
□ [fill state SCHIP program names]
□ A military health plan, such as TRICARE, CHAMPUS, VA or military health care;
□ Indian Health Service
□ Something else

17. Who in this household is covered by that plan?

18. [Intervening questions on when coverage started and stopped (not shown)]

19. At any time between MONTH1 1st and today, [were you/was anyone in this household] ALSO covered by any OTHER kind of health insurance plan (in addition to the plans you just told me about)?
□ Yes => [loop back to 13]
□ No => [end]

J. EXP2

1. These next questions are about health insurance coverage. First I'd like to ask you about yourself.
=> if 65+ or disabled go to 2; else go to 3

2. Are you covered by Medicare?
□ Yes => 13
□ No => 3

3. Are you covered by any type of health insurance? [Post-test recommendation: Do you have any type of health coverage or health plan?]
□ Yes => 6
□ No => 4

4. Are you covered by Medicaid, [Medicaid state name], Medical Assistance, or any other kind of government assistance program that helps pay for health care? [Post-test recommendation: Include a probe or separate item for Medicare (if the Medicare item above was not asked) and VA coverage]
□ Yes => 6
□ No => 5

5. Do you get a medical card from the state health or welfare department so you can go to the doctor or hospital?
□ Yes => 6
□ No => 22

6. Is that [card/coverage] provided through a job or union, through the government, is it purchased directly from the insurance company, or is it obtained some other way?
READ IF NECESSARY: If this coverage is provided through employment with the government or the military, consider that coverage through a job.
FR: CHECK ALL THAT APPLY. ENTER (N) FOR NO MORE.
□ Job, union or business => 15
□ Government => 7
□ Direct purchase from insurance company => 16
□ Other => 10

7. Is this coverage related to a **job** with the government?
□ Yes => 8
□ No => 10

8. (ASK OR VERIFY): Is this plan related to the military in any way?
□ Yes => 9
□ No => 15

9. (ASK OR VERIFY): What plan are you covered by? Is it TRICARE, CHAMPVA, VA, military health care, or something else?
□ TRICARE
□ CHAMPVA
□ VA
□ Military health care
□ Other (specify)
=> 15

10. What is the name of that [government] coverage or program?
READ IF NECESSARY: How do you refer to this coverage?
=> 11

11. (ASK OR VERIFY): Do you consider that Medicare, Medicaid or Medical Assistance, [Medicaid state name], [SCHIP state name] or something else? [post-test recommendation: add military and VA to question text and response categories]
READ IF NECESSARY: Medicare is for people 65 years old and older or people with certain disabilities; Medicaid is for low-income families, disabled and elderly people who require nursing home care; and [fill SCHIP state name] is for low-income families and children.
FR: CHECK ALL THAT APPLY
□ Medicare => 13
□ Medicaid or Medical Assistance => 13
□ [SCHIP state name] => 13
□ Other => 12

12. Is this a medical assistance-type plan? [post-test recommendation: reword as "government assistance-type plan"]
□ Yes
□ No
=> 13

13. Is anyone else (within this household) also covered by [PLAN TYPE]?
□ Yes => 14
□ No => 19

14. Who else is covered?
=> 21

15. Who has the job that provides this insurance policy?
READ IF NECESSARY: Who is the policyholder? In whose name is the policy?
FR: ENTER LINE NUMBER OF ALL POLICYHOLDERS
(O) Other(s) not currently living here
=> 17

16. Who purchases the insurance policy?
READ IF NECESSARY: Who is the policyholder? In whose name is the policy?
FR: ENTER LINE NUMBER OF ALL POLICYHOLDERS
(O) Other(s) not currently living here
=> 17

17. Is anyone else (within this household) covered by [your plan/the person not living here]?
□ Yes => 18
□ No => 19

18. Who else is covered (by your plan/the person not living here]?
=> 19

19. Is this plan related to the military or the Indian Health Service in any way?
□ Yes => 20
□ No => 21

20. What plan are you covered by? Was it TRICARE, CHAMPVA, VA, military health care, the Indian Health Service, or something else?
□ TRICARE
□ CHAMPVA
□ VA
□ Military health care
□ Indian Health Service
□ Other (specify)
=> 21

21. Other than the plans we have already talked about, are you also covered by any other type of health insurance?
□ Yes => 6
□ No => [next person]

22. OK, I have recorded that you are not covered by any kind of health insurance. Is that correct? [post-test recommendation: drop item #23 and instead go to item #6 and cycle through entire series]

23. What type of insurance are you covered by? Any other type of plan?
□ Medicare
□ Medicaid or Medical Assistance
□ TRICARE or CHAMPUS
□ CHAMPVA
□ VA health care
□ Military health care
□ SCHIP state name
□ Indian Health Service
□ Other government health care
□ Employer/union-provided (policyholder)
□ Employer/union-provided (as dependent)
□ Privately purchased (policyholder)
□ Privately purchased (as dependent)
□ Plan of someone outside the household
□ Other (specify)

K. EXP3

1. PERSON 1: **These next questions are about health insurance coverage.** [IF MULTI-PERSON HOUSEHOLD**: First I'd like to ask you about yourself.]**
PERSONS 2+: **Next I'd like to ask you about NAME.**
=> CK2

CK2:
- if NAME is 65+ => Q2
- else go to Q3

2 **Are you covered by Medicare?**
□ Yes => Q16
□ No => Q3
□ DK/REF => Q3

3 **Do you have any type of health plan or health coverage?**
□ Yes => Q8
□ No => Q4
□ DK/REF => Q4

4 **Are you covered by Medicaid, Medical Assistance, S-CHIP, or any other kind of government assistance program that helps pay for health care?**
□ Yes => Q16
□ No => CK5
□ DK/REF => CK5

CK5:
- If Medicare already asked go to Q6
- else go to Q5

5 **Are you covered by Medicare?**
□ Yes => Q16
□ No => Q6
□ DK/REF => Q6

6 **Are you covered by**
IN DC: **DC Healthy Families, DC Healthcare Alliance, the State Child Health Plan or Medical Charities?**
IN MARYLAND: **Health Choice or the Maryland Children's Health Program?**
IN VIRGINIA: FAMIS Plus?
□ Yes => Q16
□ No => Q7
□ DK/REF => Q7

7 **OK, I have recorded that you are not covered by any kind of health plan or health coverage. Is that correct?**
□ Yes (not covered) => Q28
□ No (covered) => Q8
□ DK/REF => Q28

8 **Is that coverage provided through an employer or union, the government, or some other way?**
PROBE: If this coverage is provided through employment with the government or the military, consider that coverage through an employer.
PROBE: "Employer/union" coverage includes coverage from someone's own employer or union as well as coverage from a spouse's or parent's employer or union. It also includes coverage through former employers and unions, and COBRA.
FR: CHECK ALL THAT APPLY
□ Employer, union or business (current or former) => Q15
□ Government => Q9
□ Other => Q14
□ DK/REF => Q13

9 **Is that coverage related to a JOB with the government?**
□ Yes => Q11
□ No => Q10
□ DK/REF => Q10

10 **(ASK OR VERIFY): What type of government plan is it – Medicare, Medicaid, Medical Assistance or S-CHIP, military or VA coverage, or something else?**
READ IF NECESSARY: **Some of the government programs in [STATE] are:**
IN DC: **DC Healthy Families, DC Healthcare Alliance, the State Child Health Plan and Medical Charities.**
IN MARYLAND: **Health Choice and the Maryland Children's Health Program.**
IN VIRGINIA: FAMIS Plus.
READ IF NECESSARY: **Medicare is for people 65 years old and older or people with certain disabilities; Medicaid is for low-income families, disabled and elderly people who require nursing home care; and S-CHIP is for low-income families and children.**
FR: CHECK ALL THAT APPLY
□ Medicare => CK16
□ Medicaid, Medical Assistance or S-CHIP => CK16
□ Military or VA => Q12
□ Other => Q13
□ DK/REF => Q13

11 **(ASK OR VERIFY): Is that plan related to military service in any way?**
□ Yes => Q12
□ No => Q15
□ DK/REF => Q15

12 **[Earlier you reported coverage through a military plan.] (ASK OR VERIFY): Which plan are you covered by? Is it TRICARE, CHAMPVA, VA, military health care, or something else?**
□ TRICARE
□ CHAMPVA
□ VA
□ Military health care
□ Other (specify)
□ DK/REF
=> CK16

13 **Is it a government assistance-type plan?**
□ Yes => CK16
□ No => CK16
□ DK/REF => Q27

14 **[Earlier you reported coverage through another plan.] How is that coverage provided? Is it through...**
□ **a parent or other relative**
□ **a college, university or school or**
□ **direct purchase from the insurance company or a trade association**
□ **or some other way?**
□ DK/REF
=> CK16

15 **And who is the policyholder?** [include "Someone outside household"]
Name of policyholder ____
=> CK16

CK16
- if this is a currently-held plan => Q16
- else if this is a plan not currently held but held at some point in 2007 or Q26=yes =>Q22

16 **Did that coverage start before or after January 1, 2007?**
[If this is a **job-based plan** fill: PROBE: When we say "that coverage" we mean any coverage through [policyholder's] employer. So if [policyholder] switched plans offered by the employer, or even switched employers, we still consider this all the same coverage.]
[If this is a **directly-purchased plan** fill: PROBE: When we say "that coverage" we mean any coverage directly purchased by you or another policyholder. So if you/NAME switched plans but they were all directly-purchased, we still consider this all the same coverage.]
□ Before January 1, 2007 => if Medicare =.> CK23; else => Q20
□ On or after January 1, 2007 => Q18
□ DK/REF => Q17

17 **Did you have the coverage at any time during 2007?**
□ Yes => Q22
□ No => CK23
□ DK/REF => CK23

18 **In what month did that coverage start?**
□ Month [1-12] in 2007 => Q20
□ Month [1-4] in 2008 => CK26
□ DK/REF => 19

19 **Do you know if it was before or after January 1, 2008?**
[If this is a **job-based plan** fill: PROBE: When we say "that coverage" we mean any coverage through [policyholder's] employer. So if [policyholder] switched plans offered by the employer, or even switched employers, we still consider this all the same coverage.]
[If this is a **directly-purchased plan** fill: PROBE: When we say "that coverage" we mean any coverage directly purchased by you or another policyholder. So if you/NAME switched plans but they were all directly-purchased, we still consider this all the same coverage.]
□ Before January 1, 2008 => Q22
□ On or after January 1, 2008 => CK26
□ DK/REF => Q22

20 **And has it been continuous since then?**
□ Yes => CK23
□ No => Q21
□ DK/REF => Q21

21 **In what month did this most recent spell of coverage start?**
□ Month [1-12] => CK23
□ Month [1-4] in 2008 => CK26
□ DK/REF => CK23

22 **What months in 2007 were you covered by that plan?**
□ Month [1-12] =>
□ DK/REF =>
=> CK23

CK23:

- if single-person household => CK26
- else => Q23

23 **Is anyone else within this household also covered by [if job-based or directly-purchased fill: your/policyholder's/else fill name of plan (e.g.: that Medicaid, Medicare, VA] plan?**
□ Yes => Q24
□ No => CK26
□ DK/REF => CK26

24 **Who?** (Who else is covered by that plan)? => Q25

25 **And [was NAME/were NAMES] covered during the same months in 2007 as you were?**
□ Yes, all were covered during same time=> CK26
□ No, DK, REF => Q25a

25a **What months in 2007 was NAME covered?** [repeat as needed] => CK26

CK26:

- If this is a job-based plan and NAME was covered for less than 12 months of 2007 by this plan => Q26
- else => Q27

26 **Ok now I'd like to ask you about other plans through either [your/NAME's own] or someone else's job. Were there any months in 2007 that [you were/NAME was] covered by a different job-sponsored health plan?**
□ Yes => Q15
□ No, DK, REF => Q27

27 **Other than [plan(s)], are you also covered by any other type of health plan or health coverage?**
PROBE: Do not include plans that cover only one type of care, such as dental or vision plans.
□ Yes => Q8
□ No => Q28
□ DK/REF => [Your best estimate is fine] => Q28

28 How about during 2007? (Other than [plan(s)]) were you covered by any (other) type of health plan or health coverage at any time during 2007?

PROBE: Do not include plans that cover only one type of care, such as dental or vision plans.

□ Yes => Q8
□ No => CK29a
□ DK/REF => CK29a

CK29a:

- If there are more household members on the roster who have not been asked about yet => CK29b
- else end

CK29b:

- If the next person on the roster was reported as having coverage (now or during 2007) during the course of any previous person's interview => Q29 for that person
- else => Q1 for that person

29 Now I'd like to ask you about [PERSON 2+]. Other than the [plan(s)] you reported earlier, does [PERSON 2+] have any other type of health plan or health coverage?

PROBE: Do not include plans that cover only one type of care, such as dental or vision plans.

□ Yes => Q8
□ No => Q30
□ DK/REF => Q30

30 How about during 2007? Other than the [plan(s)] you reported earlier, did [PERSON 2+] have any other type of health plan or health coverage at any time during 2007?

PROBE: Do not include plans that cover only one type of care, such as dental or vision plans.

□ Yes => Q8
□ No => go back to CK29a
□ DK/REF => go back to CK29a

APPENDIX B: METHODOLOGIES OF INDIVIDUAL COGNITIVE TESTS

Reports varied in their level of detail on methodology. All reported the number of respondents interviewed and the year conducted (see Table 2 below). Most noted that respondents varied across a range of demographic characteristics (age, sex, education, race/ethnicity and income). In the case of SLAITS, the focus was on immunizations for young children. Thus parental status and age of children were the key eligibility criteria, and all respondents were women.

Subjects were generally recruited through advertisements and word-of-mouth. However, in the case of the CPS 2000 test, which was part of a larger study to test questions on Medicaid, SCHIP and welfare, respondents were recruited from state and local government agencies that serve clients receiving those types of benefits. In the MEPS case, because the focus was on managed care questions, respondents were recruited by contacting employers and employees in order to generate a sample with a broad range of managed care plans (HMOs, PPOs, etc.). Not all reports mentioned payments and duration, but where noted respondents were paid $30-40 and interviews lasted somewhere between 45 minutes and an hour and a half.

Regarding mode, all tests were conducted face-to-face with two exceptions. In the BRFSS, which is a telephone-administered survey, some respondents were first interviewed over the telephone as if it were a production interview, and afterward subjects met face-to-face with an interviewer who administered retrospective cognitive probes. In the ACS, which is a multi-mode survey, cognitive testing was conducted in three different modes in an attempt to mimic production conditions: face-to-face self-administered (n=25); face-to-face CAPI (n =11); and CATI (n=4).

Table 2: Details on Cognitive Testing Reports

Study	**Year Conducted**	**Number of Respondents**
BRFSS	1995, 1998*	18 (1995); 16 (1998)
NHIS	2002	17
ACS	2004-2005	40
MEPS	1994-1995	58
CPS (2000)	2000	29
CPS (2006)	2004	27
SLAITS	2005	20
UMASS	1997	5
EXP1	2000	29
EXP2	2003	20
EXP3	2008	36

* These findings are based on a compilation of 10 different reports on cognitive tests conducted between 1992 and 1998 on a range of topics. Report 7 (1995) and Report 10 (1998) contained questions on health insurance coverage.

The design and testing of questions for mixed-mode surveys: Lessons learnt and considerations for the future

Michelle Gray
Questionnaire Design and Testing Hub, National Centre for Social Research

Summary

The use of mixed-mode in social surveys have become increasingly more popular over the last few decades and the norm in some countries, as survey managers seek to use data collection procedures that produce the best possible data within existing constraints of time and budget. Supporting this, and given the fact response rates are on the decline, offering the choice of mode of data collection has been shown to be an effective way of improving response rates (Shettle and Mooney, 1999). We also know that some respondents have preferences for the mode in which they receive the questionnaire and therefore mixed-mode designs allow for tailoring of mode for specific populations. If there is to be a move towards more mixed-mode designs, a concern for survey researchers is whether people who respond by one mode would have provided the same answers had they responded by another mode. When mixing modes, the goal is to achieve equivalence and one way of allowing this is to adhere to a Unimodal design, whereby questions are designed to provide the same stimulus in all survey modes to reduce differences in the way respondents respond to the survey questions in different modes.

This paper will discuss the design and testing of questions for mixed-mode surveys: some practical considerations and some lessons learned, with particular emphasis on designing and pre-testing unimodal questions. Specific examples from a recent project, which involved unimodal questionnaire design and testing to feed into a business survey toolkit, will be used to aid the discussion.

1. Introduction

Traditionally, in social surveys, data were either obtained by 1) an interviewer administering a questionnaire in a face-to-face setting or 2) through a paper self administered mail questionnaire. Two further modes joined the arena with the introduction of telephone surveys, in the 1990s, and data collection via the internet/web in more recent years. Social research organisations tended to focus their efforts on single mode studies, however declining response rates has moved us more towards opting for mixed-mode surveys: **using one or more mode to collect the same data from different people**[1].

The decision to mix modes can be made upfront (a concurrent design), with the aim of reducing coverage bias whilst still completing the survey at a reasonable cost, or it can be made further down the line (a sequential design). The former allows for respondents to be offered a choice, for example a paper mail survey with a web option, and it is assumed that offering a choice will reduce nonresponse as some people may express a preference for the

[1] 'Mixed-mode' is also used to refer to using one or more mode, in a survey, to collect **different** sorts of data from the same person, for example a self administered questionnaire component of a face-to-face interview.

mode in which they complete the survey. In establishment surveys, in particular, it is more common (compared with household surveys) to allow respondents to choose the data collection method they favour. The impact of mode on response, however, is not always easy to assess in concurrent mixed-mode designs, compared with the alternative approach to mode-mixing where additional mode(s) are added at a later stage: a sequential mixed-mode design. A sequential mixed-mode design involves one main mode with the addition of an alternative mode(s), used to follow up non-respondents. Various studies in recent years, which have used sequential mixed-mode strategies, have shown that switching to a second, and even third, mode is an effective means of improving response rates.

Regardless of *when* the different modes are introduced, a by-product of mixing modes is the reliance on different communication channels (i.e. aural vs. visual), when using alternative modes, which can introduce measurement error. The most basic cause of mode differences is the tendency to construct different questions, for example failing to offer a ‘don’t know’ response option over the telephone but providing respondents with this option in a self administered paper questionnaire. Dillman (2006) argues for a unimode construct which can reduce these differences but let us just remind ourselves of the underlying reasons for differences in mode, which are likely to occur even if the questions are designed to be identical across all modes involved.

Mode differences

There are a number of reasons why mode differences occur and these can broadly be categorised as those relating to 1) the presence, or absence, of the interviewer and 2) the stimulus delivery. Respondents are more likely to alter their behaviour in the presence of an interviewer; taking into account social norms and what they think are ‘culturally acceptable’ answers. Additionally, respondents may be less likely to divulge undesirable beliefs, opinions or behaviours to an interviewer. Interviewer administered surveys (face-to-face or telephone) are more likely to suffer from social desirability bias compared with self administered surveys (postal or web), however the effects are likely to be less for telephone surveys, as the respondent does not have to physically face the interviewer in the flesh.

The delivery of the stimulus (i.e. what respondents have to work with) differs according to modes. For example, in self administered surveys (postal and web), the stimulus are visual whereas in interviewer administered surveys (face-to-face and telephone), the stimulus are aural. The difference in stimulus delivery results in different processes through which the meaning of a question and the response options are comprehended. To achieve equivalence in mix mode surveys, these differences can be overcome to some extent in face-to-face interviewer administered surveys by using showcards, where response options can be presented to respondents visually, however telephone surveys tend to avoid showcards for practical reasons. Finally, if the interviewer has control over the delivery of the stimulus, they can read out the whole question and all response options, ensuring all respondents are exposed to the entire stimulus. If the respondent is in control however, as is the case with self administered surveys, he/she retains control of which parts of the question and response categories they read. An immediate result of lack of interviewer control in self administered surveys is higher item non-response rates.

2. Questionnaire design and testing for mixed-mode surveys

There is still a lot to be learned about those things that are inherent by-products of the different modes of data collection (i.e. the presence or absence of an interviewer), and at present there is little we can do about these. However we can focus our attention on ensuring that the stimulus respondents have to work with are consistent across modes, allowing comparisons to be made between the data collected via different modes.

Although there is a lack of theoretical or empirical knowledge available concerning how to design optimal questionnaires for mixed-mode data collection (de-Leeuw, 2005), there seem to be three schools of thought for achieving mode equivalence in questionnaire design for mixed-mode surveys. Table 1 below summarises these approaches, and some of the problems associated with each of them.

Table 1: Three schools of thought for mixed-mode questionnaire design.

School of thought	Problems
Mode Specific design: Design questions to optimise each mode seperately.	• Does not account for question format effects (e.g. check all that apply Vs yes/no response options) • Between mode comparisons are questionnable
Unified Mode Design or Uni-mode design: Design questions so that the same stimulus (question format) exists in each mode.	• Number of response alternatives is restricted • Unfolding questions must be avoided
Generalised mode design: Design questions to be different in different modes to achieve cognitive equivalence of the perceived stimulus, thereby resulting in equivalent answers across modes.	• Empirical evidence is needed to estimate what constitutes as the same perceived stumulus across different modes

Although there appears to be some guidance on designing questions for mixed-mode surveys, for example on Uni-mode design see Dillman (2006) Chapter 6, there is little guidance on pre-testing procedures for mixed-mode surveys.

