

Cognitive Interviewing as a Tool for Improving Data Quality in Surveys: Experiences in Istat.

Sebastiani G., Tinto A., Battisti A., De Palma E.¹

Abstract In recent years, the growing awareness of communication and interpretation distortions that could be produced during an interview, raised attention on *pre-testing* techniques aimed at improving the survey instrument and, consequently, data quality. The purpose of this paper is to present the potential of cognitive testing techniques for the identification of errors in the first phases of questions development. This paper will consist of two main sections: 1) description of main characteristics of cognitive testing methods, focusing on choices to make and pros and cons of each choice; 2) description of two experiences of pre-testing of questionnaires carried out in Istat using cognitive testing.

Key words: Measurement errors, data quality, cognitive testing, pre-testing methods, question evaluation.

1 Cognitive Interviewing

In recent years, the growing awareness of communication and interpretation distortions that could be produced during an interview, raised attention on pre-testing techniques aimed at improving the survey instrument and, consequently, data quality.

In particular, cognitive interviewing has emerged as one of the most prominent *pre-testing* methods for identifying and correcting problems with survey questions.

¹

Gabriella Sebastiani, ISTAT, email: gabriella.sebastiani@istat.it

Alessandra Tinto, ISTAT, email: tinto@istat.it

Alessandra Battisti, ISTAT, email: albattis@istat.it

Elena De Palma, ISTAT, email: depalma@istat.it

The interest for cognitive testing came, at the beginning of the 80s, from an interdisciplinary seminar on *Cognitive Aspects of Survey Methodology* (CASM) [6]. Although its initial paradigm was quite specific and explicitly related to psychology, in its following development it became broader. Beatty [1] proposed a definition that reflects its most common application: cognitive interviewing entails administering draft survey questions collecting at the same time verbal information about the questions and the survey responses. The information is then used to determine if the question is generating what authors intended.

The main purposes of a cognitive test can be outlined referring to the phases of the cognitive process that is followed during an interview, as described by Tourangeau [8]: comprehension of the questions, retrieval from memory of relevant information, judgement/estimation process and response process. For every phase it is then possible to identify the specific aims that can be pursued through this kind of test:

1. *Comprehension of the questions*: to test whether respondents understand the question as intended by the researcher who designed it. In particular the degree of difficulty in answering needs to be evaluated, considering if sentences and terms used are adequately understood and if they describe sufficiently the object/event they refer to or if they could lead to different interpretations.
2. *Retrieval from memory of relevant information*: to judge the recallability of the information needed in order to answer, checking the strategy used by respondents to retrieve information and evaluating if respondents are asked to provide information they forgot or they never had.
3. *Judgement/estimation process*: to evaluate if respondents devoted sufficient mental effort in answering the questions and to assess the degree of interest in the subject of the survey. To check if the question wording could lead, for example, to positive or negative answers (social desirability). To identify questions that could cause embarrassment or defensive behaviours.
4. *Response process*: to test if respondents correctly interpret the answer categories provided, and if they match with the spontaneously generated answers to those questions.

Given the general rule that implies submitting survey questions to participants and collecting verbal information on questions and answers, in practice cognitive interviewing can follow two main paradigms. One based on encouraging participants to verbalize thoughts as much as possible while answering questions (*think aloud procedure*). The other involving a more proactive role of interviewers, who ask specific additional follow-up questions, generally called *verbal probes* [7, 9].

The choice of the approach to adopt leads to relevant differences in every phase of the testing process. Cognitive interviews can be based on a written protocol, can be semi-structured or totally open and based solely on issues coming up during each interview. Also the strategy of analysis of results coming from cognitive interviews can vary, going from cases where it is based on systematic analysis of scripted interviews, to cases where it is based only on notes taken during the interview.

Furthermore, the expected role of interviewers can vary according to the approach adopted. When the protocol is less structured, either when using the *think-aloud* method or when adopting less standardised probe questions, it is necessary to assign the job to interviewers who are experts in conducting this kind of interviews.

Nowadays cognitive testing is widely used in several research institutes around the world, however, definitions and practices used vary notably. Some authors [3] consider cognitive testing as synonym of *think-aloud*, others believe that cognitive interviewing

entails a mixture of *think-aloud* and *verbal probing* techniques [9], others define cognitive interviews mainly in terms of *verbal probing* [2].

When choosing to adopt, even if only partially, verbal probes, it must be noted that two main factors rule the development and use of verbal probes: the nature of probe construction (standardised probes vs. spontaneous or emergent probes) and the conditions for probe administration (proactive vs. reactive probes).