Regardless of the intended mode of administration, it has tended to be common practice to pre-test questions, using cognitive interviewing methods, in face-to-face settings. Cognitive interviewing, as a method, strongly relies on non verbal cues from the respondent which may signify problems with the survey question and/or the response process. Of equal importance to the success of this method is both the rapport established between the cognitive interviewer and respondent and the interviewer’s encouragement throughout the cognitive interview. In fact, there are more communication resources available in face-to-face than in any other mode, and thus “communicative flexibility”, as Beatty & Schecter label it in their

1994 paper, makes conducting cognitive interviews in the face-to-face mode attractive (Lyberg and Kasprzyk, 1991), practical and logical.

Face-to-face cognitive interviews work well for questions designed to be interviewer administered, as the 'real' survey conditions can be replicated. When testing questions designed to be administered over the phone or self administered (either via a mail paper questionnaire or the web), it is important to consider whether face-to-face cognitive interviews are appropriate. Face-to-face cognitive interviews result in the questions being tested in a setting different from the final mode they will be administered in. A cognitive experience may therefore be created that is so far removed from the 'real' survey conditions. Beatty & Schecter (1994) showed that cognitive interviews in different modes are feasible and recommend "mode mimicking"; if matching the cognitive interview mode to the survey mode is felt to be important.

3. Developing uni-modal business survey tool-kits

The Department for Transport (DfT), in the UK, is actively involved in attempting to engage businesses on transport policy and related issues through involvement in a number of initiatives. Business surveys on transport are a rarity for the DfT and for Local Authorities (LAs) and Local Transport Authorities (LTAs). The DfT commissioned a programme of work to develop business survey toolkits which would allow for businesses to be effectively consulted by obtaining their attitudes to three different transport policy areas: 1) Transport; 2) Congestion and 3) The Environment. Working closely alongside Transport consultancy Faber Maunsell, the DFT designed three questionnaires which would eventually form part of these tool-kits, and would be used by LAs and LTAs in Britain to survey businesses. Uni-mode questions were designed, in line with the specification for the toolkits to be fit for use by all LAs and LTAs, regardless of the resources available to them. The DfT knew, for example, that some LAs and LTAs may be limited to conducting only mail surveys.

NatCen's Questionnaire Design and Testing (QDT) Hub was invited to be involved in the programme of work to lead on pre-testing the three questionnaires. The cognitive testing phase was used to explore data collection mode, both in terms of the mode in which the questions performed 'best' and respondents' preferred mode or modes. Additionally cognitive interviewing methods were used to assess:

- The acceptability of the questions;
- The degree to which they were understood as intended;
- Comprehension of key terms within the questions;
- Ability and willingness of respondents to provide responses; and,
- Perceived knowledge needed to answer the questions, which could also link to whether the correct person, also known as the Person Most Knowledgeable (PMK), had been identified through the screening process.

Methodology

The questions, and instructions, within each policy questionnaire were designed to be unimodal, i.e. they could be administered in exactly the same format, in any of the four main modes of data collection: 1) Interviewer Administered face-to-face survey; 2) Interviewer administered telephone survey[2]; 3) Mail self administered survey; and 4) Web self administered survey. There were two research teams from the two organisations involved in the cognitive testing (NatCen and Faber Mansuell) and a limited timescale. Due to this and the fact that there were three instruments to test, a decision was made to test subsets of questions (from each of the three policy questionnaires) in two modes: a face-to face Interviewer Administered (IA) format and a paper Self Administered (SA) format.

The testing was split across two fieldwork phases in 2008: the three IA test questionnaires were tested between August and September and the three SA test questionnaires were tested between September and October. Ninety cognitive interviews were conducted in total; the breakdown of these is shown is table 2 below.

Table 2: Total number of conitive interviews split over two phases of testing

	Transport		**Congestion**		**Environment**	
Mode of testing	IA	SA	IA	SA	IA	SA
No. of interviews	15	15	15	15	15	15

As we were only able to test the questionnaires in two of the four modes, we made assumptions that the interviewer administered (IA) questionnaires could act as a proxy for the telephone mode and the self administered (SA) paper questionnaires would act as a proxy for the web mode. Although we were relatively confident that we were replicating telephone survey administration conditions, the interviewer was in the same room as the respondent (as opposed to on the end of a telephone line), so an underlying problem was the effect the interviewer's presence may have had in these cognitive interviews, in terms of rapport, body language and social desirability.

The assumption that the SA paper questionnaire could act as a good proxy for the web is more problematic as the question format, the visual design and the physical act of filling out each instrument will be different. However, on the basis that the respondent has complete control over whether and/or how each question is read and comprehended, we expected that responses to a web questionnaire would closely resemble those to a mail questionnaire. We therefore felt that any problems the testing highlighted with the mail questionnaire could be applied to the context of a web questionnaire.

Sampling and recruitment

In order to identify businesses with a range of particular characteristics to take part in the cognitive interviews, businesses were sampled from the Experian database[3]. A quota sample was designed to ensure that a range of different businesses were included in the two phases of the pre-testing, taking into account level of business (enterprises and establishments), type of business activity and number of employees at the site. Businesses

[2] Questions were designed so that there were no showcards. Response options would be read out to respondents in both of the interviewer administered modes.

were contacted by telephone, the Person Most Knowledgeable (PMK) was identified and invited to take part. Upon agreement, respondents were sent a letter confirming the interview details. The cognitive interviews took place at the business' sites, across ten geographical areas of the country in both urban and rural locations.

Conduct of interviews and analysis

The interviews were conducted by a mix of both NatCen's and Faber Maunsell's experienced cognitive field interviewers. The same interviewers worked on the two fieldwork phases and were briefed by the research team at the start of each phase. Each interview lasted around an hour and interviews were audio digitally recorded with respondents' consent. Interviewers probed concurrently in both phases (modes) and confidentiality and anonymity were assured throughout. All interviewers made detailed notes from the recordings and the notes were analysed using a content analysis approach based on Framework, an analytic tool developed by the Qualitative Research Unit at NatCen. A matrix was set up, which listed the areas under investigation (areas probed on and explored by interviewers) across the page and cases (businesses) down the page. The matrix also included a summary of the business' characteristics; such as whether the business was an enterprise or an establishment, the nature of the business' activity and its size. Thus data could be read horizontally, as a complete case record for an individual/business, or vertically, by question or area under investigation, looking across all cases. Once the matrix was completed, the data were reviewed. In reviewing the matrix the full range of problems with each question were explored and appropriate recommendations for improving the questions made, which included recommendations regarding mode.

4. Findings

The cognitive interviews highlighted many question-specific problems that were found in both modes. These included issues such as lack of knowledge and awareness, confusion over positively and negatively worded response options in the same list, confusing question wording and unrealistic, or ignored, reference periods. There were then findings which were directly related to mode. We will now describe one of these in a little more detail.

The forced choice response approach

Questions with a long list of response options were designed to be formatted in the same way, in each mode, with the use of forced choice response options. In the IA mode, the interviewer was instructed to read out the response options pausing after each one for a Yes/No response, whilst in the SA mode respondents were forced to tick a Yes/No box for each response option. The alternative means of guiding respondents through a list of multi-code response options is the 'check all that apply' approach. However we know that with self administered questionnaires, respondents tend to be more likely to choose options towards the top of the list, sometimes failing to consider those further down – know as the 'primacy effect'. On the flip side, with a check all that apply approach, in interviewer administered telephone questionnaires where there is an absence of a showcard, respondents have a

[3] The Experian Database is a retail database which provides information on over 450,000 businesses in the UK and Ireland.

tendency of opting for response options that are read out to them last, i.e. those towards the bottom of the list. This is known as the ‘receny effect’.

There were three main issues which the testing highlighted in relation to the forced choice approach. Firstly, and particularly for questions with long lists of response options, IA respondents often felt that there were far too many response options to consider, reporting difficulty retaining the question stem as the interviewer read further down the list. Secondly, IA respondents volunteered ‘don’t know’ responses whilst SA respondents had no place to record ‘don’t know’ and expressed a desire for this option during probing[4]. Finally, SA respondents did not always follow the instruction to tick one box on each line and failing this, either:

- Only ticked ‘Yes’ to those which applied and omitted the No boxes all together;
- Only ticked ‘No’ to those options that did not apply and omitted the ‘Yes’ boxes all together;
- Ticked the ‘No’ boxes at all options and then realised that they should have ticked the ‘none of these’ option at the end; or they,
- Ticked the 'none of these' option where it was offered, despite having ticked 'Yes' at other response options.

Interestingly, for one question, an option which came second in a long list was chosen far more often by SA respondents than IA respondents, highlighting a possibly primacy effect, however as our sample was very small we can not be sure about this.

There were other mode related problems that the testing highlighted. Ranking tasks, for example, were difficult in the IA mode. IA respondents found it impossible to retain all of the information in their heads, claiming they would only have been able to complete the task with a visual list of the items for ranking in front of them. There were problems with the routing in places in the SA versions of the questionnaires, where respondents got lost and confused or incorrectly skipped questions as they proceeded through the questionnaire. As mentioned above, the long list of response options were found to be difficult for IA respondents. Interviewers found them tedious to read out and respondents got lost and missed the point of the question.

Mode preference

It is common practice in the UK to survey businesses via the telephone, as it is deemed the most cost effective method and telephone numbers are usually available for sampling purposes (this is unfortunately not the case for most of our household surveys). Businesses are also surveyed by mail and web and occasionally in face-to-face interviews. Respondents in our study were asked if they had a preference for the mode in which they would complete the questionnaires that were tested, and interestingly there was a greatest preference shown for a face-to-face or a web survey. Respondents expressed a desire for the web as it would allow them flexibility for when they could complete the questionnaire and the complicated nature of the questionnaires seemed to be a main reason why a face-to-face interview was a

[4] Despite there being no instruction to do so, nor a code for such a response, interviewers recorded don’t knows on the test questionnaires in the IA mode.

preferred choice for some. Contrary to common practice in the UK, respondents in our cognitive interview sample displayed an overall resistance to a telephone survey. The reasons for this included:

- Difficulty answering such complicated questions without being able to see the questions;
- Being on the phone to a survey interviewer would block up the businesses' phone line, stopping potential customers from getting through (particularly so for smaller businesses);
- Gatekeepers influencing the likelihood of survey interviewers being able to get through to the correct person to speak to, i.e. interview;
- Lengthy nature of the questions and the questionnaires would make them difficult to answer over the phone: resulting in loss of interest and lack of enthusiasm to give questions proper consideration.

5. Lessons learned

As with any first attempt, there are a number of key lessons that we feel we learned which can be applied to designing and testing questions for mixed-mode surveys. First, and probably foremost, it became evident even part way through the project that plenty of the questions were, by design, inappropriate for some of the modes of data collection. It should not have been assumed, for example, that questions with long lists of response options (in some cases up to 15), could be asked over the telephone, with the absence of a visual list. Similarly, questions which involve complicated and thought provoking tasks, such as ranking items, should be kept well away from telephone surveys! We feel that it might not always be necessary to format forced choice response options, especially for questions that have a simple and short list of options, however in taking this approach we had in mind Dillman's principle of 'keeping everything the same'.

Next time we would seriously consider testing questions which have been designed for a telephone survey over the phone, and those designed for a web survey on a computer screen. We feel that 'mode-mimicking' in cognitive interviews would be appropriate in these circumstances for the following reasons. From the point of view of the telephone survey questions, it is possible that the presence of an interviewer in our study may well have influenced the way respondents went about answering. Visual cues and the building of rapport, for example, are just two of the ways the supposedly 'telephone survey' conditions could have been affected. In future, we might conduct cognitive interviews over the telephone with the interviewer in a different room to the respondent, at the same location. This would provide an opportunity for replication of 'real' telephone survey conditions whilst allowing the interviewer to come back into the same room as the respondent to carry out a retrospective probing session, following the administration of the test questionnaire. An additional, and similar, approach could be to have an interviewer phone the respondent and administer the questions over the phone just before the cognitive interviewer arrives at their home to conduct the retrospective probing exercise.

Mode mimicking for web survey questions would be beneficial as routing issues found in paper questionnaires could be avoided. Although it would be advantageous to have the interviewer present - to observe, encourage ‘think-aloud’ and to follow up with probing, having the respondent fill out the questionnaire on the screen would most certainly pick up on visual design and layout issues and potential problems. Since the project mentioned in this paper, we have tested a web survey questionnaire and a Computer Assisted Self Administered (CASI) questionnaire in this way and it proved to be highly successful and useful, whilst remaining relatively cheap and easy to set up.

Finally, we have learned that the more modes that are thrown into the mix, the more likely it is that mode effects will occur. It is hard enough optimising the design of just two modes, whilst still attempting to retain equivalence. When the design is for two or more modes, you have to make even more compromises. This highlights the importance of fully considering the mode at the outset and if a mixed-mode design is necessary, and the data needs are complex, whether it might be more appropriate to only ask certain questions in certain modes. What you would lose out on in quantity, you would potentially gain in quality.

6. Next steps

This paper has highlighted just some of the considerations that need to be taken into account when designing and testing questionnaires for use on mixed-mode surveys. This area of practice is still relatively new however and further research is needed to provide more guidance and to develop our understanding. As a starting point I would suggest the following:

- Further experimental research, supported by cognitive research, to explore how respondents process survey questions under the condition of different modes; and,
- Research and experiments to explore whether mode mimicking (on the telephone for example) is the most effective means for testing questions for mixed-mode surveys.

References

Beatty, P. and Schechter, S. (1994). ‘*An examination of mode effects in cognitive laboratory research*’. 1994 Proceedings of the Section on Survey Research Methods, American Statistical Association, 1275 - 1280.

Shettle, C, and G Mooney. (1999). ‘*Monetary Incentives in U.S. Government Surveys*’. Journal of Official Statistics 15:231-250

Don A. Dillman. (2006). *Mail and Internet surveys*, New York: Wiley

Lyberg, L., and Kasprzyk, D. (1991). "*Data Collection Methods and Measurement Error: An Overview*." In Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S., eds., Measurement Error In Surveys. New York: Wiley

de-Leeuw, E. (2005) ‘*To Mix or Not to Mix Data Collection Modes in Surveys*’, Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 233–255

What do respondents want and expect from electronic self-completion surveys? A discussion.

Lucy Tinkler

The use of electronic modes as a means of collecting self-completion survey data has become increasingly popular. Surveys that use electronic modes of data collection are most often thought of as those which are completed directly on the internet, and those that are administered via email. The Office for National Statistics (ONS) has recently completed a literature review into research investigating respondent perspectives on electronic modes of data collection, the focus of which was respondent preferences. It was found that there has been little research carried out into whether respondents prefer to complete questionnaires using electronic modes, what respondents understand an electronic 'questionnaire' to be, and, the level of burden incurred compared with paper self completion. ONS proposes to investigate this topic further focusing on business surveys. In this session Lucy Tinkler will briefly outline the research that has been carried out into this topic, this will be followed by a discussion exploring the type of research and methods ONS data collection methodology branch could employ to carry out this research.

Presenting 'don't know' in web surveys

Rachel Vis-Visschers[1]

Summary: In 2007 the Questionnaire laboratory of Statistics Netherlands conducted an experiment in which four versions of presenting “don’t know” in a web questionnaire were compared. In this paper we discuss the experiment and its results. The experiment showed that presenting “don’t know” always on screen for questions to which it is a relevant answer is to be recommended.

Keywords: Don’t know, Web survey, Mixed-mode, Questionnaire laboratory.

1. Introduction

In 2007 the Questionnaire laboratory of Statistics Netherlands conducted two laboratory tests on how to present “don’t know” and help texts in web questionnaires. The first of these tests was discussed and presented at the Quest Workshop of 2007 in Ottawa (Vis-Visschers, 2007b). The present paper will deal with the results of the second test, and we will focus on the results of presenting the answer category “don’t know” in a web questionnaire. The complete set of results of both tests is presented in a Dutch report (Vis-Visschers et al., 2008). Firstly we give a short background of the laboratory test and the importance of presenting “don’t know” in surveys, in relation to developments concerning mixed-mode research at Statistics Netherlands. Then in section 2 we discuss the method and the set up of the laboratory test. In section 3 we present the results. In section 4 we draw our conclusions and give some recommendations. We conclude this paper with plans for future research.

1.1 Background of the laboratory test

In 2007 a project was started at Statistics Netherlands (StatNeth) to gather information and experiences on the best way to conduct mixed-mode surveys, since StatNeth envisions executing most surveys from 2010 onward in a mixed-mode context for efficiency reasons. Most social surveys will be executed in a sequential mixed-mode setting, in which the respondents are first asked to participate in a web questionnaire, after which the non-respondents are either approached for a telephone interview or a personal interview. This is considered to be a cost efficient way of data collection (De Leeuw, 2005).

Yet there are also drawbacks to this way of collecting data; it is expected that survey errors occur, for instance discontinuities in the time series of a statistic. These errors can have several causes, but in this paper we will focus on the ‘mode effects’, i.e. errors that occur because different modes of data collection are used, for example differences due to oral or visual presentation of the questions. Three types of non-sampling error are distinguished[2] (Groves, 1989; Roberts, 2007). These are “the three main types of ‘mode effect’ researchers interested in mixing modes should be aware of”, according to Robert (2007).

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

[1] Based on a Dutch report by Rachel Vis-Visschers, Judit Arends-Tóth, Deirdre Giesen & Vivian Meertens (2007a).

[2] The total survey error consists of both non-sampling and sampling errors. The latter arise because estimates are based on a sample rather than a full census of the population.

1. **Coverage errors**, e.g. differences occur because telephone interviewing often only includes households with a known landline (fixed) telephone connection.
2. **Nonresponse errors**, e.g. differences occur because older people are less likely to participate in an internet survey.
3. **Measurement errors**, e.g. question interpretation may vary in self-administered modes in comparison with interviewer-assisted modes in which interviewers can control uniformity in question interpretation. It is also possible that the presence of an interviewer may cause the respondents to edit their response since they are reluctant to disclose sensitive or embarrassing behaviour or opinions.

To prevent, or at least to minimize the impact of, mode effects it is necessary to gather relevant information. For this purpose the Questionnaire laboratory (Q-lab) formulated a research proposal (Vis-Visschers, 2007a). In this proposal several ways to accumulate knowledge about mode effects and experiences with mixed-mode research are described. One way to gather information is an extensive literature study; the results of this study are described in Ariel et al. (2008). A second way is a laboratory test (based on what we learned form the literature review) on how to present help texts and the answer category "don't know" (DK) in a web survey, hence this paper.

1.2 The importance of the way "don't know" is presented.

When respondents answer survey questions, theoretically speaking, they follow the four phases of the response process (Tourangeau & Rasinski, 1988; Snijkers, 2002): 1) Interpretation and comprehension; 2) Information retrieval; 3) Judgement; and 4) Report. Still, we also know that many respondents do not (always) perform the four phases carefully and completely. This behaviour is known as 'satisficing' (Krosnick, 1991). Satisficing manifests itself more often in self-administered modes than in interviewer-administered modes. An interviewer can assist a respondent in giving a correct answer, and can motivate the respondent to focus on the task. One way in which the influence of the interviewer can be seen is in the way a respondent deals with the answer category "don't know". An interviewer can persuade a respondent to change an initial DK-answer; this is not possible in a self-administered questionnaire.

The way in which DK is presented in a web questionnaire can effect the (quality of the) data (Presser & Schuman, 1989; Gilljam & Granberg, 1993; DeRouvray & Couper, 2002). If DK is always presented as an answer category it will be chosen significantly more often (Van den Brakel et al., 2006). If DK is never offered, it is possible that a respondent is forced to give an answer that doesn't fit his opinion (i.e. "pseudo opinions", Bishop et al., 1980) or even that a respondent will quit without finishing the questionnaire (i.e. drop-outs). The idea is that a balance must be achieved. For this reason the Q-lab executed a laboratory test in which four versions of presenting DK were compared.

2. Method

2.1 Four versions of presenting "don't know"

The following four versions of presenting "don't know" in a web questionnaire were compared in the laboratory test (in the appendix we included screenshots of the web questionnaire):

Version 1 “Always” (Appendix figure 1). The response option is presented on screen for each question. It appears directly under the other response options. From previous experiments we know that this increases the chance that the option will be chosen. For this laboratory test we expect that the amount of DK’s will be highest in this version.

Version 2 “Never” (Appendix figure 2). In this version “don’t know” is never an option. The respondents are forced to select an answer, even if they do not know, in order to continue with the questionnaire. This will possibly result in non-attitudes or pseudo attitudes. We expect that some respondents could get irritated. In a lab setting it will be highly unlikely, but in reality it is plausible that a respondent will quit before finishing the questionnaire.

Version 3 “Hidden” (Appendix figure 2 and 3). If the respondent tries to skip the question a warning appears that all questions have to be answered, and the DK-option is presented under the other answer categories. The procedure is explained at the beginning of the questionnaire. This option requires a deliberate action of the respondent. It resembles the procedure in an interviewer administered survey: the respondent first has to indicate that he or she[3] does not know an answer after which the interviewer tries to motivate the respondent to still choose between the response options and only to accept DK as a last alternative. In this test we want to find out whether a respondent is aware of this hidden option; how do respondents react to the warning that they cannot skip a question; do they see the DK-option appear after they have tried to skip a question?

Version 4 “Button” (Appendix figure 4). DK is presented less visually prominent. At the bottom left side of every screen a DK-button is presented. The idea behind this version is that it is always possible to answer “don’t know”, as it is in an interviewer administrated survey, but choosing it is made a bit less inviting because the option is put away from the other response categories. A similar option was tested in DeRouvray & Couper (2002), in which DK was presented together with the other response options, but in a lighter colour. In this test we want to find out whether this version is indeed “less visually prominent” and whether it really does resemble an interviewer administrated survey.

2.2 Test respondents

To get an as accurate as possible picture of problems that can arise during the interview and the questionnaire, we strived to recruit a diverse group of test respondents based on gender, age, education, computer and web survey experience (Table 1.).

For this laboratory test people who had participated in previous tests were addressed. A few days after the letters were posted they were contacted by telephone to ask whether they wanted to participate and/or they knew other people who were willing to participate. At the end of the telephone call an appointment for a test interview was made. Shortly after the call a letter was sent to confirm the appointment.

Eventually, 36 persons completed the test interview. The group of test-respondents consisted 1) partially of people who had participated in previous tests themselves; 2) partially of people recruited via them and 3) partially of acquaintances of the researchers. Due to this ‘snowball’-method for recruitment it is not possible to give a response percentage.

[3] Henceforward we mean “he or she” when we refer to the test respondent as “he”.

Table 1. Some background information on the 36 test respondents.

Gender		
	Men	11
	Women	25
Age		
	17 – 35	11
	36 – 50	13
	51 - 68	12
Education		
	Low	12
	Middle	14
	High	10
Experience with computers		
	Little	10
	Much	26
Experience with web surveys		
	Little	20
	Much	16

2.3 Test questionnaire

The questionnaire was built in BlaiseIS and consisted of several blocks that each dealt with a separate subject: Background questions (mainly factual questions); Health and food (mainly knowledge questions); Dutch politics (mainly knowledge questions); Family and role divisions (mainly attitude questions). On the last page of the questionnaire all versions of "don't know" were presented and explained and the respondents were asked whether they preferred one version above another.

For this experiment we set out to provoke the use of DK, and thus different kinds of questions (factual, knowledge and attitude) and especially questions that would be difficult to answer were selected. We deliberately included:

- questions with difficult words ("Do you know your Quetelet-index?")
- questions with controversial propositions ("In a family the man should be the provider and the woman should take care of the household and the children.")
- questions about events in the past and requiring special interest ("At the end of June 2006 the Dutch government resigned. Which incident caused the cabinet to fall according to you?")
- questions for which records have to be consulted and calculations have to made ("How much money did you spend on medical costs this year?")

2.4 Experimental design

The four versions of presenting DK were built into four versions of the test questionnaire. The test respondents were randomly assigned to one of the four versions. Table 2 shows the distribution of the respondents and the respondents' characteristics over the four experimental conditions.

Table 2. Distribution of 36 respondents over the four experimental conditions.

	Total	**Gender**		**Age group**			**Education**		
	(36)	**Male**	**Female**	**17-35**	**36-50**	**51-68**	**Low**	**Mean**	**High**
Version 1 "Always"	6	1	5	3	1	2	1	3	2
Version 2 "Never"	6	2	4	2	2	2	2	3	1
Version 3 "Hidden"	13	3	10	3	6	4	3	4	6
Version 4 "Button"	11	5	6	3	4	4	6	4	1

Previous research has shown that either explicitly showing or never showing DK is not the best option (e.g. Krosnick et al., 2002; Van den Brakel et al., 2006). Nevertheless these versions are still tested in this experiment in order to set reference points for the results for the other two versions.

2.5 The laboratory test

In a laboratory setting people behave differently than in real life. It is impossible to pretend you are at home, if there is an interviewer observing all your moves and scrutinizing all your remarks. Moreover, there often will be a camera directly aimed at the respondent. Test respondents are very cooperative and motivated to "do it right". They know they are participating in an experiment and know that their behaviour is judged, thus they want to present themselves as favourably as possible. It is difficult to copy a real life situation, yet in a laboratory setting one can observe other aspects. For instance, one can try to uncover underlying reasons for choosing or not choosing "don't know". Since respondents know that we want to know everything about them, it is not awkward to probe and ask many follow-up questions.

For the test interviews the respondents were invited to the office of StatNeth in Heerlen. All interviews were recorded on tape, and were conducted by several colleagues of the Q-lab. In order to standardize the way in which the interviews are conducted, a fixed protocol was followed:

- **Filling out the questionnaire.** First the interviewer shortly explains the importance of pre-testing questionnaires and the task of the test respondent. Next the test respondent fills out the questionnaire, which is already set out on the computer. The respondent is asked to think aloud, and is motivated to mention anything that seems noteworthy. Otherwise he has to act as if the interviewer is not present. In case the respondent wants to ask something, the interviewer is instructed to say: "What would you do if you were alone at home?". The interviewer observes the respondent's behaviour using a paper scoring form.
- **Cognitive interview.** When the respondent finishes the questionnaire, the cognitive part of the interview starts. By retracing the respondent's steps through the questionnaire and asking meta-questions[4] the interviewer tries to unveil the respondent's thoughts and decisions while he was answering the survey questions.
- **Last question about preference.** After the in-depth interview, the final page of the questionnaire, in which all versions of presenting "don't know" are shown, is opened. The interviewer explains the aim of the test and the different versions of "don't know" and then asks the respondent whether he prefers one of the versions and why. The interview is concluded and the respondent receives his participation fee.

[4] For instance: "You changed an initial answer here, can you explain why you did this?" or "Could you rephrase this question using your own words?".

The results of the individual interviews were gathered in one spread sheet for the analyses. Finally the results are written down in a report (Vis-Visschers et al., 2008).

3. Results

3.1 In general

The design of the laboratory test was successful. The test respondents often chose to answer "don't know", even though it was a test situation and they were motivated to be 'good respondents'.

It turns out that people deal differently with "don't know". Some want to avoid it, because they see that it is not a useful answer for the researchers: "*That's not useful for Statistics Netherlands*." Others avoid it, because they hesitate to admit they do not know something: "*It is outrageous that I don't know this. My husband is a nurse, so I should know these medical terms*." Again others prefer answering "don't know", as an alternative to randomly guessing an answer. And finally there are respondents who answer "don't know" because it is just the most honest answer.

In Table 3 we have shown some details on the respondents' use of "don't know". Version 2 "Never" is not shown since it was not possible for these six respondents to answer "don't know". As can be seen, it was possible for 30 respondents to answer "don't know". Of these respondents, 16 persons actually used the response option 67 times. That is a mean use of 4.2 times per respondent who ever used DK.

Table 3. Details on the use of "don't know".

	"Don't know" versions			
	Version 1 "Always" (n=6)	**Version 3 "Hidden" (n=13)**	**Version 4 "Button" (n=11)**	**Total (n=30)**
Respondents that use DK	6	7	3	16
Percentage that use DK	100%	54%	27%	53%
Total number of DK's	28	25	14	67
Mean use of DK (if used)	4,7	3,6	4,7	4,2

In table 4 we have divided the DK-answers over the different blocks of the questionnaire or the different kinds of questions. The first block, Background, consists mostly of factual questions. Here we see that none of the respondents answered "don't know". In the blocks Health and food and Dutch politics we find the most knowledge questions and the most DK-answers. The block Family and role divisions contains mainly attitude questions. Table 4 shows that only a few times DK was answered in this block.

Table 4. The number of "don't know" answers per questionnaire block.

	"Don't know" versions			
	Version 1 "Always" (n=6)	**Version 3 "Hidden" (n=13)**	**Version 4 "Button" (n=11)**	**Total (n=30)**
Background questions	0	0	0	0
Health and food	6	10	2	18
Dutch politics	20	14	11	45
Family and role divisions	2	1	1	4

In the following sections we will discuss the results per version of DK.

3.2 The use of “don’t know” in Version 1 “Always”

All respondents in this version used the response option DK. Moreover in this version, DK was used most often. This confirms our expectation that presenting DK always on screen increases the chance that DK is chosen.

The amount of times DK is used varies between one (female, 22, high level education) and eight times (female, 64, low level education). The mean amount of times DK is used is 4.7.

When asked how the respondents appreciate their version of presenting DK, all respondent answer that they liked the fact that they knew DK was possible.

3.3 The use of “don’t know” in Version 2 “Never”

In this version is was not possible to answer DK or even to skip a question. Theoretically it was possible that drop-outs could occur, but since this was a laboratory setting this didn’t happen. Yet some respondents spontaneously mentioned that it was annoying that DK was missing.

Verbally the respondents confirmed that they randomly chose an answer, because there was no real fitting response option in the list: “*I don’t know anything about politics. I’ll just pick an answer.*” Or “*I have to pick something… I don’t know anything about this. I’ll just pick the middle option.*”

Another interesting finding here is that, compared to the other versions, these respondents consulted the help texts the most. It is possible that the absence of DK motivates the respondents to search for more information to be able to better choose between the response options.

3.4 The use of “don’t know” in Version 3 “Hidden”

Not all respondents in this version used DK. Seven of 13 respondents chose “don’t know” a total of 25 times. Even though at the beginning of the questionnaire it is explained how DK could be chosen, it appeared that not all respondents remember this at relevant moments. Some find the response option by accident, because they try to skip the question, one respondent remarked: “*Oh yeah, don’t know is now possible, I have seen it in the introduction, but had forgotten about it.*”. Another accidentally skips a question, but did not see that the response option DK had appeared. She thinks she really has skipped the question and browses back and forth a couple of times. Eventually she chooses an answer (not DK). Another respondent had remembered that she could skip a question (though not that DK would appear), yet this was not an option for her: “*No, I would rather make a guess than skip a question.*”

The respondents in this version have mixed feelings about this way of presenting DK. Five of them used the option more often and appreciated the possibility. Yet the other two only found the option by accident, and are less appreciative: “*It only hampers the flow.*” Or “*It is not clear, you think: there is that question again?*”

Actually you can say that the respondents in this version can be divided in two groups: 1) those who don’t know there is a DK-option, and who only find it by accident or not at all; and 2) those who know there is a DK-option, and use it as if the option was not ‘hidden’.

3.5 The use of “don’t know” in Version 4 “Button”

The DK-button was meant to be less visually present, though it was not meant to be invisible. Still only one respondent found the button early in the questionnaire. Two others found the button when they were well on their way in the questionnaire and the other eight never found it. The reason why they never found it was that the button was too far out of focus. The respondents had their eyes focussed on the question on the upper half of the computer screen, since there was never any relevant information at the bottom of the screen they were not enticed to look there.

Just as is the case for version 3 ′Hidden′, we can divide the respondents into two groups: 1) those who know that DK is an option and 2) those who don't know.

3.6 Preference for a specific version of presenting "don't know"?

At the end of each test interview the respondents were asked whether they preferred one of the versions of presenting DK. In Table 5 we have presented the answers to this question. Most respondents (20) prefer version1 "Always", followed by 14 respondents who preferred version 3 "Hidden". The obvious losers are both versions 2 and 4. There does not seem to be a relation between the version the respondent had in the questionnaire and his preference.

Table 5. The respondents′ preference for the way to present DK.

	Respondents in the versions:				
	Version 1 "Always" (n=6)	**Version 2 "Never" (n=6)**	**Version 3 "Hidden" (n=13)**	**Version 4 "Button" (n=11)**	**Total (n=36)**
Preference for Version 1 "Always"	3	2	6	9	20
Preference for Version 2 "Never"	0	0	1	0	1
Preference for Version 3 "Hidden"	2	4	6	2	14
Preference for Version 4 "Button"	1	0	0	0	1

4. Discussion

4.1 In general

During the test it became evident that the respondents' character also influences the way they deal with answering "don't know". There are people who consciously answer DK, because it is:

1. the most honest answer, or
2. better than taking a guess.

There are also people who avoid answering DK, because

1. they do not want to seem uninformed, or
2. they think the answer is not useful for the researcher.

Whichever way DK is presented in a questionnaire will influence the data. In a real data collection situation the respondent's point of view towards DK can never be retrieved. This could imply that it is fairly impossible to gather comparable results.

Since it cannot be avoided, researchers have to be aware of it and have to make a decision whether to do something about it: e.g. inserting a selection question (Bishop et al., 1980); or probing after a DK-answer; or reassuring a respondent at the beginning of a questionnaire that DK is a possible and acceptable answer; or never presenting DK and forcing all respondents to give an answer. Still, all of these methods will have their own advantages and disadvantages.

4.2 In relation to different kinds of questions

The laboratory test showed that "don't know" is answered most to questions for which certain knowledge is presumed (that is, the blocks Dutch politics and Health and food). For attitude questions (e.g. the block Family and role division) DK is less often chosen. In case of factual questions (in the block Background questions) no test respondent had the wish to answer DK. There is a certain logic to this: it is possible that one does not have the knowledge to answer a specific knowledge question, while for attitudes or opinions it is often the case that one has never thought about it until that moment. At that moment one does not know an answer, but an opinion is formed on the spot.

4.3 What does a respondent's answer mean?

Not all respondents use DK, because not all respondents or have the urge to use it. Or, as is especially true for DK-version 3 and 4, the respondents do not find the response option. This means that the respondents can be divided into two groups: 1) the respondents who know there is a DK-option and either use it or not; and 2) the respondents who do not know there is a DK-option. This results in the disturbing fact that a researcher does not know for sure that the response to a survey question corresponds to the respondent's opinion. That is to say, there are several possible explanations if a respondent did not choose "don't know":

- Either, the respondent really does know the answer.
- Or, the respondent does know there is a DK-option, but doesn't want to answer DK.
- Or, the respondent does not know there is a DK-option and is forced to choose another response.
- Or, the respondent is not participating seriously and is just answering at random.

4.4 To conclude

The results show that the way in which DK is presented influences the amount of times DK is answered. Our expectation that DK would be used by most respondents in Version 1 "Always" has been confirmed. In the other two versions, "Hidden" and "Button", less respondents used DK. Still when you look at the mean amount of times a respondent uses DK, there is little difference between the versions. There seems to be little difference between the 'discouraging to use DK'-effect of either version 3 and 4.

In table 6 we have listed advantages and disadvantages of all versions investigated in this test. From this we cannot definitely conclude that there is one best way of presenting "don't know" in a web questionnaire. A researcher has to consider all advantages and disadvantages and decide for his research which risks he is willing to take.

When presenting DK in a web survey a researcher should also take into account which kinds of questions are posed. We recommend that DK is always presented with the other response options if there is a reasonable possibility that respondents do not know the answer to the question, i.e. for questions to which certain knowledge that not all respondents have is presumed. For attitude questions or factual questions it does not seem crucial.

Table 6. Advantages and disadvantages of all versions of presenting "don't know".

	Advantage	**Disadvantage**
Version 1 "Always"	• Respondents can always answer DK if they want to.	• Danger of DK being used as an escape (satisficing).
Version 2 "Never"	• No missing values in the data.	• It is unknown whether the answer corresponds with the respondent's actual opinion. • Danger of drop-out.
Version 3 "Hidden"	• Satisficing is discouraged.	• Not all respondents know DK is an option, notwithstanding the announcement. • A lot of mouse clicks in order to answer DK. • It is unknown whether the answer corresponds with the respondent's actual opinion.
Version 4 "Button"	• Satisficing is discouraged.	• Not all respondents know DK is an option, because the button is too far out of sight. • It is unknown whether the answer corresponds with the respondent's actual opinion.

5. Future plans for experiments

In February 2009 the Q-lab executed a mixed-mode experiment together with CentERdata of the University of Tilburg. For this experiment we used the questionnaire of the Dutch Consumer Sentiment Survey. We wanted to compare the following variables in CATI-mode versus Web-mode:

1. Always presenting DK vs. never presenting DK.
2. Probe after DK-answer vs. no probe.
3. "Unfolded" question vs. 5-point scale question.

Ad. 1. In this experiment it was possible to test with a substantial amount of persons the effect of always presenting DK versus never presenting DK, and to see whether it is possible to never present DK in a telephone survey.

Ad.2. In order to make a web survey more similar to a telephone interview it is possible to probe "It is very important for us if you try to answer this question." after DK is answered. We want to see whether and how this works in a web survey.

Ad. 3. This survey has been executed by StatNeth for many years. Before the 1980s it was a PAPI survey, after that a CATI survey and now we want to investigate whether it is possible to execute this survey by web. For this we need to change a specific kind of question of this survey: the "unfolded question" (Table 7). It seems weird to ask an "unfolded" question in a web survey, thus we will change this into a 5-point scale question (Table 7). To see the effect of the 5-point scale we included it in both the web survey and the telephone interview.

The results of this experiment will probably be available by the end of 2009.

Table 7. Unfolded question vs. 5-point scale.

Unfolded question

Question A.

"What do you think will happen to the unemployment rate in the Netherlands the next 12 months? Do you think it will increase, decrease or stay the same?

1. Increase → got to Question A1
2. Decrease → got to Question A2
3. Stay the same

Question A1.

"Will it increase a little or a lot?

1. Increase a little.
2. Increase a lot.

Question A2.

"Will it decrease a little or a lot?

1. Decrease a little.
2. Decrease a lot.

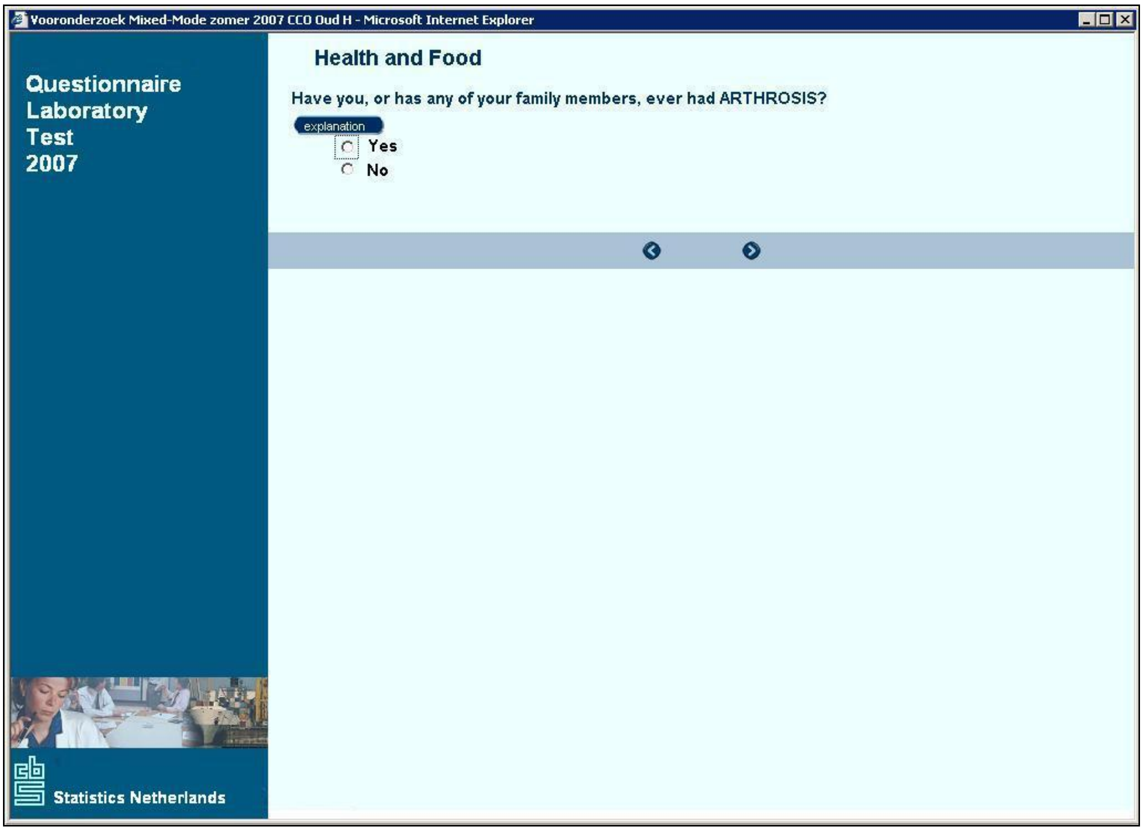
5-point scale question

Question A*.

"What do you think will happen to the unemployment rate in the Netherlands the next 12 months? Do you think it will increase, decrease or stay the same?