Concerning the nature of probe construction, the main choice to make is whether probes should be specifically planned prior to interviews or if they could be made up during the course of the interview. In the first case we talk about standardised or anticipated probes, which generally refer to issues identified by researchers prior to the testing phase and which are used in a standardised fashion by interviewers. For example, when researchers doubt that a specific term used could be unclear or difficult to understand, specific probes could be aimed at shedding light on the issue. A probe of this type could be, for instance, “What does ‘learning opportunity’ mean to you?”. In other cases, the aim of the test could be also evaluating the formulation of questions, by using anticipated probes that ask respondents to repeat in their own words the questions (*paraphrasing*). Another method used to identify problematic questions is to ask respondents to assess the degree of reliability of their answers (*confidence ratings*) [5].

On the other hand, spontaneous or emergent probes are made up during the interview by each interviewer, according to his/hers evaluation of the appropriateness of going in more depth for issues which emerged in the particular interview. Interviewers must be able to make up these questions using an unambiguous language, which could not possibly lead to distortions. It is quite clear that the choice to make is tightly linked to the ongoing debate about the style to be adopted by interviewers in cognitive testing. In particular, that can range from a more standardised approach, which allows to reduce as much as possible the interviewer effect being able to compare outputs across interviewers and across different studies, to a more flexible and spontaneous approach, which gives more emphasis to the need of creating a “conversation” between interviewer and participant. Naturally this second approach requires particularly skilled interviewers.

The second important decision to be made is whether to opt for a proactive probing, based on interviewer initiative, or reactive probing, according to the behaviour of the subject interviewed. The first approach is based on the assumption that by restricting probing only to cases in which clear difficulties emerge, we could miss problems that exist below the surface. The second approach, instead, is extremely useful to identify unanticipated problems. In practice very often a mixture of these approaches is the most productive way.

2 Experiences in Istat

We are going to describe two experiences, which were both carried out in Istat within international projects. These experiences were carried out in collaboration with expert working partners, who work specifically at the development and improvement of these methods in their institutes (for example the NCHS – National Center for Health Statistics in the USA).

2.1 *The Cognitive Test Carried out in ISTAT within the Work of the Task Force WHO/Eurostat/UNECE*

In 2004, with the aim of developing, within Official Statistics, common tools for measuring and comparing statistics on the health state of the world population, a joint WHO/Eurostat/UNECE Task Force (TF) was set up. Among the activities of this TF a cognitive test was carried out in ISTAT in 2006, on a draft set of questions, identified by the TF itself. The questions covered 8 domains: Vision, Hearing, Mobility, Cognition, Affect, Pain, Fatigue and Social Relationship.

Cognitive interviews were carried out in parallel in Australia (25 interviews), Canada (22 interviews), Italy (18 interviews) and the United States (40 interviews).

The approach adopted was a mix of think-aloud and verbal probing based on a common testing protocol developed by the TF, in which pre-scripted general probes were included (such as “How did you come up with this answer?” or “What were you thinking?”). However it was left to interviewers to integrate and fill in with spontaneous probes where the scripted ones were not sufficient to pursue the aims that wanted to be reached, eliciting respondents to produce narratives.

In particular, the TF intended to evaluate the general feasibility of operationalisation of domains such as “Cognition”, “Affect” and “Social Relationship”, for which it was well known the measuring difficulty. Moreover, the TF identified specific overarching issues that were to be explored for all of the questions tested. In particular:

1. *Answer categories*: to assess how answer categories proposed fit the different levels experienced by respondents, each question was first administered in an open fashion, and then administered again proposing a set of categories for respondents to choose from. Two alternative sets were administered. This allowed to evaluate the match between spontaneous and chosen answers, and at the same time to evaluate the performance of the two proposed sets.
2. *Recall period*: a specific objective was to judge if respondents answered considering strictly the 4 weeks period mentioned in the questions, and if they answered, as requested, averaging the situation experienced over the period.
3. *Distances*: some of the questions made reference to specific distances (for example 4 meters for vision, 100 and 500 meters for mobility). The test aimed at evaluating if respondents knew how long those distances were and if they actually referred specifically to them when answering the questions.
4. *Use of aids*: the test also aimed at identifying if respondents accurately followed instructions given in the questions concerning the use of aids.

Finally, given the peculiar nature of this experience, with parallel tests in four different countries, there was a special attention to cross-cultural issues and to difficulties in understanding due to problems encountered in the translation of question from the source language (English) to Italian and to French.

In Italy each interview was conducted by one interviewer and one observer. The group of respondents was selected to have a variety of different situations, according to age, level of education and type of health problem/disability. After a first round of 10 interviews, interviewers and researchers met to discuss about evident problems, and after a few changes, the remaining 8 interviews were carried out.