1. Increase a lot.
2. Increase a little.
3. Stay the same.
4. Decrease a little.
5. Decrease a lot.

Appendix. Screenshots of the questionnaire

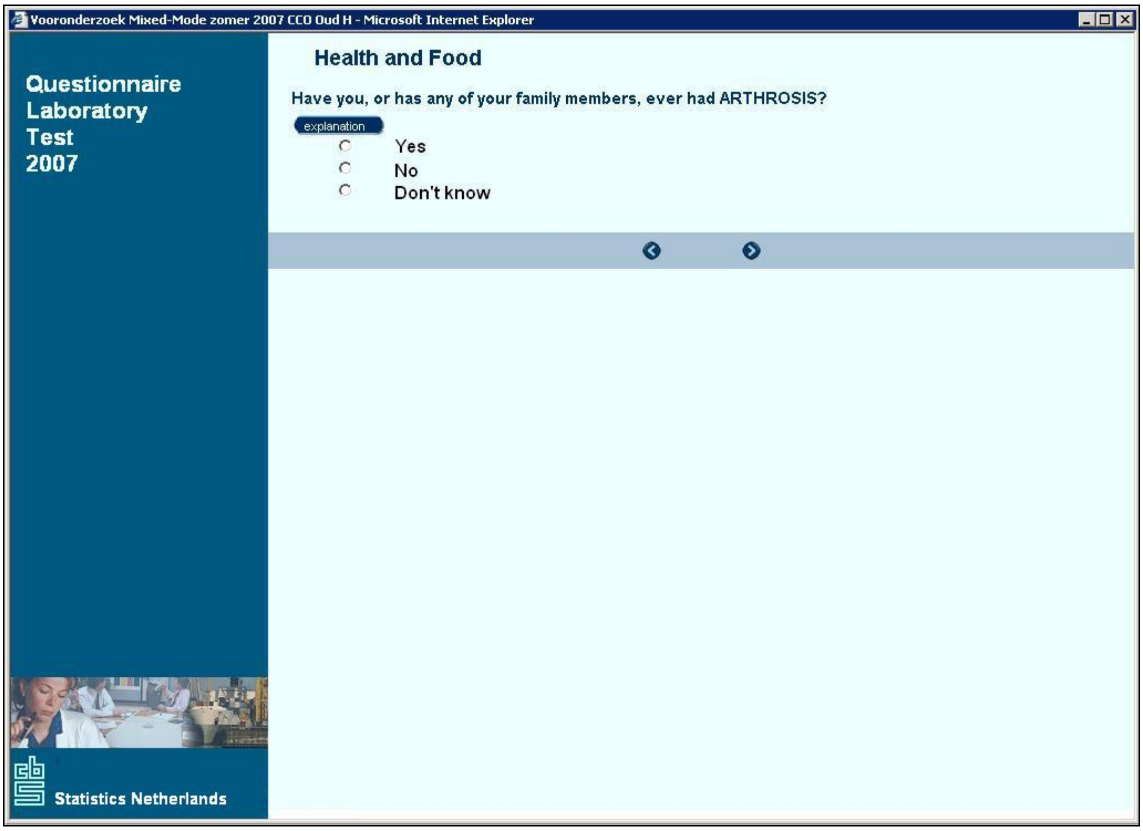


Vooronderzoek Mixed-Mode zomer 2007 CCO Oud H - Microsoft Internet Explorer
Questionnaire Laboratory Test 2007
Health and Food
explanation
Statistics Netherlands

Have you, or has any of your family members, ever had ARTHROSIS?

○ Yes
○ No

Figure 1. "Don't know" is not presented on screen.

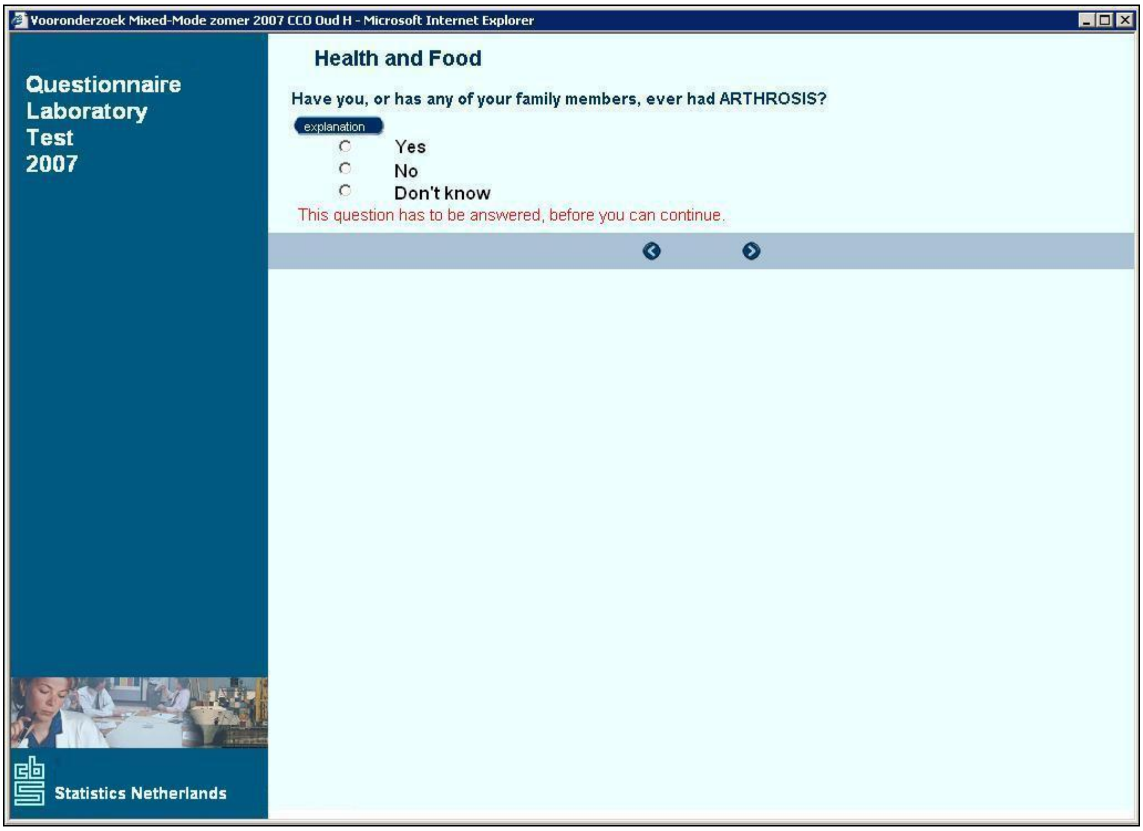

Vooronderzoek Mixed-Mode zomer 2007 CCO Oud H - Microsoft Internet Explorer
Questionnaire Laboratory Test 2007
explanation
Statistics Netherlands

Health and Food

Have you, or has any of your family members, ever had ARTHROSIS?

○ Yes
○ No
○ Don't know

Figure 2. "Don't know" is presented together with the other response options.


Vooronderzoek Mixed-Mode zomer 2007 CCO Oud H - Microsoft Internet Explorer
Questionnaire Laboratory Test 2007
explanation
Statistics Netherlands

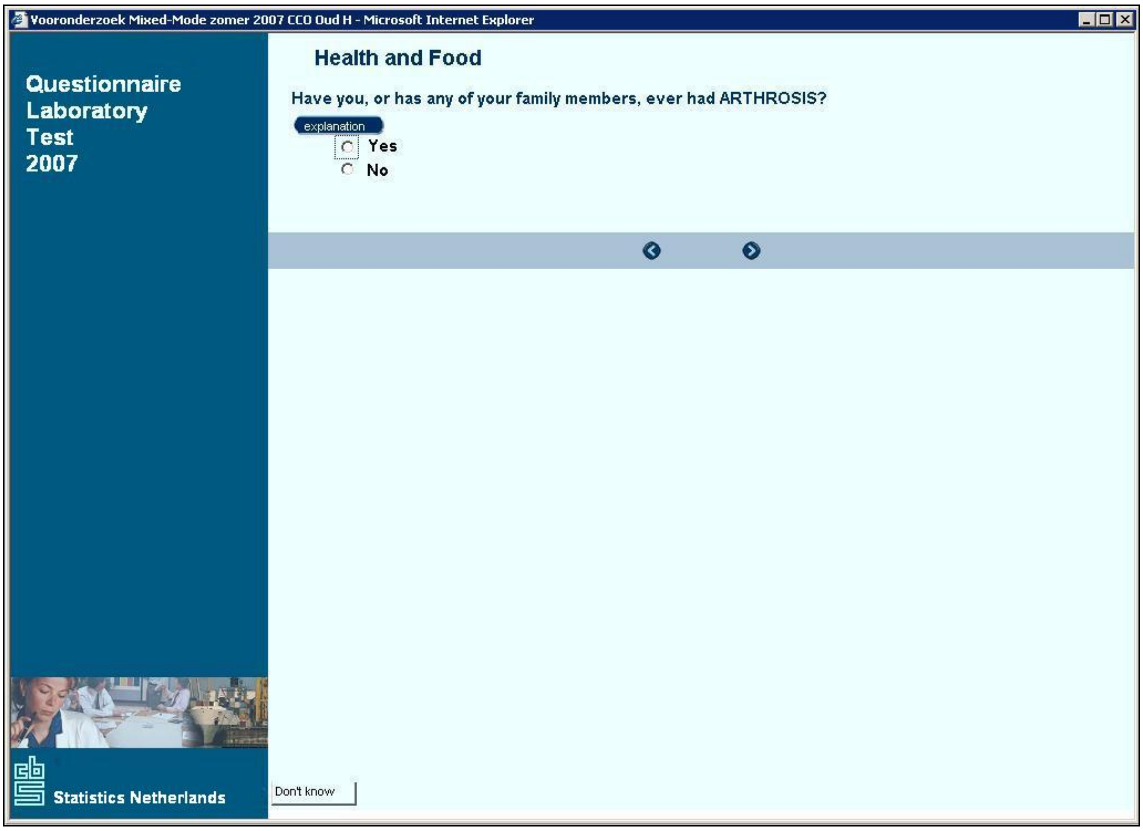
Health and Food

Have you, or has any of your family members, ever had ARTHROSIS?

○ Yes
○ No
○ Don't know

This question has to be answered, before you can continue.

Figure 3. A warning (in red) appears if you try to skip a question, and "Don't know" is presented together with the other response options.



Vooronderzoek Mixed-Mode zomer 2007 CCO Oud H - Microsoft Internet Explorer

Questionnaire Laboratory Test 2007

Health and Food

Have you, or has any of your family members, ever had ARTHROSIS?

○ Yes
○ No

Figure 4. "Don't know" is presented as a button on the bottom left side of the screen.

References

- Ariel, A., Giesen, D., Kerssemakers, F. & Vis-Visschers, R. (2008) Literature Review on Mixed-Mode Studies. StatNeth Internal report.
- Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. (eds.). (1991). Measurement Errors in Surveys. New Jersey: John Wiley & Sons.
- Bishop, G.F., Oldendick, R.W., Tuchfarber, A.J., & Bennett, S.E. (1980) Pseudo-Opinions on Public Affairs. Public Opinion Quarterly 44, 198-209.
- Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. Journal of Public Health, 27, 281-291.
- Brakel, J.A., van den, Vis-Visschers, R., & Schmeets, J.J.G. (2006). An experiment with data collection modes and incentives in the Dutch family and fertility survey for young Moroccans and Turks. Field Methods, 18, 321-334.
- DeRouvray, C. & Couper, M.P. (2002). Designing a strategy for reducing “no opinion” responses in web-based surveys. Social Science Computer Review, 20, 3-9.
- Gilljam, M. & Granberg, D. (1993). Should we take don’t know for an answer? Public Opinion Quarterly, 57, 348-357.
- Groves, R.M. (1989) Survey Errors and Survey Costs. New York: John Wiley and Sons.
- Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5, 213-236.
- Krosnick, J.A., Holbrook A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., Mitchell, R.C., Presser, S., Ruud, P.A., Smith, V.K., Moody, W.R., Green, M.C. & Conaway, M. (2002). The Impact of “No Opinion” Response Options on Data Quality. Non-Attitude Reduction or an Invitation to Satisfice? Public Opinion Quarterly, 66, pp. 371-403.
- Leeuw, E.D., de (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21, 233-255.
- Presser, S., & Schuman, H. (1989). *The management of a middle position in attitude surveys*. In E. Singer & S. Presser (Eds.), Survey Research Methods (pp. 108-123). University of Chicago.
- Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. ESRC National Centre for Research Methods, NCRM review papers 008.
- Snijkers, G.J.M.E. (2002). Cognitive laboratory experiences: On pre-testing computerised questionnaires and data quality. Doctoral thesis: Utrecht University.
- Tourangeau, R. & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 103, 299-314.
- Vis-Visschers, R. (2007a). Vooronderzoek Mixed-Mode. Project Initiatie Document. Versie 4.0, 10 mei 2007. Interne nota.
- Vis-Visschers, R. (2007b) Cognitive Pre-Tests on How to Present “Don’t Know” Help Texts in Web Surveys - Results from the first test. Proceedings of the 6th Conference on Questionnaire Evaluation Standards (Quest) 24-26 April, 2007, Ottawa, Canada, pp 15-23.
- Vis-Visschers, R., Arends-Tóth, J., Giesen, D. & Meertens, V. (2008) Het Aanbieden van ‘Weet Niet’ en Toelichtingen in een Webvragenlijst. StatNeth Internal report.

Statistisk sentralbyrå
Statistics Norway

Notat

Testing questions on sexual identity: experiences from optional data collection modes

Paper presented for the QUEST workshop in Bergen, May, 2009 – updated version

By Elisabeth Gulløy and Gustav Haraldsen[1]
Statistics Norway

In the following paper, we will present results from a Statistics Norway development project to measure sexual identity in the living conditions survey 2008. This includes both planning and test results and its effect on the final question and data collection design. We also present some experiences with the optional mixed mode design.

After presenting the test results and the final question and data collection design, the following questions will be discussed:

- Non-response errors and measurement errors
- What kind of feedback from respondents did the interviewers report?
- Optional mixed mode: how did it work out for the organisation?

Introduction

For the first time in 2008/2009, the Norwegian living conditions survey included questions on sexual identity. The initiative came from external users of statistics; government, research and interest organisations, as a response to the increased attention towards both identity issues and sexual minorities. Estimates of the amount of non-heterosexuals in Norway differ from 2 to 8 percent of the adult population, and there are substantial differences in how research agencies and other producers of statistics measure the phenomenon (Gulløy et. al. 2009). The management of Statistics Norway decided to meet the demand for information by taking up a development project, financed by the Ministry of Children, Equality and Social Inclusion.

First, the project consisted of a literature review, mainly looking at empirical studies from the Nordic countries. The review concluded that there was some evidence that sexual identity does affect living conditions in a substantial way, particularly in the area of mental health and exposure to violence. On this background, it was decided to run a pilot survey where questions on sexual identity were included. The pilot should be a part of the Norwegian living conditions survey.

Second, the Division for data collection in Statistics Norway was given the task to develop a good measure instrument and effective data collection procedures, all in the pursuit of avoiding high levels of item non-response (and of course to avoid respondents who abrupt the interview) and also to collect valid and reliable data. To achieve this, the questions should have valid and mutually exclusive response categories, and generate a minimum of negative attitudes towards responding. The careful design should take note of the sensitivity of these issues in various respondent groups, and the data collection mode or combination of modes also should reflect this. Finally, the development project was supposed to come up with ideas and drafts on how to introduce the survey issue of sexual identity and these particular questions for the respondents.

[1] Thanks to Aina Holmøy and Marit Wilhelmsen, Statistics Norway, who contributed to the presentation at the QUEST conference.

Background

To ask questions on sexual identity is traditionally seen as problematic in social surveys. In general, survey planners are advised to reduce to a minimum the request for sensitive information (Dillman et.al 2009). If absolutely necessary to include, such questions should come with a thorough information strategy, explaining directly or indirectly why this is so important, and how the organisers plan to secure confidentiality for each respondent. Dillman et. al. also states that "…the choice of question wording can help "soften" the requests for personal information" (2009).

Sensitive issues are a possible threat to response rates, but the relationship between sensitivity and non-response is complicated. For instance, a study performed by Office for National Statistics (ONS) in Great Britain showed that non-respondents on questions on sexual identity did not refuse to answer other types of sensitive questions, like income (Betts 2007, Wilmot 2007). Apparently there is no straightforward co-variation in the tendency to respond to questions on different types of sensitive issues. For some groups, questions on ethnic background are more sensitive than questions on income or education.

From previous studies we know that four groups of respondents are expected to be particularly sceptical to questions on sexual identity:

- The elderly, since they seldom speak about these issues, or are not sexually active
- Persons from immigrant societies with strong taboos on sexuality
- Persons being uncertain about their own sexuality, and therefore uncertain about what to answer
- Persons with an established sexual identity unknown to their families, friends and colleagues

It is reason to believe that the various ways we communicate with respondents in a survey process affects the perception of particular questions as sensitive or not. Thus, it was important for the development project to identify a set of strategies to meet the different respondent groups according to their possible source for scepticism.

The development phase

Statistics Norway decided to develop a draft set of questions for a pilot to be included in the living conditions survey focusing on health 2008-2009. The project included both literature studies and careful testing of questions. This included both formerly used questions picked up form other researchers or statistical agencies, and our own newly developed questions. The process can be characterised as an iterative process of testing – development – new testing – revision etc. Cognitive and focus group interviews included both potential respondents and Statistics Norway's own interviewers.

The development and testing gave the following conclusions:

1. we must secure an understanding among the respondents about sexual identity as a relevant dimension of living conditions and health, and that the questions are an integral part of a living conditions survey framework
2. sexual *attraction, orientation and identity* are different concepts and the survey questionnaire has to reflect this
3. sensitivity is a larger problem for interviewers than for most of the respondents

Questionnaire and data collection design

We ended up with four questions, ideally to be asked in a personal interview/telephone interview, but with an option for including them in a paper questionnaire distributed by mail after the interview. The final wording of the questions was as follows;

Sid1

To what extent do you feel that your sexuality affects your quality of life? Would you say it affects it….

1... a great deal
2 ... somewhat
3. ... slightly or not at all
4. DO NOT WISH TO ANSWER

If Sid1 is not equal to 4
Sid2

Which sex do you feel attracted to? Would you say…
READ THE NUMBERS IN FRONT OF EACH RESPONSE CATEGORY SO INTERVIEWEE CAN RESPOND BY STATING NUMBER ONLY

1. **..that you only feel attracted to men**
2. **..that you feel attracted to both men and women**
3. **..that you only feel attracted to women**
4. FEEL NO ATTRACTION TO ANYONE
5. DON'T KNOW
6. DO NOT WISH TO ANSWER

If (Interviewee is female and Sid2 = 2 or 3) or (Interviewee is male and Sid2 = 1 or 2)
Sid3

Do you regard yourself as being gay/lesbian, bisexual or heterosexual?
READ OUT THE NUMBERS IN FRONT OF EACH RESPONSE CATEGORY SO INTERVIEWEE CAN RESPOND BY STATING NUMBER ONLY

1. Gay or lesbian
2. Bisexual
3. Heterosexual
4. NO, NONE OF THE CATEGORIES
5. DON'T KNOW
6. DO NOT WISH TO ANSWER

If Sid1 = 4 or Sid2 = 6 or Sid3 = 6
Sid_post

I appreciate that it may be uncomfortable to talk about this topic. So, I suggest we send you the questions in a questionnaire by post. You will then have a chance to look at them in peace and quiet, before deciding whether or not you wish to answer them.

Yes/No

If Sid3 = 1 or 2
Sid4a

Has your sexual orientation ever caused problems for you in relation to your family?
Yes/No

Sid4a

Has your sexual orientation ever caused problems for you in relation to friends?
Yes/No

If (Work1a or Work1b = Yes) or IY1 = Yes
Sid4a

Has your sexual orientation ever caused problems for you in relation to co-workers?
Yes/No

The questions aim to measure what sexuality means for quality of life, sexual attraction to same/opposite sex, sexual identity, and finally whether sexual orientation has ever caused problems in relation to family/friends/colleagues. Heterosexuals will only be asked the first two questions. The last question is only asked to homosexuals, lesbians, bisexuals, or those who claim they are none of these alternatives. Thus, the sub-sample for question 3 on sexual identity is thus made up only of those who feel sexually attracted to the same or both sexes.

In addition, the data collection instrument was designed as follows:

- Topic is presented as one of several questions on lifestyle and quality of life
- Response given by stating numbers
- The first question emphasises the relevance
- Distinction between sexual attraction and sexual identity
- Also questions on perceived problems resulting from sexual identity

The data collection procedure should consist of:

- Telephone interview
- Option of self administered postal questionnaire if respondent declines to answer one of the questions
- Interviewer debriefing

Survey of Living Conditions 2008 – a multi-mode survey

The survey comprised 10,000 persons, aged 16 years and over, selected for personal in-home interview and in accordance with Statistics Norway's two-step sampling plan[2]. The interviews were conducted with a personal computer. Data collection lasted for six months in the winter of 2008/2009. Although the development project concluded that the questions on sexual identity should ideally be asked in telephone interviews, the survey was designed as a personal-in-home survey. However, in practice, the majority of the interviews were conducted as telephone interviews. In sum, the survey method was a combination of personal interview, *either by phone or in-home*, and a postal self-administered questionnaire form sent out subsequently. In other words, it was a *multi-mode survey*.