The cognitive interviews in the 4 testing sites allowed, first of all to give an answer to the issues identified by researchers prior to the test:

1. *Answer categories*: researchers who developed the module set up the priority of having the same answer categories for all questions. However, the test showed that answer categories did not work in the same way for every question. Moreover in several cases respondents had difficulties distinguishing among the middle categories. This led to the conclusion that the initial priority had to be abandoned and that, for most of the questions, it was better to go for 4 answer categories instead of 5.
2. *Recall period*: the test showed that respondents did not consistently refer to the period of 4 weeks given in the questions. Some of them answered in general, not thinking of any specific time frame, others referring to a different period which was more appropriate to describe their situation. In some other cases, the explicit reference to the previous 4 weeks, made respondents consider only specific episodes occurred in that period, instead of making an average over the period. According to these results it was decided to restrict the recall period to 1 week, and to make it explicit in the question only for those areas showing more evident time variability, such as “Pain” and “Affect”.
3. *Distances*: respondents seemed to have difficulties figuring out and taking into account the distances proposed in some of the questions. These difficulties varied in the different testing sites, showing to be culture-dependent. The TF agreed that it was important to facilitate the job of respondents, providing culturally appropriate examples of the distances considered.
4. *Use of aids*: overall, question clauses specifying when the use of aids had or had not to be considered seemed to work quite well.

Testing results allowed also evaluating the set of questions proposed to measure domains of difficult operationalisation. In particular, cognitive tests results were not sufficient to be able to finalise satisfactory questions on three aspects of health: “thinking clearly”, “fatigue” and “social relationship”. The TF agreed that more work will be necessary in these domains.

Moreover, through cognitive testing it was possible to identify further problems with some questions, which were not anticipated by researchers. In particular, as far as the interpretation of question meaning is concerned, in a question aimed at measuring vision ability from far, testing results showed that the expression used “recognising the face of someone 4 meters away” was often misunderstood as cognitive ability to recognise more than visual ability. The test allowed then to improve the wording of the question referring to the concept of “clearly seeing someone’s face”. In the mobility domain, the TF intended to test the feasibility of measuring the concept of “*moving around/getting out and about*”. Results showed that, besides creating problems for translation into other languages, this formulation of the concept was not clear for several respondents, who referred to a wide variety of different situations, going from “using public transport” to little movements such as “combing one’s hair” or “getting out of bed”. In the cognition domain one question referred to difficulty in “concentrating or remembering things”. In all testing sites, with the exception of the US, having two concept in the same question created some problem to respondents who, sometimes, wanted to give two different answers in relation to the two activities. This, in some cases, resulted in the fact that the effective response referred only to one of the two activities. Moreover, the range of different memory problems mentioned was very wide, going from very little to more severe problems. The TF agreed to focus only on memory for the moment.

The main output of the work of testing carried out by the TF was the definition of a module, called “Budapest Initiative – Mark 1”, that was submitted to Eurostat for inclusion in the European Health Interview Survey in February 2007.

2.2 The Cognitive Test Carried out in ISTAT within the Project “Disability and Social Integration Module - EDSIM”

The Italian National Institute of Statistics, in 2008, carried out cognitive interviews in order to test the draft survey questions on disability and social integration, developed among the activities planned by Eurostat for the implementation of the “European Survey Module on Disability and Social Integration”. The questions were designed by the Department of Health Sciences at Leicester University and similar tests were undertaken first in the UK and, in parallel with the Italian test, in Lithuania.

The main goals of the cognitive test were, besides the general aim of testing respondents understanding of the concepts underlying each question, to evaluate the comprehension of specific terms identified by the researchers, as well as the appropriateness of the existing response options and to check the accuracy of the translation of questions from English.

The testing protocol was based on verbal probing technique with pre-scripted general probes and other probes that each interviewer was free to use if further investigation was needed.

Respondents were selected in order to interview a variety of people, according to sex, age, level of education and health conditions. A total of 30 respondents were recruited in Rome, 22 of which reported a disability or a longstanding health problem. Respondents with disability were selected through contacts with organisations working in the field of disability; non-disabled respondents were contacted through researchers’ networks. Interviews lasted 1 hour on average. Each interview was carried out by an interviewer and an observer who took notes and observed the respondent’s behaviour. All interviews were recorded by a digital recorder.

After each interview the person who took notes wrote them down on a question-by-question basis using an electronic form of the survey questionnaire and listened to the audio file integrating the notes. The interviewer then added in his comments and listened to the taped interview for any doubt.

The cognitive test pointed out different problems [4], grouped as follow:

1. *Questions with a complex structure.* Some questions generated confusion among respondents and sometimes seemed to induce respondents to answer thinking in general, or maybe referring to difficulties reported by other people, instead of referring to their personal experience. This is the case, for example, of a question in which respondents were requested to answer about difficulties getting to, getting into, going through and using facilities in the buildings. Other questions seemed long and respondents showed signs of tiredness. For example questions in which respondents had to express the level of satisfaction for a list of leisure activities. In addition the activities listed in the question were often considered non-homogeneous among them. In both cases several respondents asked to repeat the questions as well as the answer categories.
2. *Questions with terms which were difficult to understand.* Some questions included words that were difficult to understand for respondents, for example

‘Do you ever find that you have difficulty using the facilities in the building? Very often, quite often, occasionally, not at all, bed bound’. The word “facilities” was misunderstood by several respondents, who were not sure what to include or exclude in their answers.

3. *Questions which were difficult to understand because their scope was not clear to respondents.* It appeared in general quite difficult to ask questions about cognitive abilities to people with no cognitive difficulties. The majority of respondents seemed to struggle getting a frame of reference for the set of questions. For questions asking about the capacity to read, to understand, to write and count, respondents gave a wide range of interpretations of what they thought the purpose of these questions was. The most serious problems were encountered with the questions on reading and counting, for which very few people with no cognitive difficulties managed to understand the scope.
4. *Problems with answer categories.* Sometimes the answer categories did not seem to be satisfactory for respondents, notably when there were dichotomic items such as yes/no. In some cases respondents said that they would prefer to have more options as answering items, while in others they said that options in the answer categories were missed and proposed to add.

This cognitive testing experience contributed to understanding some differences in the answering process of people with health problems and people without health problems. For example the level of satisfaction for activities done in the leisure time was with any doubt influenced by personal expectations, which seemed to be lower for people with health problems than people without problems. The first group, indeed, answered questions taking into account only the activities that they already know to be able to do and had no consideration for activities they would like to do.

Based on the Italian and Lithuanian cognitive test results several changes were made to the module (i.e. questions on cognitive abilities were dropped, the order of questions was modified), to the wording of some questions (i.e. by changing words or taking out sentences in the introductory part of the questions) and to the answer categories (using frequency scales, when appropriate, or adding new answer items).

3 Discussion and Conclusions

The cognitive testing method has potential limitations because of specific characteristics inherent to the methodology, such as small sample size, non-representative sample, non-standardized interviewing protocol, requirement of trained interviewers, lack of consensus on methods of analysis, lack of standardized criteria for what constitutes a cognitive interview finding. In particular, with reference to the last issue, in order to determine whether or not a particular problem discovered is relevant enough to attempt finding a solution to it, the researcher must take into account various aspects such as the nature of the problem, whether or not the problem is tied to specific characteristics of respondents or possible experiences, and evaluate how the survey data may ultimately be affected by the flaw. This level of insight can only come from a systematic analysis across all of the cognitive interviews.

However, even though not allowing to estimate the actual magnitude of impact that the particular problem may have on survey data, the strength of the method is that it

reveals interpretive patterns of the question-response process which could lead to response error. As a qualitative study, cognitive test findings provide relevant pieces of insight from various perspectives that, when brought together, can assist the question design analyst in assessing the quality of the question as it pertains to the type of phenomena that it should capture. These findings can be considered also in the context of subsequent phases of field test of the questionnaire, which involve a larger number of respondents. In these phases findings and possible solutions identified through cognitive interviews can be validated using various techniques, including for example interviewers' *debriefing* or *behaviour coding* [5].

References

1. Beatty, P. Answerable questions: Advances in the methods for identifying and resolving questionnaire problems in survey research. Doctoral Dissertation, University of Michigan (2003). Dissertation Abstracts International, 64 (019), 3504A.
2. Bolton, R.N., Bronkhorst T.M. Questionnaire pretesting: computer-assisted coding of concurrent protocols. In Schwarz, N & Sudman, S (eds). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass Publishers, 37-64 (1996).
3. Conrad, F., Blair J. From impressions to data: Increasing the objectivity of cognitive interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9, (1996).
4. De Palma, E., Battisti, A., Tinto, A., Sebastiani, G., Results of the cognitive testing performed on the Italian version of the "Disability and social integration module" (2008), published on: http://circa.europa.eu/Public/irc/dsis/health/library?l=/reports/disability/edsim/annex_part_1pdf/EN_1.0_&a=d
5. Eurostat, Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System (2006) in: <http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPSQDET27062006.pdf>
6. Jabine, T.B., Straf M.L., Tanur J.M., Tourangeau R. (Eds), *Cognitive aspects of survey methodology: Building a bridge between disciplines*, Washington, DC: National Academy Press (1984).
7. Presser, S. et al. *Methods for testing and evaluating survey questionnaires* J. Wiley & Sons, (2004).
8. Tourangeau, R. Cognitive science and survey methods: A cognitive perspective. In Jabine T. B., Straf M. L., Tanur J. M., Tourangeau R. (Eds), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press (1984).
9. Willis, G.B. *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks. London, Sage publications (2005).