The postal questionnaire was designed to contain the questions which by tradition would be regarded as sensitive in a personal interview setting; concerning life situation and coping, psychological mood and worries, use of medicinal drugs, serious life events, use of health services, alcohol, drugs and gambling. The fact that the pilot (which included questions on sexual identity) concluded that those persons who declined to respond to questions on sexual identity during personal interview were to be offered a self-administered variant of the questions instead was therefore consistent with the customary practice for any living conditions survey dealing with sensitive topics.

The result was therefore two versions of the postal supplement: one containing questions on sexual identity designed for those who preferred to respond to such questions by self-administered questionnaire form rather than in a personal interview setting, and one without these questions.

In sum, this means that the questions on sexual identity were made in three different modes: during a personal in-home interview, during a telephone interview or in a postal self-administered questionnaire form. The choice between the first two modes was made primarily by the interview organisation, although the interviewees themselves could influence this by saying how they preferred to be interviewed. The decision regarding telephone or personal in-home interview was informed partly by financial-administrative considerations (the lower cost of telephone interviews) and partly by statistical considerations (optimal response quality from personal in-home interview for vulnerable groups in terms of the survey's principal topic, which was health). Experience also indicates that certain sample groups (elderly persons especially) find telephone interviews more difficult than in-home interviews. In-home interviews are thus a "scarce commodity" and to some extent restricted to certain respondent categories.

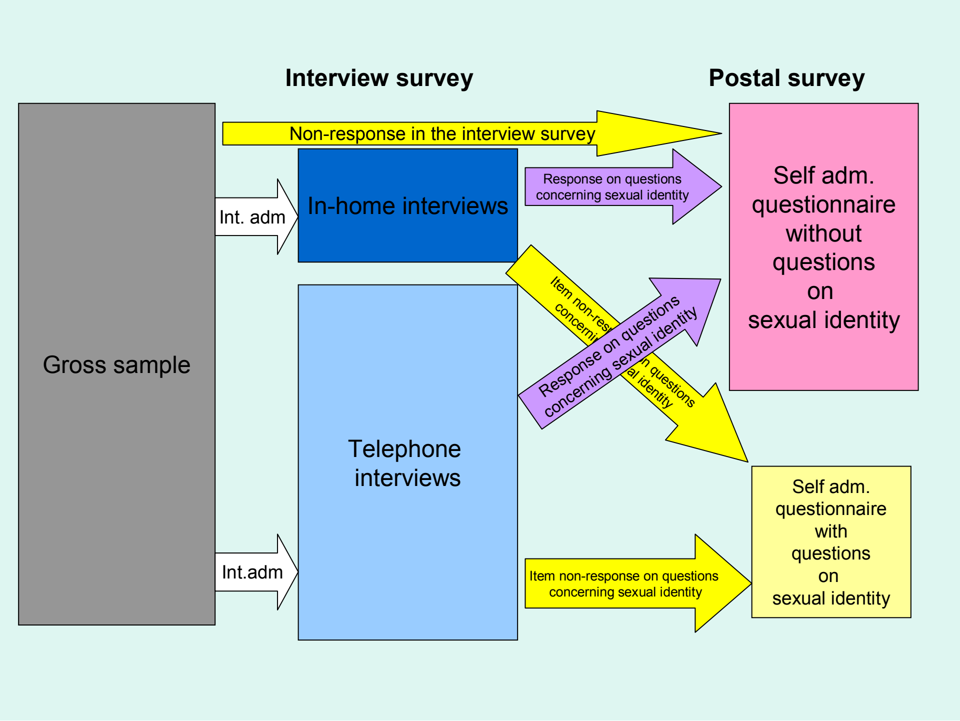
[2] In the documentation report for the survey (Wilhelmsen, 2009) the sampling and data collection are presented in more detail.

Accordingly, the distribution of personal in-home interviews is not random. At the start of the data collection, the plan was for the local interviewers in different parts of the country mainly to conduct personal in-home interviews, while the centralised staff (*CATI-interviewers*) were to deal with the telephone interviews. Ultimately, the strategy was for elderly persons and long-term ill persons to be prioritised for in-home interviews. In addition, individuals who specifically requested them were granted in-home interviews.

Unlike the in-home interviews, a respondent's assignment to the postal sample for questions on sexual identity was solely a result of the interviewee's response (or strictly: unwillingness to respond) during interview. But nonetheless there were thus two "paths" to the self-administered postal questionnaire: either via phone interview or via in-home interview.

Figure 1.1 presents an overview of the survey's various components and the spread of the gross sample of 9,684 individuals across the different components. [3]

Figure 1.1 Survey of Living Conditions 2008 – different components of the survey


Interview survey
Postal survey
Non-response in the interview survey
Gross sample
Int. adm
Int.adm
In-home interviews
Telephone interviews
Response on questions concerning sexual identity
Item non-res
concerni
n questions
al identity
Response on questions concerning sexual identity
Self adm. questionnaire without questions on sexual identity
Item non-response on questions concerning sexual identity
Self adm. questionnaire with questions on sexual identity

Experiences from data collection

The effects of including questions on sexual identity in Survey of Living Conditions 2008 should be assessed in terms of the impacts on *representation* and *measurement*. On the one hand, we have the potential effects the questions on sexual identity may have had on representativity in the survey through *increased and/or biased non-response*. What level of representation was achieved for the survey's different components and what level for these questions compared with the rest of the survey?

Then we have possible effects of non-response on the estimates, both the non-response that "always" occurs in the living conditions surveys, and those that occurred in this particular survey because we asked questions about what may be assumed to be a highly sensitive issue.

Finally, we have the potential effects on the estimates of the questions in themselves. Is there reason to believe that measurement errors occurred as a result of the content and design of the questions?

The problems are thus as follows:

[3] 316 people were characterised as absent.

- Have the questions on sexual identity resulted in a higher unit non-response or item non-response in Survey of Living Conditions 2008 (than would otherwise have been the case)?
- How do the scale and composition of non-response affect the estimates regarding sexual identity?
- How do the questions themselves affect the estimates?

To answer these questions, we split the analysis into four parts:

1. presentation of non-response at the different stages and in the different modes of the collection (both unit and item non-response)
2. assessment of whether the questions affect unit and item non-response
3. assessment of the effect of non-response on the estimates
4. assessment of the effect of the questions on the estimates

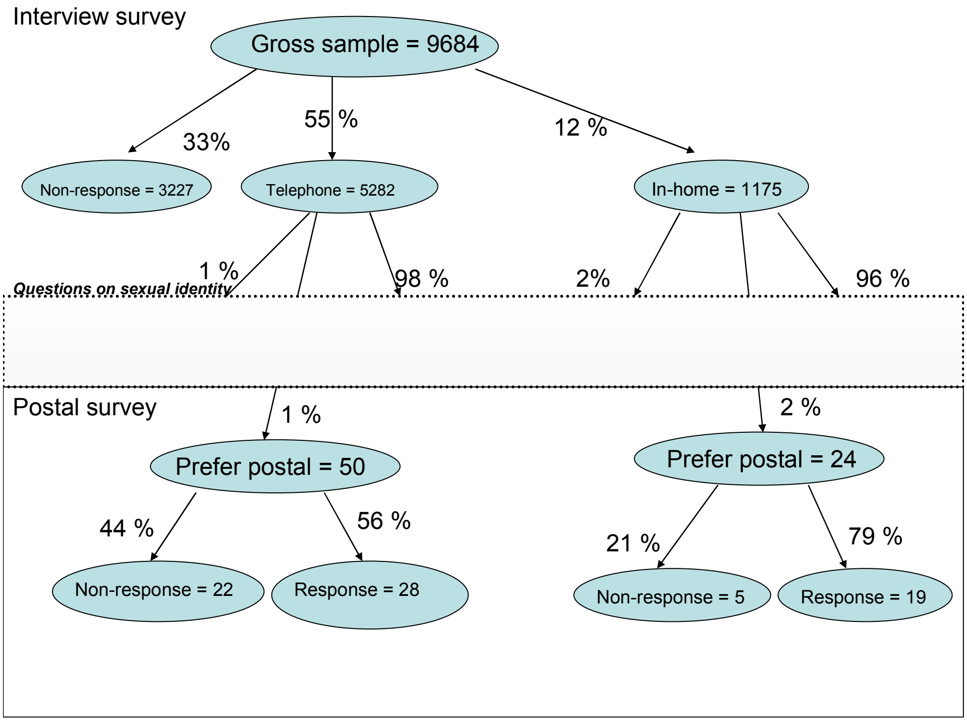
Non-response in the different stages and modes of the collection

In the information letter for the survey, respondents were informed that the topic of the survey was “health, care-giving and social contact” and that the questions would include ’how you rate your state of health, health services you have used and care-giving you have provided’. No detailed information was provided about the specific sections in the questionnaire. In other words, the respondents were not informed that sexual identity was one of the topics until during the interview itself. In terms of the sensitivity issue, unit non-response is therefore only of interest as regards the postal supplements, *unless the questions caused respondents to terminate the entire interview on reaching this point in the interview session.* However, the results show that no-one terminated the interview after the section on sexual identity, hence this effect is equal to nil.

But what was the scale of total non-response in the different phases and sections of the survey?

In Figure 1.2 we provide an overview of all the response and non-response rates over the course of the survey[4].

Figure 1.2 Responses, unit non-response and item non-response for the questions on sexual identity in the different sections of the survey. Absolute numbers and per cent.



The interview survey yields a response rate of 67 per cent, which is what one might expect and a fairly good result. The non-response in the interview survey varies according to gender, age and region. Women have a somewhat higher response rate, while the oldest age group (80 years and over) shows a good deal higher non-response than the average for the survey (more than 10 percentage points). The response rate is also higher than the average in the Trøndelag (Central Norway) and Northern Norway regions (Wilhelmsen 2009). Women in general are more willing to participate than men. Broken down by gender and age, we find that among women 80 years and over, many decline to participate (response rate of 50 for this age-group), and that men between the ages of 25 and 44 also account for a relatively low response rate. In this group there are both many who decline to participate and many that Statistics Norway is unable to make contact with.

Unit non-response in the interview survey increases marginally (less than 0.5 percentage points) as a consequence of introducing questions on sexual identity. Unit non-response in the postal survey containing questions on sexual identity is substantial if considered as an "isolated" survey. For those who fail to return the postal questionnaire, this also means item non-response for the other postal questions in the survey (mental health, drugs etc). However, this does not represent more than 27 individuals, or 0.4 per cent of the net sample.

Item non-response

The rate of item non-response is a quality determinant for the data material. How many respondents declined to respond to the questions on sexual identity but otherwise participated in the survey by phone interview, in-home interview or by postal questionnaire? In order to estimate this, we have to go back to the figures that emerged in Figure 1.2. Here we saw that 1 and 2 per cent respectively of phone and in-

[4] Note that the figure presents aggregate figures for the number of observations who decline to respond to one or more of the questions in the section – for individual questions this varies.

home interviews ended with “outright” refusal to answer the questions on sexual identity. But in addition, there was then also a proportion of respondents who agreed to complete the postal questionnaire, which is also a form of item non-response in the interview survey. *This pushes the item non-response rate up to 2 and 4 per cent respectively.*

However, these aggregate figures mask the trend for each of the questions in the section. Note that the option to respond via a self-administered questionnaire form was only to be offered if the respondent answered “don’t know”, or declined to answer one or more of the three first individual questions in this section. In this way, those who end up in the sample for the postal survey may still have responded substantively to one of the first questions. Even if they did, it is the responses from the postal questionnaire that are included here.

The question-skip structure was designed so that everyone would be asked the first two questions on quality of life and sexual attraction, while only those attracted to the same or both sexes would be asked the questions on identity and problems. Starting from the question on sexual identity, the net sample thus consisted exclusively of persons who state that they feel attracted to the same or both sexes. This group may include persons who live with a hidden sexual identity, persons who are vulnerable to the sensitivity in the interview setting or the content of the questions.

In Table 1.1 we show item non-response for individual questions, both as regards questions on sexual identity and other potentially sensitive questions in the survey. We also break results down by phone, in-home and postal mode. Note that “Don’t know” is counted as a substantive response in this analysis, since we regard it as a valid expression for an individual unsure of his or her own sexual orientation or identity.

Table 1.1 Item non-response for sensitive individual questions in different response modes

	Total	Phone interview	In-home interview	Postal supplement
Questions on the importance of sexuality for quality of life	1,7 %	1,4 %	2,0 %	23,4 %
Questions on sexual attraction	0,3 %	0,2 %	0,1 %	19,1 %
Questions on sexual identity	8,5 %	5,4 %	4,9 %	19,1 %
Questions on problems in relation to family due to sexual identity	13,8 %	15,4 %	6,7 %	17,0 %
Questions on problems in relation to friends due to sexual identity	12,1 %	10,3 %	6,7 %	17,0 %
Questions on problems in relation to co-workers due to sexual identity	18,1 %	20,5 %	16,7 %	17,0 %
Questions on contact with doctors	0,8 %	0,8 %	0,4 %	
Question on height	0,4 %	0,5 %	0,2 %	
Question on weight	2,0 %	2,1 %	1,6 %	

Source: Survey of Living Conditions 2008, Statistics Norway.

The first question, concerning the extent to which sexuality affects quality of life, produces an item non-response of 1.7 per cent in total; that is, below the level for item non-response for the question as to how much the person weighs. If we then move on to the question concerning attraction, there is a drop to below 0.5 per cent, but a rise above 8 per cent when we ask about sexual identity. A relatively high number decline to answer the questions on problems this has caused in relationships with others. The highest level of non-response is to the last question, concerning problems in relation to co-workers, with item non-response at 18 per cent overall. The most obvious explanation for this is that a certain “response fatigue” had set in among respondents by this stage. It may be the case that respondents find it wearing and/or unpleasant to take such personal and relatively intrusive follow-up questions three in a row. The last three questions underwent less testing in the development project than the first questions. There is the possibility that they are ill-considered in cognitive terms, and that what the respondent is being asked to consider is simply too difficult. This should perhaps be investigated in more detail subsequently.

The tendency is seen in all modes. The phone interviews result in a higher level of item non-response than the in-home interviews as we approach the end of the section. The greatest effects are seen in the small postal sample. Here the proportion of non-responders is much higher across the board than in the interview samples. This is not surprising, since the postal sample is made up of individuals who reacted negatively at interview. Besides which, the self-administered questionnaire form lends itself more readily to item non-response, since the respondent has time to (re)consider the content of the questionnaire before, during and after responding. But again, we find that all the respondents revert to responding after this section; no one opts out of the entire survey after these questions.

Those who decline to answer *and* to receive a postal questionnaire containing these questions are hereby called the "firm refusers". They were instead sent the ordinary postal questionnaire. For this group, item non-response is limited to the questions on sexual identity.

In the interview survey, the first two questions in the section, on how sexuality affects quality of life, and sexual attraction, produce low item non-response; so low that it may be characterised as negligible, seen in the context of other sensitive topics also covered by the survey. The scale of non-response to questions aimed solely at respondents who feel attracted to the same or both sexes (question on sexual identity and any social consequences of this) is substantial.

In the postal survey, item non-response to the questions on sexual identity is extremely high. This confirms that the questions were perceived as highly sensitive to the small group of respondents who declined to answer these questions during interview.

Introducing the questions on sexual identity has not had any appreciable effect on the ordinary postal supplementary survey. Indeed, for a postal supplementary survey, this was a highly satisfactory response rate. Even in the category of "firm refusers", for whom we only have interview data, half of the postal questionnaires were returned. These results are as good as those achieved by sending out postal questionnaires as part of the non-response follow-up on the survey.

Effects of the questions on non-response

The introduction of the questions on sexual identity in Survey of Living Conditions 2008 did not result in any appreciable increase in unit non-response for the interview survey and the ordinary postal survey. The former is not surprising, the latter is perhaps more unexpected.

If the questions were felt to be so sensitive that their very presence in the survey caused offence to the respondent, then this should have resulted in greater non-response than otherwise to the ordinary postal survey. But this was not the case. There is however substantial unit non-response to the special postal survey containing questions on sexual identity. The "firm refusers" push non-response up dramatically even though the response rate among those who agreed to receive the questionnaire form was in fact very satisfactory (64 per cent).

One might expect a greater sense of obligation bound up with receipt of the postal questionnaire containing questions on sexual identity than the ordinary postal questionnaire. The recipients, in full knowledge of their content, have committed to them in advance, by agreeing to receive them. The ordinary postal questionnaires, however, are sent out without the respondents being 'forewarned' of their arrival. Still, fair results were achieved for both versions, although still poorer for the version dealing with sexual identity.

In addition, there is the seriously high item non-response to the last questions in the section, especially in the postal questionnaires. For the majority, however, the two preliminary questions, on how sexuality and sexual attraction affect quality of life, are quite straightforward.

The generally positive results in terms of item non-response in the postal surveys may be interpreted as a confirmation of the fact that *efforts to stress relevance and consistency had an encouraging effect for the great majority of the sample.*

However, we would stress that the sensitivity of the questions in this section was a reason for item non-response. Moreover, the marginal increase in unit non-response among the "firm refusers" is obviously connected with the topic of the questions. People who react strongly against these questions tend to refuse to respond at the first instance and many of them then also fail to respond by post. *This means* *that we have a small core-group of respondents for whom the questions are not conducive to participation in an interview setting.* This amounted to 185 persons and approximately 2.9 per cent of the net sample for interview survey. In this group, we can then draw a distinction between those who opt to respond to questions by post, and those who either refuse to receive them as a postal questionnaire or who decline to answer the individual questions in the self-administered questionnaire form.

Mode-effect?

Did the mode in which the questions were originally asked have any effect? In Figure 1.2 we saw that the proportion of "firm refusers" and the proportion of those who prefer a postal questionnaire was twice as high for the in-home interviews as for the phone interviews. The closer the interviewer gets to the respondent, the more difficult it seems to be to obtain a response to these questions. The in-home interview clearly makes the interview setting more personal, and for those who in the first place find these to be problematical questions, there is correspondingly a greater risk that they will refuse to respond to them. *This outcome seem to confirm the conclusion of the development project; that the questions are best suited for telephone interviews.*

Effects of the non-response on the estimates

The results from the survey show that less than two per cent of the population aged 16 and over feels attracted to the same or both sexes. Just over one per cent categorise themselves as gay, lesbian or bisexual. This however varies depending on gender and age, while educational attainment and place of residence have little influence. Because the proportions are small and the sample relatively limited, the figures are statistically uncertain however. The estimates are below those from a number of research environments, but are still consistent with comparable surveys, among others from Sweden (Swedish National Institute of Public Health 2005).

We have concluded that the questions on sexual identity have had some influence on non-response in this survey, but this varies depending on which questions we are surveying. When we come to assess the effect this has on the estimates for sexual attraction and identity, we find it differs from one question to the next. However, we are not in a position to estimate with any certainly the *scale of the effect* of non response. The trends tend in different directions. Furthermore, we cannot rely on the presumptions when assessing potential bias as a result of non-response: that there is a similarity between those who respond and those who decline to. Since homosexuality remains taboo in many circles and cultures, we should expect a certain amount of systematic under-reporting whatever the case. There is therefore (more so than usual) uncertainty as to whether one may conclude that respondents in the non-response group would have produced the same distribution as for those who did respond. *There is reason to assume that the proportion of gays, lesbians and bisexuals is larger in the non-response group than in the sample as a whole. We can then assume that the estimates are too low. It is impossible to know how much lower, and the significance of the non-response can only be roughly suggested.*

For the first two questions in the section, on the importance of sexuality and sexual attraction, the level of item non-response was in fact of very limited significance for the estimates because it was so small. If we assume nonetheless that the tendency for sexual attraction has the same distribution among women in the group of "firm refusers" as it has among the women who did respond, the proportion that is attracted to both sexes or the same sex would be slightly higher in the estimate. At the same time, it is the case that respondents from the two oldest age-groups (aged 67 and over) show a tendency to be less attracted by both sexes or the same sex. These age groups are also under-represented. As such, we see that the two non-response tendencies are contrary to each other. Two other groups are more weakly represented: respondents from sparsely populated districts and with the lowest level of educational attainment produce higher item non-response, while the tendency to feel attracted to the same sex or both sexes shows little variation under the influence of place of residence and educational attainment.

The estimate for how many people feel no sexual attraction at all varies greatly depending on age. The proportion is a great deal higher among the oldest respondents, and especially in the age-group 80 and over. It is also somewhat higher among persons with low educational attainment and women, but this presumably correlates with the age variable. With a more even representation of elderly persons, women and persons with low educational attainment, the estimate of how many are "asexual" would presumably have been higher, provided that respondents and non-respondents share the same characteristics in this area.

For the last two questions, on sexual identity and social consequences, item non-response has had greater significance for the estimates. Here we examine the case for only the first of these two questions. The question(s) concerning social consequences give rise to such a high volume of non-response that we will not concern ourselves with analysing them. In any case, our primary interest is the identity question.

In the first place, only a very few will be asked these questions: only those who initially respond that they are attracted to the same or both sexes. This means that each non-response becomes more important. In addition, there is reason to believe that those who do not respond would in fact have influenced the estimate in the direction of more gays, lesbians and bisexuals, since a proportion of those who decline, presumably are concealing such an identity or orientation. But although “firm refusers” are quite distinct from the rest of the sample when it comes to the characteristics we surveyed for, there is no way of establishing *the true* distribution of gays, lesbians and bisexuals in the groups of elderly persons, women, people living in sparsely populated districts and people with low educational attainment.

Before looking at the estimate for gays, lesbians and bisexuals, it will be useful to look at bias in the group of “firm refusers” in terms of common characteristics.

Table 1.2 Gender, age, area of residence and education for “firm refusers” and the entire sample for questions on sexual identity

		Total	"Firm refusers"
No.		6457	108
Gender			
	Females	49,0 %	60,2 %
	Males	51,0 %	39,8 %
Age			
	16-24 years	13,8 %	23,1 %
	25-44 years	34,3 %	41,7 %
	45-66 years	37,1 %	25,0 %
	67-79 years	10,3 %	5,6 %
	80 years and older	4,5 %	4,6 %
Area of residence*			
	Sparsely populated areas	20,5 %	17,8 %
	Densely populated < 2 000 inhab.	8,4 %	4,7 %
	Densely populated 2 000-20 000 inhab.	25,7 %	23,4 %
	Densely populated 20 000-100 000 inhab.	21,9 %	16,8 %
	Densely populated 100 000 or more inhab.	22,5 %	37,4 %
Educational attainment**			
	1	26,3 %	32,3 %
	2	43,0 %	33,3 %
	3	31,0 %	34,3 %

*For the area of residence variable, register information was missing for 59 persons. These are not included in the calculation basis here

**For the educational attainment variable, register information was missing for 360 persons. These are not included in the calculation basis here

Source: Survey of Living Conditions 2008, Statistics Norway.

When it comes to questions concerning sexual identity there is still a higher proportion of women than men among the “firm refusers”, but the difference is not quite as large. The level is ten percentage points higher than for the entire sample. There are relatively more younger individuals up to the age of 44 in the group of firm refusers. The proportion of firm refusers is also higher in densely populated districts. A relatively larger proportion of persons with the lowest level of educational attainment decline to answer these questions than in the sample as a whole.

How does this affect the estimates for sexual identity? Women have a somewhat greater tendency to be homosexual or bisexual, and if we had had a less biased respondent group in terms of gender for this question, the differences would have been reinforced, assuming there was the same pattern in the non-response group. In relation to age, the proportion of gays, lesbians and bisexuals decreases with increasing age. If the response rate from the younger generation had been higher, then this would also have driven the estimates for non-heterosexuals up. When it comes to geography, correlations between place of residence and identity are uncertain; there is a slightly higher proportion of gays, lesbians and bisexuals in both sparsely populated districts and in the largest towns/cities. For sparsely populated districts, the "firm-refuser" group is not more heavily represented here than elsewhere, while the largest towns/cities are over-represented in the "firm refuser" group. This perhaps increases the likelihood that lower non-response from the major towns/cities would also have influenced the estimates in the direction of a higher proportion of gays, lesbians and bisexuals. Finally, we have the educational attainment factor, which does not appear to be a significant background variable for sexual identity either. Here we have the same pattern as for place of residence: the proportion of non-heterosexuals is somewhat higher in the lowest and highest level educational attainment group and both of these are also slightly over-represented in the firm refuser group.

The differences are too small to be given any emphasis, but all told, the composition of the firm refuser group, for all the four background variables we surveyed for, tend in the same direction. *The estimates for gays, lesbians and bisexuals are too low.*

In the postal supplement which includes the questions on sexual identity, the estimates will be greatly biased as a result both of non-response due to the fact that the preceding interview contained these sensitive questions ("firm refusers"); as a result of the fact that high non-response in postal surveys is generally not randomly distributed in the population; and as a result of the high item non-response for individual questions. This sample cannot be regarded as representative, or as an independent postal survey. The units must only be included as part of the main sample. They are however still of interest for the analysis, because they consist of respondents who represent one of the two main challenges for the project: how to secure responses from those who are willing to take part in the survey but sceptical about responding to these specific questions in an interview setting.

Effects of the questions on the estimates

Is it possible that our wording of the questions affected the estimates? The cognitive testing in the planning phase was intended to ensure that the wording and substance of the questions were comprehensible and perceived as relevant, while not causing offence. The results suggest that these aims were by and large achieved in that item non-response was not higher than it was – at least for the initial questions. People largely responded to the questions on the effect of sexuality and sexual attraction. The fact that item non-response begins to increase with the question on identity is due to the sensitivity of the topic, albeit for a relatively small proportion of the population.

The last questions, on social consequences of a gay, lesbian or bisexual identity, were not however worded well enough to yield representative results. These questions should probably undergo further cognitive testing before any attempt to use them again. In fact, this had already emerged in the very last round of cognitive testing before the survey commenced. The questions were nevertheless included, among other things because they were seen as important in shedding light on the relevance of the questions in the living conditions context.

We lack responses from many elderly persons. This is both because they object to these questions and because they are over-represented in the non-response group anyway. One of the reasons for the latter is bound up with the design of the survey in that we allowed persons who feel no sexual attraction to anyone skip the question on sexual identity. This then resulted in a less robust sample for the identity question. In the event of a repeat of this survey, the effectiveness of this solution could be discussed, but the need for representativity must be weighed against the need to make the respondents feel that the structure of the questionnaire is meaningful. For a person who is not sexually active, and who might perceive him/herself to be "asexual", a battery of questions on sexual identity may seem intrusive and interrogative.

Against that, it is of interest to reflect on what we were actually measuring with the questions on sexual attraction, and then to have let one of the response categories be "feel no attraction at all". Those who then selected this response alternative might be both persons who lead an "asexual" existence because this stage of life is a thing of the past or because they are not sexually active or seeking to be so. The difference between the two types will be temporal aspect; we might call it the "widow-type" and "single person without sexual interest-type". It is not hard to imagine that the latter group might also include persons who conceal homosexual tendencies. As a result of the design of the questionnaire form, such individuals were never asked the question on sexual identity, but even if they had, it is perhaps unlikely that they would have responded with anything but the most conventional answer.

Optional mixed mode: how did it work out for the organisation?

How successful was data collection with optional response mode for the organisation? No particularly negative experiences were reported although a certain added effort is involved in dealing with multiple versions of questionnaires and managing the postal supplements alongside them.

What lessons where learned that could be applied to optimise data collection in future?
In order to improve our ability to analyse the effects of multi-mode design, we need to keep more records on how data collection was done. Examples would be what mode (phone or in-home) was used in each of the different contact attempts. This would also have allowed us to calculate non-response in each of the two modes. We also lack data on how these questions worked in settings where they may have been disclosed to other family members, i.e. information about anyone else sitting in on the interview. This has been done in the past, for instance in the living conditions survey among immigrants, and it might be of interest to investigate on another occasion, for example in follow-up interviews with the respondents.

Was it wise to create valid "Do not wish to answer" options throughout the entire section even if these were not read out to the respondent during the interviews? There was a relatively high proportion of non-substantive responses throughout the section. The reason for this format was on grounds of sensitivity, and in order to satisfy respondents who are unsure of their own sexuality or of disclosing it, by making this a valid response category. In the UK, they have currently decided to abandon this type of response category. In the ONS cognitive interviews, there was no indication that this should have been necessary. On the contrary: in all groups, it was indicated that a category of this kind would tend to direct attention at non-heterosexuality. However, we still maintain that being unsure of one's sexuality can be regarded as a genuine situation, and should therefore also be included among the valid response categories.

Ultimately it may be worth discussing whether field work should be organised so that the gender of the interviewers is the same as the respondent's. This is good practice in surveys of immigrants, but will still be adapted to other needs, such as the interviewers' language skills, geography etc. However, there is some indication that this may reduce under-reporting of non-heterosexual identity.

The interviewers' expectations
The interviewers were known in advance to be sceptical. Accordingly, a general skills coaching programme on sensitivity issues was run prior to the survey. After data collection had been completed, focus group interviews were held to debrief the interviewers on their experiences. The outcome of these was that there had been very few problems in asking these questions in Survey of Living Conditions 2008. The interviewers were pleasantly surprised at how little "fuss" the topics had caused. It was pointed out that a number had been surprised; they had dreaded asking the questions, but discovered that their concerns were groundless. There were also no reports of any particular reactions from respondents to Statistics Norway as a result of these questions

Litterature

Betts, P. (2007) *Developing survey questions on sexual identity: UK experiences of administering survey questions on sexual identity/orientation*. Data Collection Methodology – Social Surveys, Census and Social Methodology Division, Office for National Statistics.

Dillman, D., J. D. Smyth and L. M Christian. 2009. *Internet, mail, and mixed-mode surveys. The Tailored Design Method. Third edition.* Wiley and Sons, New Jersey.

Gulløy, E., G. Haraldsen og T. M. Normann. 2009. *Kartlegging av seksuell identitet i Statistisk sentralbyrås levekårsundersøkelse. Dokumentasjon av bakgrunn og utvikling av spørsmål.* Notater 2009/22. Statistisk sentralbyrå.

Normann, T.M. and E. Gulløy. 2010. *Seksuell identitet og levekår. Evaluering av levekårsrelevans og datafangst.* Rapporter 2010/? (*forthcoming*). Statistisk sentralbyrå.

Statens Folkhälsoinstitut (2005) *Homosexuellas, bisexuellas och transpersoners hälsosituation. Återrapportering av regeringsuppdrag att undersöka och analysera hälsosituationen blant hbt-personer.* Statens Folkhälsoinstitut. Rapport nr A 2005:19

Wilhelmsen, Marit. (2009) *Samordnet levekårsundersøkelse 2008 – Tverrsnittsundersøkelsen. Dokumentasjonsrapport.* Notater 2009/40, Statistisk sentralbyrå Oslo/Kongsvinger.

Wilmot, A. (2007) *In search of a question on sexual identity*. Paper presented at the 62nd Annual Conference of the American Association of Public Opinion Research. Office for National Statistics, UK. May 2007.

What Kinds of Problems Does Cross-Cultural Pretesting Reveal?

Gordon Willis, Ph.D.

Applied Research Program

Division of Cancer Control and Population Sciences

National Cancer Institute

Abstract

Attempts to attain cross-cultural equivalence of survey questions have increasingly relied on the adaptation of pretesting techniques such as cognitive interviewing and behavior coding. I this paper, I summarize the general categories of problems with survey questions that are identified through such studies, and in particular those relating to translation, culturally-oriented issues, and general problems of question design. Two models have been proposed to encompass these concepts – one by Willis and colleagues (2007; 2008) and the other by Fitzgerald et al. (2009). I compare and contrast these models, and present a hybrid that encompasses the terminology of the former, and the conceptual structure of the latter.

What Kinds of Problems Does Cross-Cultural Pretesting Reveal?

Increasingly, researchers have conducted pretesting – cognitive interviewing and behavior coding in particular – to assess sources of error in multicultural, multilingual, and multiregional studies. Typically these investigations produce a range of problems that have the effect of precluding measure comparability across groups. That is, sources of variation are identified in which the survey questions function differently across groups, such that estimates from these groups cannot be meaningfully compared. Given the range of problems identified, it would be useful to develop a characterization or coding system that systematically organizes and chronicles the types of findings that have been observed. Rather than simply listing the specific errors that occurred for particular questions in each study, researchers could then ascertain patterns across studies in a way that relies on a common measurement system.

Several such error models do exist in the Cognition and Survey Methodology (CASM) field. Tourangeau (1984) developed the original four-stage model of the survey response process (involving cognitive processes of comprehension, retrieval, judgment, and response) that has formed the underpinning of much work within CASM, and in cognitive laboratories. Other models have focused more on characteristics of survey questions, rather than on the psychological processes giving rise to response error - e.g., the Question Appraisal System, or QAS (Willis and Lessler, 1999); and the Q-BANK Response Error model (Beatty, Willis, Hunter,, & Miller, 2005).

These models, however, do not have a strong cross-cultural emphasis. Hence, it is necessary to determine whether a revised model is needed when dealing with problems identified across languages and cultures. In particular, the application of cross-cultural pretesting has made clear that additional problems are introduced through the practice of language translation of the source questions into target versions. These additional elements have been incorporated into two simple models that have been developed independently- one by Willis and colleagues (Willis & Zahnd, 2007; Willis, Lawrence, Hartman, Kudela, Levin, Forsyth, (2008), and the other by Fitzgerald, Widdop, Grey, and Collins (2009). Beyond translation difficulty, such studies have revealed problems that appear to invoke cultural elements that are not well-represented by standard models of the survey response process.

Based on a study of Asians and Non-Asians in pretesting a tobacco questionnaire, Willis et al. (2008) proposed a simple, three-category model. The TCG model, incorporates (1) Translation errors; (2) Problems of Cultural Adaptation; and (3) Generic Problems of Questionnaire Design (note that the authors also included a fourth category – Question Misreading – but this most directly involved interviewer behavior as opposed to intrinsic question design, although it clearly could contain elements of the latter).

> *1) Translation problems* involve items that fail to express the meaning of a term, phrase, or entire question, as intended by designers, due to defects in the translation process. For example, Willis et al. (2008) found that the question*:* *"What price did you pay for the LAST pack/carton of cigarettes you bought? - Please report the cost after using discounts or coupons"* was mistranslated into Chinese, such that it directed the participant to report *only* the cost of the coupon (the opposite of the intended meaning). In response, the translation was modified to express the intended concept.

2) Problems of cultural adaptation, which supersede translation issues, as they are not clearly related to literal mis-translation. For example, Willis et al. (2008) found that few Asian respondents were familiar with snuff (however translated). A definition was therefore added for the interviewer to use as needed. Further, Asians were found to have problems with providing a response on a 0 to 10 scale, and survey interviewers and behavior coders suggested that Asians are not familiar with the use of this scale. The designers added a standardized instruction to Asian versions: "*Overall, on a scale from 1 to 10, where 1 is <u>not at all</u> interested and 10 is <u>extremely</u> interested, how interested are you in quitting smoking. Please indicate how interested you are in quitting by picking a number from 1 to 10.*"

3) As a final problem category, *Generic Design Problems* are not specific to any particular tested group. Willis et al. (2008) reported that the question **"***How soon after you wake up do you typically smoke your first cigarette of the day*?" tended to produce the same reaction for every version (including English): Rather than answering with a number such as "one hour," respondents tended to report "before breakfast" or "as soon as I open my eyes." Generic problems presumably include all those types that would fit into an existing error models, such as the QAS.

More recently Fitzgerald, et al. (2009) have introduced a fundamentally similar model, completely independently. In that model, both Cultural Issues and Generic problems -- which they title ‘Problems with the Source Question’ -- are recognized, and appear to be fundamentally similar to the Willis et al.’s respective concepts. However, the Fitzgerald model adds a fourth category, which can be viewed as the result of further sub-dividing a Translation category into two sub-types: (a) Translator Error, and (b) Source Question Interaction with Translation.

The first authors of these two models met at the Comparative Survey Design and Implementation (CSDI) conference in 2009, and discussed these similarities and differences. On that basis, I propose a modified, hybrid TCG model that reconciles the models, by including the four conceptual categories contained in the Fitzgerald model, but relying mainly on language and terminology from the Willis et al. model (See Figure 1).

Figure 1: Revised TCG classification of error in translated questionnaires.

1) **Translation problems**

   1a: Translator Error

   1b: Translation Difficulty
   (Consistent with Fitzgerald, et al.: “Source question interaction with the translation”)

2) **Problems of Cultural Adaptation**
   (Consistent with Fitzgerald et al.: “Cultural Issues”)

3) **Generic Problems**
   (Consistent with Fitzgerald, et al.: “Problems with Source question”)

Revised TCG model. To reiterate, according to the revised, reconciled model I propose here, Translation problems can be sub-divided into two major types: Translator Error; and Translation Difficulty. Translator Errors – as conceptualized by Fitzgerald et al. (2009) -- are due to outright flaws of language translation that, upon discovery and review, can be fixed without much further difficulty. Translation Difficulty, on the other hand, corresponds to what Fitzgerald et al. characterized as an interactive effect between question and translation -- where fundamental differences between languages or cultures precludes a straightforward translation from the source to the target language (significantly, Fitzgerald et al. use the word "difficulty" in describing such problems). In these cases, the translator is not seen as the source of the error. Rather, the problem is viewed as based on the fact that for whatever reason, the initial linguistic unit, as expressed in language *A*, cannot easily be linguistically recoded into language *B*, in a way that retains its meaning. Resolving such problems requires more than simply re-translating, and may in fact lead to a modification of the target question (i.e., decentering) in order to ease the task of producing a meaningful analog in another language or culture.

Model Utility

The usefulness of the distinctions made by the above models is in part determined by how easily specific instances of behavior or occurrences can be coded unambiguously. Further, the model must be non-trivial, and of more than academic interest, in that it provides a viewpoint of the relevant phenomena that is useful to practitioners of the science. I close by discussing these issues, with respect to the basic elements within the modeling approach taken by Willis and colleagues (2007, 2008) and by Fitzgerald et al. (2009).

a) Translation-related codes. Translation problems tend to present particular challenges of labeling, and coding. As an example, cross-cultural researchers sometimes find that terms such as "excellent" … "poor" are notoriously difficult to translate in such a way that the scaling differences existing in English convey similarly in different languages. Hence, for a question asking Chinese respondents to rate their health care, the term "excellent" was in one study found to have been mis-translated so that it came across as "god-like" (Brian Claridge, University of Massachusetts Survey Research Center, personal communication). In this case, one might apply a code of "Translator Error." However, given the intrinsic difficulty of translating these items, one might argue instead that the more refined category of "Translation Difficulty" (or alternatively, "Source question interaction with the translation") pertains better, as the problem is not necessary due to an error, but rather to difficulties in establishing language correspondence such that the categories function equivalently. That is, a translator may have made an "error" because the term is difficult to translate, and it is not clear where one flaw ends, and the other begins.

Hence, it may be difficult to assign a code unambiguously, beyond the simple level of "Translation problem" (as in the original TCG model). Although I believe that the further distinction originally made by Fitzgerald and colleagues is useful, and should be retained, I also propose that these more specific subtypes often shade into one another. Hence, I propose that these two subtypes share a common superordinate label (more concretely, by retaining the "T" in an overall three-category TCG model). In any event, it seems evident that at least conceptually, problems associated with translation can be meaningfully separated from the other two major elements of this model.

b) Codes dedicated to culturally-related problems. Whether termed "Problems of Cultural Adaptation", or "Cultural Issues," critics with an anthropological perspective might argue that the notion of "culture" represented in these models is sparse, and not especially well-developed. As such, it may not be clear how the models relate to existing, extensive work within the fields of either anthropology or sociology. However, I suggest that even an initial movement in this direction is a significant step for survey methodologists, who are accustomed to thinking in either statistical or psychological terms, or about errors of translations. As such, "culture," as an explanatory process has not been well-represented. Further, the fact that this general error class has been regarded as central, by independent developers of the current models, suggests that it has some meaningful substance within the realm of survey questionnaire design and evaluation - even if the concept demands further development.

In particular, the detailed descriptions of problems of this type serve to distinguish them from either purely psychological effects, or errors associated with translation. Willis et al. (2008) found that questions about tar/nicotine level on cigarette brand smoked the most failed to account for the fact that Korean and Chinese brands are unlabeled in this regard, so asking about strength of cigarettes that respondents had smoked in earlier stages of their lives produced considerable difficulties for immigrants to the U.S. Similarly, Miller, Willis, Eason, Moses, & Canfield (2005) determined that dietary questions assuming the Western three-meal "breakfast-lunch-dinner" pattern do not apply to some traditional Mexicans, who eat four rather than three meals. The key factor here is of course culture, rather than the individual respondent's cognitive processes.

Additionally, Fitzgerald et al, (2009) cite an example of Cultural Issues arising in the cognitive testing of the European Social Survey (ESS), where respondents were asked to select the type of national taxation system they favored, from a list of three possibilities. They found that countries differed markedly in the degree to which their populations were generally knowledgeable about variations in tax systems, or even how the tax system functions within their own country. Hence, the question functioned dissimilarly across countries, but in ways that could not be addressed through re-translation (as if a Translation error were at root) or through direct modification of the target question that targeted individual cognitive processes.

c) Usefulness of "Generic Problem" code. A final reaction to the models proposed here is that the notion of Generic problems of question design (TCG model), or problems in the source question (Fitzgerald et al. model), adds nothing new in terms of definition or description of error, beyond previous models of questionnaire-based error. This is of course true, but also allows the newer models to encompass previous ones – by expansion as opposed to replacement, of key conceptual elements. In fact, it is significant that general problems tend to remain dominant (or at the least, well-represented), even in cross-cultural studies that also feature pronounced translation and cultural issues (see Figure 2 for an illustrative example). In some ways, the persistence of such general flaws is unsurprising, given how difficult it is to craft error-free survey questions, in any language. The fact that such challenges appear to be common across languages and cultures seems to indicate that, despite the differences that are imposed by translation and cultural variation, the process of answering survey questions in fact exhibits more commonality than differentiation. That is, problems such as question vagueness, recall difficulty, and presence of response categories that do not well match the question text, are frequent problems that appear to afflict survey respondents generally. To a great extent, problems in survey questions are simply problems, that afflict a wide swath of the inhabitants of our world.

Figure 2. Berrigan et al., (in preparation) – Percent of questions exhibiting Translation, Cultural, and Generic problems in a cognitive interviewing study of physical activity and respondent acculturation to U.S. society.

Physical Activity questions:

Translation problems:	2%
Problems of Cultural Adaptation:	8%
Generic problems:	52%

Acculturation questions:

Translation problems:	2%
Problems of Cultural Adaptation :	26%
Generic problems:	49%

Further research. Ideally, authors of cross-cultural pretesting studies could attempt to apply the revised TCG (or Fitzgerald et al.) models to their results, in order to determine (a) Whether these appear to fit their findings; (b) How difficult it proves to be to make coding decisions; (c) Whether this type of conceptualization contributes to an understanding of the nature of the problems observed; and (d) Whether the frequencies of observed problems match those of other investigators.

References

Beatty, P., Willis, G., Hunter, J., and Miller, K. (2005). Design of the Q-Bank: Determining Content, Concepts, and Standards, *Proceedings of the ASA Section on Government Statistics*, Alexandria, VA: American Statistical Association, pp.981-988.

Fitzgerald, R., Widdop, S, Grey, M., & Collins.. D. (2009). Testing for equivalence using cross-national cognitive interviewing. *Center for Comparative Social Surveys, Working Paper Series* , No. 01.

Miller, K., Willis, G., Eason, C., Moses, L., & Canfield, B. (2005). Interpreting the results of cross-cultural cognitive interviews: A mixed-method approach. *ZUMA-Nachrichten Spezial, Issue #10,* pp. 79-92.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Willis, G., Lawrence, D., Hartman, A., Kudela, M., Levin, K., & Forsyth, B. (2008). Translation of a Tobacco Survey into Spanish and Asian Languages: The Tobacco Use Supplement to the Current Population Survey. *Nicotine and Tobacco Research, 10(6),* 1075-1084.

Willis, G. B., & Lessler, J. (1999). *The BRFSS-QAS: A guide for systematically evaluating survey question wording*. Rockville, MD: Research Triangle Institute.

Willis, G., & Zahnd, E. (2007). Questionnaire Design from a Cross-Cultural Perspective: An Empirical Investigation of Koreans and Non-Koreans. *Journal of Health Care for the Poor and Underserved, 18:* 197-217.

Challenges in designing and testing questionnaires to be administered in multiple languages

Margaret Blake, NatCen

Introduction

This paper examines the issues surrounding designing and testing questionnaires in multiple languages, with a particular focus on conducting cognitive testing in multiple languages. The paper first looks at why questionnaires might need to be translated, briefly covers the links between language and culture in translation, before looking at why cognitive testing should be carried out in the languages in which the questionnaire will be administered and what this can tell us. The paper raises some of the key challenges associated with cognitive pre-testing in multiple languages and suggests some alternative approaches, based on experience at NatCen.

Why are translated survey questions needed?

Most questionnaires are initially designed in a "source" language. Questionnaires need to be translated from this source language into "target" languages when members of the population to be surveyed cannot communicate in the source language. When cross national research is being carried out, it is essential for the questionnaire to be translated into the languages of the countries involved in the research. For example, the European Social Survey (ESS) is carried out in 34 countries and involves translation into 31 languages from the source language, English. Within a single country there may be more than one official language, for example in Switzerland, German, French, Italian and Romansh are spoken. In some countries there is a legal requirement to provide documents in regional or national languages, for example the Welsh Language Act, 1993 in the UK required public bodies providing services in Wales to prepare a Welsh Language Scheme and this extends to survey documents.

In many counties there are sizeable ethnic minority populations, some of whom do not speak the main language(s) of the country well enough to participate in a survey. Since each language group forms a very small proportion of the population it is not usually practical or cost effective to provide translated questionnaires for minority groups in general population surveys. However, where the survey focuses on the needs of the ethnic minority population and may include a boost sample so that ethnic minority respondents are over-represented in the survey, translated questionnaires may be provided. In the UK ethnic boosts and translated questionnaires are included on a number of large scale surveys including the UK Household Longitudinal Survey, the Health Survey for England and the Communities Study.

The rest of this paper focuses mainly on the challenges involved in developing and testing translated questionnaires in multiple ethnic minority languages.

Decisions about translation

Once a decision has been made to offer the questionnaire and associated survey materials in languages other than the source language, the next decision is which languages should be offered. The choice of languages will be determined by the prevalence of each language speaking group in the survey population; their ability to complete the questionnaire in the source language; the availability of a written form of the language and translators to provide it; and political and equality issues. In the UK, the number and range of languages spoken by the ethnic minority population poses a challenge, together with the lack of Census data on language and the constantly changing language map. The inclusion of a language question on the UK Census is under consultation for 2011 and it seems likely that something will be included (ONS, 2007; Cabinet Office, 2008) but at present there are limited national data on language available.

In 2008 a schools census was conducted which identified the main language spoken at home for English school children. This found that 14.3% of school children in England had a first language known, or believed to be, a language other than English and among the 79% of this group who provided a named language, 240 languages were reported. Table 1 below shows the main languages spoken by at least 0.2% of children and the percentage who speak it. Of course these data do not show what percentage of the population need translated survey questions in each language but the data provide an idea of the sheer range of languages and demonstrate that there are no clear second or third languages spoken, unlike in the US where Spanish is spoken at home by 12% of the population, making it a clear second language (US Census Bureau 2007 American Community Survey).

Table 1: 2008 Schools Census: Main languages other than English for school children in England: languages spoken by at least 0.2% of the population of school children

Language	**Number**	**%**
Punjabi	102,570	1.6
Urdu	85,250	1.3
Bengali	70,320	1.1
Gujarati	40,880	0.6
Somali	32,030	0.5
Polish	26,840	0.4
Arabic	25,800	0.4
Portuguese	16,560	0.3
Turkish	16,460	0.3
Tamil	15,460	0.2
French	15,310	0.2
Yoruba	13,920	0.2
Chinese	13,380	0.2
Spanish	10,000	0.2
Base	*6,549,300*	*100*

The languages offered on NatCen surveys vary and are changing as the characteristics of the UK population changes. For example, a new survey which started in 2009 (UK Longitudinal Household Survey) involved an ethnic boost and offered Urdu, Bengali, Punjabi (both Gurmukhi and Urdu scripts), Gujarati, Arabic (Egyptian), Cantonese (simplified), Somali (Latin script) and Welsh. French, Hindi, Polish and Tamil were considered but not provided. Somali and Arabic are new languages which are being introduced when translations are carried out. Five years ago only South Asian languages and Chinese were offered on NatCen surveys (for example on the 2004 Health Survey for England (Craig et al, 2006)).

The choice of languages is further complicated by the fact that among the most prevalent minority languages in the UK, there are complex dialect and script issues. For example, Punjabi is spoken by immigrants from Pakistan who mainly read the Urdu script and by immigrants from India who mainly read the Gurmukhi script. This means that the one language has to be provided in two scripts. Most Bengali speakers in the UK do not speak formal Bengali but a regional dialect called Sylheti. Sylheti, however, has no modern written form so surveys are often translated into formal Bengali and then translated into the Sylheti dialect during the interview. An alternative approach, taken by Hunt and Bhopal (2004) was to use phonetic Latin script for a Sylheti translation.

Political and cultural sensitivities also need to be considered. For example, if translation is being provided in a group of languages such as the South Asian languages it may not be acceptable to exclude a particular language such as Hindi, even if few members of that language group will choose a translated questionnaire. Translation into African languages is often difficult given the wide range of languages and dialects and lack of written forms. Furthermore, among African migrants English may be favoured as it is more neutral and has fewer tribal or religious associations (McManus et al, 2006).

Translation and cultural issues

Once a decision has been made about which languages to cover, translation of survey questions, and documents into target languages, is not simply a matter of translating the words. The concepts intended in the source language must be translated (Agans et al, 2006; Hunt and Bhopal, 2004; Blais and Gidengil, E, 1993; Carlson, 2000). As Hunt and Bhopal (2004) wrote “If the data are to be used to make comparisons between groups then the questions must be conceptually and functionally equivalent and salient for all the groups compared”. This involves translating the question as a whole and translating the intended meaning. However, beyond this, it also involves considering whether the concepts included in the question can be translated at all.

Concepts can be divided into two groups: emic and etic (Warnecke et al, 1997; Hunt and Bhopal, 2004). Emic concepts are those which are salient and meaningful only in certain cultures (such as Chinese cultural traditions of filial piety and maintaining social harmony). Etic concepts are those which are universal and found across all or most cultures (such as the welfare of children or the concept of reciprocity). When translating questionnaires, consideration needs to be given to whether the question will be salient and meaningful in the target language. Some languages may be spoken by people from a range of cultural, religious and national backgrounds so the answer may be that the concept will translate for some groups but not others. Where a concept does not translate a decision needs to be made about whether to ask an alternative question for different language groups which is more appropriate. However, for large scale government funded surveys this may not be a feasible option since comparison between ethnic minority groups is difficult where the questions are different. In practice, standard questions are usually translated with minor modifications such as the inclusion of alternative examples or answer categories to clarify meaning (e.g. including ghee in a question about fat consumption and chewing paan in a question about tobacco use in the Health Survey for England 2004) (Craig et al, 2006).

As in many countries, the practice for UK government funded surveys is for the questionnaire to be developed, tested and finalised in the source language (British English) before translations are made into the target languages. This is largely for practical and cost reasons since translation is expensive and time-consuming. An alternative approach, however, is to develop questions in the different languages in parallel so there is no source language. For surveys with ten or more languages this is unlikely to be practical but it can be effective where only two languages are involved, as was demonstrated in the presentation about the development of the Welsh language census in the UK (Wallis, 2009).

Beyond the issues of whether words and concepts can be translated into a certain language, researchers have also questioned whether the survey process is culturally transferable (Agans et al, 2006). Even where the survey process is culturally transferable there is evidence that some forms of bias are more prevalent among certain groups. For example, Agans et al (2006) report on research showing that Mexican immigrants in the US tend to underreport sensitive behaviours and may be more susceptible to social desirability effects. Warnecke et al (1997) argue that Asian respondents avoid extreme responses. These differences can affect the comparability of survey responses, however well the questionnaire has been translated. Researchers have also raised the question of whether cognitive testing can be carried out effectively in all cultures, but conclude the method is culturally transferable if consideration is given to cultural differences (Willis et al, 2005; Agans, 2006; Willis and Zahnd, 2007)

Cognitive testing of translated questions

Assuming that cognitive testing, as a method, is cross culturally transferable, cognitive testing of translated questionnaires is desirable for a number of reasons. Importantly, cognitive testing can be used to identify whether the same concept is being measured in the source and target languages by exploring how respondents interpret and go about answering the survey question. It is in this way that cognitive testing in target languages is most useful. Alternative methods such as pilots or testing in the source language cannot explore whether the same concept is being measured in the target languages. The process can also explore whether inconsistencies in the concept being measured relate to the fact that 1) the concept does not transfer across cultures, or whether 2) the concept has not been appropriately translated. A number of researchers have developed classifications of problems with translated survey questions. Fitzgerald et al (forthcoming) and Levin et al (2009), both identify four main sources of error: 1) problems with the source language version of the question in terms of general design, 2) translation problems, 3) the characteristics of the source language questionnaire impeding translation, and 4) cultural issues. All of these error types can contribute to the target language question not covering the same concept as the source language question. The ways in which cognitive testing can identify these types of problem are explored below.

Cognitive testing in the target languages has been found to be useful in identifying problems with the source language questionnaire in terms of question wording or design problems. Experience reported in the literature suggests that in a first round of cognitive testing in translation, a substantial proportion of the problems identified relate to general design issues. In testing a Spanish language dietary questionnaire, Levin et al (2009) found in the first round of cognitive testing that 75% of the problems identified related to general design issues. Ambiguities with the source questionnaire may become apparent when respondents attempt to use a translated version of the question. There may be an overlap between poor design in general and characteristics which impede translation. Brislin (1973, 1980 cited in Behling and Law, 2000) has suggested a check list of 12 issues to consider in designing a questionnaire which lends itself to successful translation. The checklist includes keeping questions short and simple, avoiding both idioms the passive and subjunctive. Many of these principles are also useful when designing a questionnaire to be used in just one language.

Cognitive testing can identify translation errors which even committee approaches to translation can miss, by including a larger number of individuals in testing the questions (Levin et al, 2009). The literature provides examples of translation errors which slipped through even the most rigorous processes and cognitive testing provides an additional opportunity to pick up on these (Willis et al, 2005; Levin et al, 2009). Translation problems can arise from three main sources, which can all be picked up by cognitive testing: 1) human error in translation, 2) translation which is correct but is sub-optimal and 3) regional or cultural variation in language.

Cognitive testing allows questions to be tested in the survey context. This is also possible during pilot testing but survey timetables allow limited time for changes to question wording after pilots. The process of cognitive testing also provides a greater insight into why the questions do not work well in the target languages than is possible from pilot testing. After cognitive testing, changes can be made to target language questions and potentially to the source language questions.

Cognitive testing also allows the questions to be tested with the target population for the survey. Translators tend to be more educated and accultured within the dominant culture than respondents to surveys who do not speak the source language, which affects how they interpret and understand survey questions (Harkness et al, 2004). Translators are necessarily bilingual and there is evidence that bilingual people understand language in a different way from monolingual speakers of a language (Hanna et al, 2006). Cognitive testing also offers the chance to test the questionnaire with different language subgroups. Among those who speak a particular language there may be regional variations or cultural differences. For example, Spanish speakers from different parts of Latin America use different forms of the language which can present problems for translations of survey questions intended for the Hispanic population in the United States (Levin et al, 2009). In the UK, translators of Bengali questionnaires may be from Dhaka and speak the formal version of the language, but the translation is predominantly intended for Bengali speakers of the Sylheti dialect.

Sometimes problems with the translated questionnaire relate not to semantic problems with the source questionnaire but with cultural problems in that the questionnaire covers concepts which are not culturally transferable (emic concepts). This does raise the question of whether cognitive testing in translation is needed or whether cognitive testing in the source language among the various ethnic or cultural groups to be included in the survey could be just as effective at a much lower cost since the issues are cultural rather than just linguistic.

Questions and issues raised by cognitive testing

Where cognitive testing in translation is a practical option, a number of decisions need to be made and a range of challenges faced. "Extensions to languages other than English, and especially to multiple languages, pose particular challenges in terms of staffing, analysis and interpretation of results" (Willis et al, 2005).

Firstly, where problems with the target language questionnaire are identified will it be possible to make changes only to the target language questionnaire or also to the source questionnaire? This relates to the second issue which is about the timing of cognitive testing of target language instruments. If this takes place in parallel with testing in the source language this offers greater scope for changing the source language questionnaire. This would be particularly useful where the concepts included in the source questionnaire turn out not to be culturally transferable or where the design of the source language questionnaire could be improved to aid translation. However, there are cost and timetable implications of doing this and in practice, source language questionnaires are often finalised first (Levin et al, 2009).

As described earlier, the number of languages for translation may be considerable and quite commonly ten or more languages may be included for UK surveys focusing on ethnic minority populations. Where some cognitive testing in translation is possible, a key question is how many languages should be included in this process and how those languages should be selected. It would seem sensible to include languages which represent different linguistic or cultural groups as some translation issues may be common to several languages and several languages may be spoken among the same ethnic group (for example South Asians in the UK).

Since the questionnaire in the target language is most likely to be used by those who do not speak the source language, ideally the cognitive testing should be carried out with a similar group (respondents who speak the target but not the source language). In a study testing a Korean language questionnaire, Willis and Zahnd (2007) found that cognitive testing with a monolingual Korean speaking sample uncovered more translation problems than testing with a bilingual group. However comprehension problems seemed to be uncovered more by interviews with bilingual Koreans, regardless of whether their interview was in English or Korean.

As well as the availability of suitable respondents there are also issues about the availability of suitable interviewers and researchers (Levin et al, 2009; Willis et al, 2008). Particularly where several languages are required it is very unlikely that bilingual researchers covering all the languages will be available to work on the project and frequently the project researchers are not familiar with any of the languages. Bilingual interviewers have to be available if testing is to be carried out in translation, but it should be remembered that cognitive interviewing relies on some very specific skills and experience and it cannot be assumed that any field interviewer will make a good cognitive interviewer. This does raise the question of whether using source language speaking expert cognitive interviewers, to interview respondents from the ethnic or cultural groups of interest, could be more effective than using less experienced target language speaking interviewers.

Where cognitive interviews in translation are provided, it is not just the questionnaire which needs to be translated, but also the probes to be used by interviewers. Standard translated probes will be particularly important where relatively inexperienced cognitive interviewers are involved. There is also a question about how the interviewer will record the notes of the interview. The audio-recordings will be in the target languages but the notes for analysis will need to be in the source language, requiring a process of translation during the interviewer note taking, which is another skill which not all interviewers have and again relies on experience.

In a cognitive interview lasting about an hour, it is usually only desirable to test around 20 survey questions. Thus for a long questionnaire, or survey interview, it will not be possible to cognitively test every question. Even where it is used, cognitive testing cannot be relied on as the only method of checking and evaluating the translation. Other methods such as independent checking of the questionnaire by a bilingual native speaker of the target language and field piloting are still needed.

For the reasons discussed above, cognitive testing in multiple languages can be extremely costly and time consuming. At NatCen the time and cost implications of cognitive testing in multiple languages mean that we have not yet been able to do this, even on projects where we have tested a questionnaire which will be offered in translation in the survey. This has led us to explore other options which are presented in the following section.

Alternatives to cognitive testing in translation

The literature about the experience of cognitive testing of questionnaires in translation demonstrates that many of the problems identified arise from cultural issues and poor design of the source language questionnaire in terms of the ability to translate it semantically and culturally (Levin et al, 2009; Behling and Law, 2000). This suggests that inclusion of members of the relevant ethnic, cultural and language groups early on in the research process could reduce the incidence of these types of problem in the source questionnaire prior to translation. Cultural issues can also be identified through cognitive testing of the English versions of the questionnaire with members of the relevant cultural groups. It should be remembered that among the ethnic groups of interest in the UK, English speakers will be more accultured than those who do not speak English, but cognitive testing in English does allow many of the major cultural issues to be identified prior to translation.

In the Questionnaire Development and Testing (QDT) Hub at NatCen, we have conducted focus groups with members of the populations of interest prior to developing or finalising questions in English to ensure that the concepts and terms used to describe them are meaningful to members of those groups. This informs the questions to be cognitively tested and the probes to be used. We regularly conduct cognitive testing with members of different ethnic groups in English as our purposive sample designs for cognitive testing are designed to include members of particular groups relevant to the research study (for example, members of particular ethnic or religious groups). Where relevant, respondents may be chosen to represent first and second generation migrants. Where problems with the questions related to cultural or potential language issues are identified these can be reported and tackled by changing the source language questions.

These approaches mainly tackle cultural issues. One approach to making the source language questionnaire suitable for translation, apart from following general guidelines such as those provided by Brislin, is to involve translators in a discussion about the questionnaire prior to finalising it in the source language. On a project at NatCen we convened a round table discussion involving the survey researchers, the question testing researchers, the manager of a translation agency regularly used by the organisation and four freelancer translators, representing four languages spoken by the main groups of interest for the research. This discussion was useful in identifying, among other things, where the sentence structure did not translate well in one or more languages, where the answer categories would not translate well and the distance between categories might not be the same in different languages, where particular words could not be translated and might have to be included in English or transliterated (e.g. "community"), where the words could be translated literally but where the intended meaning was culturally specific and would not be translated (for example the political "far right"), and where terms such as "involvement", "feel" and "political" could be translated but the nuance of meaning would be different in a way which could affect the concept being measured. As a result of this meeting some changes were made to the source questionnaire and notes for translators were added.

Once the questionnaire has been translated into multiple target languages, NatCen interviewers are employed to check the translations. While this is no substitute for cognitive interviewing involves using bilinguals, this approach does have the advantage that such interviewers are familiar with the types of language used by respondents and the types of questions or terms which may be less well understood by respondents.

In the future cognitive testing of target language questionnaires is something we would like to develop at NatCen. We would be interested to hear the experiences of other researchers involved in this type of research.

Conclusions

Where survey research focuses on particular ethnic or cultural groups in the population, questionnaires need to be provided in translation if those groups are to be properly represented in the research. However providing surveys in translation is a complex, time consuming and costly process, involving many decisions including choice of language(s), method of translation and method of testing or evaluating translated survey instruments. Cognitive testing offers an important means to test the translated questions and can be invaluable in testing whether the concepts covered in the target language are consistent with those in the source questionnaire. However cognitive testing may not always be feasible for a number of reasons. This paper has presented some alternative approaches (inclusion of the relevant language, ethnic or religious groups early on the questionnaire development process, testing in English with the groups of interest and discussions with translators prior to finalising the source questionnaire) which have been used at NatCen where cognitive testing has not been an option.

References

Agans, L, Deeb-Sossa, N, Kalsbeek, W D (2006) "Mexican immigrants and the use of cognitive assessment techniques in questionnaire development" in Hispanic Journal of Behavioural Sciences 2006: 28: 209.

Behling, O. & Law, K. S., (2000) Translating Questionnaires and Other Research Instruments: Problems and Solutions, *Solving Semantic Problems,* **pp.24-28**. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-131. Thousand Oaks, CA:Sage

Blais, A and Gidengil, E, (1993) "Things are not always what they seem: French- English differences and the problem of measurement equivalence" Canadian Journal of Political Science XXVi:3

Cabinet Office (2008) Helping to shape Tomorrow: The 2011 Census of Population and Housing in England and Wales: Census White Paper. Cabinet Office

Carlson, E.D (2000) " A case study in translation methodology using the health promotion lifestyle profile II" Public health nursing 2000: 17 (1):61-70.

Craig, R, Deverill,C, and Pickering,K (2006) Health Survey for England 2004: Survey Methodology and Documentation. Stationary Office.

Fitzgerald, R, Widdop, S, Gray, M, Collins D (forthcoming) "Identifying sources of error in cross-national questionnaires: application of an error source typology to cognitive interview data."

Hanna, L, Hunt, S, Bhopal, R.S (2006) "Cross-cultural adaptation of a tobacco questionnaire for Punjabi, Cantonese, Urdu and Sylheti speakers: qualitative research for better clinical practice, cessation services and research." Journal of Epidemiology and Community Health 2006: 60: 1034-1039.

Harkness, J, Pennell,B.E, and Shaua-Glusberg, A (2004) "Survey questionnaire translation and assessment." in Presser S, Rothgeb J.M, Cooper,M.P, Lessler, J.T, Martin,E, Martin,J and Singer, E (eds) Methods for testing and evaluating survey questionnaires pp453-473. Hoboken, NJ: John Wiley and Sons.

Hunt, S.H. and Bhopal. R. (2004) "Self report in clinical and epidemiological studies with non-English speakers: the challenge of language and culture." Jnl of Epidemiology and Community Health 2004: 58: 618-622.

Levin, K, Willis, G, Forsyth, B.H, Norbery, A, Kudela, M.S, Stark,D, Thompson, F.E.(2009) "Using cognitive interviews to evaluate the Spanish language translation of a dietary questionnaire" Survey Research Methods 2009: 3 (1): 13-25.

McManus,S, Erens, B, Bajekal, M (2006) "Conducting surveys among ethnic minority groups n Britain" in Nazroo, J ed Health and Social Research in Multiethnic Societies. Routledge.

Office for National Statistics (2007) 2011 Census: Ethnic group, national identify, religion and language consultation. Experts, community and special interest group responses to the 2011 Census stakeholders consultation 2006/7. National Statistics.

US census bureau 2007 Community Survey

http://factfinder.census.gov/servlet/ADPTable?_bm=y&-geo_id=01000US&-qr_name=ACS_2007_1YR_G00_DP2&-context=adp&-ds_name=ACS_2007_1YR_G00_&-tree_id=306&-_lang=en&-redoLog=false&-format=

Wallis, R (2009) Dual development of the 2009 Rehearsal England and Wales Population Census Questionnaires. Presentation to Quest Workshop, May 19th, 2009.

Warnecke, R.B., Johnson, T.P, Chavez, N, Sudman, S, O'Rourke,D.P, Lacey, L and Horm J (1997) "Improving question wording in a survey of culturally diverse populations"Ann Epidemiol 1997: 7: 334-342.

Willis, G, Lawrence, D, Thompson, F, Kudela, M, Levin, L and Miller,K (2005). " The use of cognitive interviewing to evaluate translates survey questions: lessons learned." Paper presened at the 2005 Conference of the Federal Committee on Statistical Methodology, Arlington USA.

Willis, G and Zahnd, (2007) "Questionnaire design from a cross-cultural perspective: an empirical investigation of Koreans and non-Koreans." Journal of Helath care for the poor and underserved, 2007: 18: 197-217.

Willis, G, Lawrence, D, Hartman, A, Kudela, M.S, Levin, K, Forsyth, B (2008) "Translation of a tobacco survey into Spanish and Asian languages: The Tobacco Use Supplement to the Current Population Survey" in Nicotine and Tobacco Research, 2008: 20(6):1075-1084.

Statistisk sentralbyrå
Statistics Norway

Notat

February 11, 2009

Experiences from testing public user surveys "adapted" to the immigrant population in Norway - abstract

Paper prepared for the QUEST workshop in Bergen, Norway, May 2009

By Elisabeth Gulløy, Statistics Norway

Do respondents from different cultures in Norway understand questions on their opinions regarding public sector services in a similar way? What kind of challenges do we meet when we try to adapt a standardised public user satisfaction survey to a diverse immigrant population? These questions will be discussed when we present the test results from a development project commissioned by the Directorate of Integration and Diversity (IMDI) and performed by Statistics Norway in cooperation with Institute for Labour and Social Research (FAFO).

The project’s aim is to develop guidelines for standardised public user satisfactions surveys specially adapted to measure the experiences and opinions of an increasingly diversified minority population in Norway. Such surveys are already established for monitoring public user satisfaction in the population as such. In spring 2009, we will develop a questionnaire more or less adapted, and more or less similar to, the ordinary version. This questionnaire will be tested in a row of cognitive interviews. Results from the testing will be important in developing guidelines for future surveys in this field.

Important dimensions in the testing will be

- interpretation of key constructs: public service, public servant, availability, service information, service quality, service satisfaction
- comparing key constructs in origin culture and Norwegian culture
- interpretations and cognition of common user satisfaction survey questions
- meaning of responses
- meaning of scales
- social meaning of response process

Dual-Development of the 2009 Rehearsal England and Wales Population Census Questionnaires

Ruth Wallis, Office for National Statistics

A Welsh language Household questionnaire has been used in Wales to collect population statistics in the Census since at least 1841, but, in 1993, the Welsh Language Act made the use of a Welsh language questionnaire a legal requirement. Until now, the process used for developing the Welsh language Census questionnaires has been to translate the English questionnaire into Welsh, at the end of the English questionnaire development process.

For the 2011 Census, a new method has been implemented to develop equivalent Welsh and English questionnaires. This method builds on that used for the bilingual New Zealand Population Census, using a dual-development approach to designing questions in Welsh and English. Core to this methodology is the parallel development of questions in both languages. By using this process, issues unique to each language can be given consideration at every stage of the development cycle. Both languages are given equal status throughout the development cycle, so that compromises are not made at the expense of either language, and ultimately both versions of the questionnaire should meet an equivalent quality standard. Testing so far shows that the resulting questionnaires will be easier to complete for the Welsh-speaking population, and will collect data of a higher quality than previous Welsh Census questionnaires. This paper outlines the dual-development process.

Data Quality in Cognitive Interviewing: A new perspective on standardizing probes in multi-language and multi-cultural projects

Submitted for inclusion in 2009 QUEST Conference
Bergen, Norway

Stephanie Willson, Ph.D.
National Center for Health Statistics

The strength of cognitive interviewing as a method of question evaluation lies in the ability of the analyst to examine the construct validity of survey items. This is accomplished using in-depth interviews whereby interviewers explore how respondents interpret and answer survey questions. The advantage of the method is the flexibility that allows interviewers to explore issues as they arise during data collection. However, data quality as well as the ability to compare and summarize findings across interviews conducted by multiple interviewers in different languages among culturally distinct groups requires some degree of standardization. The question is how to achieve standardization without losing the benefits of the qualitative method.

Most efforts at achieving standardization have focused on how to word probes (i.e., follow-up questions). Informed by quantitative methods, this approach suggests that in order to obtain conceptually specific and comparable data from respondents, all must receive exactly the same follow-up questions. It's believed that imposing this level of control over the interview process improves cognitive interview data quality by minimizing the effects of interviewer skill level and maximizing the chances that specific topics will be covered.

This paper is an examination and evaluation of probes, with special emphasis on how they may be used to collect standardized and comparable data in cross-national or cross-cultural projects. Using data from cognitive interview evaluation studies conducted by the Questionnaire Design Research Laboratory at the National Center for Health Statistics, this paper will show that there are two different styles of probing techniques, each motivated by competing assumptions about what respondents know and how they are able to report it. The first approach assumes that respondents use logic and reason to understand survey questions. The second assumes that respondents base their understandings of questions on personal experiences. This paper will argue that the key to standardization lies in implementing a consistent approach to probes, shifting attention away from the specific wording of probes to acknowledging the underlying assumptions that probes make about how respondents are able to answer questions. Furthermore, the paper will demonstrate that the quality of interview data obtained by probes that assume an experiential epistemology is higher than the quality of data obtained by probes assuming a logic-based epistemology.

A proposal of best practices for conducting cognitive interviewing in cross-cultural/national surveys

José L. Padilla

University of Granada (Spain)

Due to the growing number of surveys in which different cultural and linguistic groups are involved, several organizations and research groups have recently tried to provide information on best practices or guidelines, across the multiple phases of cross-cultural/national survey. Among the most significant achievements, it is worth mentioning the cooperative Survey Design and Implementation (CSDI) Guideline. The CSDI Guidelines take care of the various organizational and operational aspects that should be considered in the structural design of a cross-national project. Focusing on the cognitive interviewing, the purpose of this paper is to propose detailed best practices for conducting cognitive interviewing in cross-cultural/national surveys. Together with a review of the "state-of-art" in the field, the proposal is mainly based on the "lessons learned" from the multi-national testing project conducted by the Comparative Cognitive Testing Workgroup. The project was coordinated by Kristen Miller, representing the Budapest Initiative, and Rory Fitzgerald from the European Social Survey. The paper will intend to stimulate discussions among QUEST meeting participants on the aims of the best practices (to provide criteria for evaluating cognitive interviewing or to guide cognitive interviewing designs); how to organize best practices in meaningful categories; how to describe best practices, etc. Finally, the design of a survey on Internet intended to get expert comments on the strengths and weaknesses of the proposal will be presented.

Qualitative testing: How to manage collected data

Karen Blanke and Sabine Sattelberger (Federal Statistical Office, Germany)

1. Background

For several years, the Federal Statistical Office (FSO) has been working on the systematic implementation of questionnaire testing. Marking an important step within this development, its own pretest laboratory[1] was established at the end of 2007. A year later, an eye-tracker was added that enables to follow eye-movements of test persons on the computer screen. Questionnaires of paper-and-pencil as well as online surveys of German official statistics are now increasingly evaluated by qualitative testing methods.[2]

In the long run, the aim is to reduce the burden for respondents and to increase data quality. However, while pretesting certainly is a challenge, this also holds for organizing the overall evaluation process as systematically as possible. Therefore, it is intended to develop efficient standards for questionnaire testing and to distribute tasks among several colleagues.

This paper gives an insight into work in progress. The FSO's strategies of analysis are put up for discussion.

2. Qualitative testing: three sources of information

To the majority of self-administered questionnaires, the FSO applies a three step approach in qualitative testing. Consequently, three different sources of information are to be analyzed in the end – this is done for each of the (usually 15-20) test persons individually.

Step 1): Observation during self-completion

The test person is observed while filling in a questionnaire on his/her own in the pretest laboratory ("reality without words"). Concurrently, the observer next door notes peculiarities – like shaking of the head, laughing, complaining and moving backward and forward through the form. These observational notes are directly handed on to the cognitive interviewer to be shortly discussed before the following interview. So the cognitive interviewer knows in advance which of the questionnaire pages seem to be particularly difficult for the test person. The observation protocol sometimes gives important hints on problems with the questionnaire which had not been anticipated as the potential respondents look differently at the instrument than experts.
The observation can be watched again and again, as it is video-taped.[3] However, the analysis is rather time-consuming and it is often difficult to decide what is worthy to note.[4]

Step 2): Cognitive interview after self-completion

After the self-completion of the questionnaire, the test person is interviewed, using a cognitive testing protocol. The utilized techniques are mainly probing, think-aloud and sorting, when it comes to difficult terms. At the beginning of the interview, the test person is invited to describe briefly his/her personal situation in a narrative

[1] The laboratory looks like an ordinary conference room. This is on purpose to create a comfortable atmosphere (thus, the FSO rejected a "window" connecting the two rooms). Cameras and microphones are hardly noticeable on the ceiling and controlled in the observation room located next to the laboratory.

[2] This paper concentrates on the qualitative testing of paper-and-pencil-questionnaires for household surveys as the FSO has only just started to implement usability-testing of websites and online questionnaires.

[3] Test persons are asked for their agreement to be video-taped. They receive €30 for participation and transportation costs.

[4] So far social sciences have hardly paid any attention to standards for the analysis of video recordings. It is even more difficult to find related research applicable to the context of official statistics. For a first approach see: H. Knoblauch; B. Schnettler; J. Raab; H.-G. Soeffner (eds.) (2006): Video Analysis. Methodology and Methods. Qualitative Audiovisual Data Analysis in Sociology. Frankfurt am Main: Peter Lang Verlag.

manner. This gives an interesting insight into the individual reality that has to be transferred into the grid of a questionnaire. The interview is video-taped so that the interviewer needs to write down only keywords.
The cognitive interview is regarded as the step to uncovering what respondents have understood in theory and after some consideration. The test person gives, for instance, reasons for misunderstanding instructions, skips or technical terms. Besides, the retrieval of information is often complicated and causes incorrect or imprecise entries. However, much of the success of this step depends on the communication skills of the test person that are often, but not always connected to the level of education.
The cognitive interviewer has to be well-trained and highly experienced, as cognitive techniques should be applied reasonably (e. g. probing) – after all, it is just a cognitive interview, not an interrogation.
You have to be aware that the results of the cognitive interview are sometimes misleading: For example, some test persons succeed in paraphrasing a definition but do not fill in the questionnaire accordingly because they do not reflect on the issue or they do not care about instructions. In addition, answers during the cognitive interview may correspond neither with the completed questionnaire nor with the observation.

Step 3): Evaluation of the self-completed questionnaire

The interviewer and the test person already look at the entries in the questionnaire together during the interview. What follows directly after the interview is a thorough evaluation of whether the form was completed correctly, taking into consideration the results of the observation, the "warming-up" at the beginning of the interview (narrative description of the personal situation) and the cognitive protocol.
This step deals with the actual handling of the questionnaire beyond what respondents thought they had understood. Quite often difficulties are discovered that remained unnoticed during the interview or assumptions concerning problems with the questionnaire can be verified by the additional information available.
The analysis of the questionnaire is rather time-consuming and some entries are hard to follow without further explanation on the part of the test person, especially if they do not correspond with the answers given during the interview.

To sum up, combining these three sources of information for each test person (observation, cognitive interview and self-completed questionnaire) provides useful qualitative data that should be analyzed and stored in an efficient way.

3. How to manage collected data

It is definitely a major challenge to organize qualitative data as systematically and as effectively as possible. This is due to the huge amount of collected data as well as to the fact that the data exist for the most part just on tape, which means they are not available in written form at first.

The FSO seeks to develop efficient ways of dealing with qualitative data by pursuing five main objectives:

- less expenditure of time for analyzing the observation,
- organizing teamwork,
- structuring and storing qualitative data,
- obtaining a fast overview of results, and
- providing a less subjective interpretation.

Less expenditure of time for analysing the observation

The FSO observes the process of filling in a self-administered questionnaire for all test persons using a computer-based tool programmed specially for this purpose.[5] One of the two cameras is focused on the test person, the other one is focused on the questionnaire. Both views can be seen at the same time in one video-recording.
The video recording can be structured by adding so called "events". With the help of these "markers", either special events occurring during the completion of the questionnaire (e.g. moving backward and forward through the form) or specific sections within the questionnaire can be defined. When markers have been set, the

[5] The computer-based tool is called "tsmlogger".

analyzers do not have to watch the whole recording again later; they can directly skip to the event they are interested in.
Standard events are usually predefined. The test person moves through the questionnaire and the observer follows him/her by defining events for every page – beginning with the cover sheet of the questionnaire and ending with the final page with the legal basis. Furthermore, particular events can be added for each test person, for example "test person sighs" or "test person complains".

Organizing teamwork

Pretesting implies teamwork in order to provide findings soon. Teamwork has to be organized well to go well. At the moment the FSO's pretest team consists of two social scientists who conduct cognitive interviews and analyze them. Two other colleagues support them by coding the cognitive interview concurrently. This is done with a computer-based analysis tool.[6]
This division of work makes the subsequent analysis a lot easier and interpretation starts directly after the cognitive interviews. The coding process may be supervised: Different colleagues can work within the analysis tool one after another and the software reports the name of the respective coder.

Structuring and storing qualitative data

Due to lack of time and human resources, the FSO does not transcript whole cognitive interviews. During the preparation phase of the pretest, a complete code system is created based on the structure of the cognitive testing protocol, so answers of probes for all test persons are coded in the end. Within the code system, you finally have an overview of the number of cases per code. Predefined codes can be rearranged after the interviews as well, if necessary.
When the cognitive interview starts, the coder already has a complete code system to work with on his/her hands. During the interview, the coder observes the situation and assigns codes from the predefined code system by drag and drop to the cognitive testing protocol. Moreover, the coder can attach memos in case of open answers to probes with no fixed response categories. Only in these cases, the coder has to write short transcriptions.
After each pretest, there exists a complete electronically stored project with all coded cognitive interview protocols, one for each test person.

Fast overview of results and less subjective interpretation

The feeling might be familiar that first general impressions on the part of the analyzers directly after the cognitive interviews can sometimes be misleading. To be as objective as possible, the cognitive interviewers of the FSO exchange experience and go for a collective interpretation.
In order to obtain a fast overview of the results, the analyzers have a look at tables containing all assigned codes and the frequencies of coded answers for the test persons. These tables can be printed or exported in order to continue working with them (e. g. for integration into the final report).
Additionally, there are tables with all memos that can be printed or exported as well. This means that you can have a closer look at the short transcriptions for open answers of all test persons to one particular question during the interview. The next step is to assign codes to these memos in order to further reduce the data.

4. Conclusions

Analyzing video-taped qualitative tests (observation and cognitive interview for each test person) is highly time-consuming. Transcribing cognitive interviews completely is impossible due to lack of time and human resources. However, in order to avoid the risk of drawing final conclusions from first general impressions, the FSO uses a three step approach by means of computer-based tools.

Apart from organizing team work in a more efficient way, the underlying idea is to structure and to store data permanently, to accelerate analysis and to simplify interpretation. On the whole, the FSO set great store by the results of qualitative testing following objective criteria and being verifiable by others.

To sum up, these are – in our opinion – the main advantages of our approach:

[6] The computer-based tool is called "MaxQDA".

- The main findings of the cognitive interviews are available directly after the cognitive interviews, as another colleague deals with coding concurrently.
- Overall transcription is limited to open questions in the cognitive testing protocol. In these cases, memos are included in the predefined coding system, which means that the coder writes short summaries of the open answers. Finally, codes are assigned to these memos in order to further reduce the complexity of data.
- Analyzers – who also conduct the cognitive interviews – have a closer look at frequencies of assigned codes in order to base their interpretation upon more objective criteria. Results can be verified by others.
- Pretesting implies teamwork. Different tasks are distributed among several colleagues –interviewers and analyzers on the one hand, observers and coders on the other hand.

However, we are also confronted with some disadvantages:

- Reducing qualitative data so much seems to be transforming it into purely quantitative data. Does this process go too far?
- While it is less time-consuming for the researcher not to code the observation and the interview by him-/herself, it is necessary to supervise the work of coders. Moreover, some information is certainly lost when this task is delegated.

Further exchange of experience between members of the QUEST group shall be initiated with regard to the benefits and constraints of managing qualitative data electronically.

Q-Notes: Development and use of analysis software for cognitive interviews to examine survey question comparability

Kristen Miller, Ph.D.
National Center for Health Statistics

Recent work has shown that cognitive interviewing studies can provide essential information regarding the comparability of survey questions, specifically, how respondents interpret and process questions and whether particular sub-populations or groups may process questions differently from others. To achieve this goal, however, studies must be based on empirical evidence and systematically analyzed across interviews and sub-populations—a process which can yield a massive amount of qualitative data across numerous countries and in multiple languages. To be sure, one of the biggest challenges for comparative, multinational cognitive testing is data management, that is, the organization and reduction of cognitive interview data such that it can be analyzed systematically. This paper will describe software that was specifically developed to analyze cognitive interviews in this capacity. To illustrate the software's use, the paper will draw heavily from the Washington Group's testing project to evaluate an extended set of disability questions in a global context.