The value of cognitive interviewing for optimizing a patient experience survey.

CORINE BUERS [A,B], MATTANJA TRIEMSTRA [A*], EVELIEN BLOEMENDAL [A]. NICOLIEN C. ZWIJNENBERG [A], MICHELLE HENDRIKS [A] AND DIANA M.J. DELNOIJ [C,D].

[a] Netherlands Institute for Health Services Research (NIVEL), Utrecht, The Netherlands;
[b] School of Governance, Utrecht University, Utrecht, The Netherlands;
[c] Dutch Centre for Consumer Experience in Health Care, Utrecht, The Netherlands;
[d] Scientific Centre for Transformation in Care and Welfare (TRANZO), Tilburg, The Netherlands

This mixed-methods study uses both cognitive interviewing and a quantitative field test to provide empirical evidence on the value of cognitive interviewing for questionnaire development. Ten interviews were conducted with a questionnaire on patient experiences with cataract surgery (75-item consumer quality index cataract), using both thinking-aloud and probing techniques. Interviews were recorded and transcribed verbatim, problems were coded with the commonly used systems of Levine et al. and Willis, and results were compared with item non-response in a field test. The coding systems revealed similar numbers and type of problems: 55 items showed a total of 174 problems. However, most problematic items (67%) had an adequate response in the field test. Results stress the importance of cognitive interviewing as a pre-survey evaluation method to early identification of questionnaire problems, and it is recommended to use the coding system of Willis for it provides specific directions for questionnaire optimization.

INTRODUCTION.

Various methods exist to pre-test and optimize questionnaires in order to improve data quality such as cognitive interviewing, expert reviews and psychometric testing. Several researchers have addressed the potential strengths and weaknesses of these different pretesting methods (e.g. Horwood, Pollard, Ayis, McIlvenna, & Johnston, 2010; Presser et al., 2004; Rothgeb, Willis, & Forsyth, 2001; Willis, Schechter, & Whitaker, 1999). However, to date, there is little empirical evidence on the value of qualitative vs. quantitative methods of pre-testing.

Field testing or psychometric testing can be a useful method for identifying items with a high item non-response, whereas qualitative evaluation methods provide information about the nature of problems and possible directions for revising problematic items (Drennan, 2003; Knafl et al., 2007). Horwood et al. (2010)

recently compared findings of psychometric testing and cognitive interviewing (i.e.'think-aloud' method) and showed that these methods largely identified the same number of problematic items. Only four out of 59 items with more than one problem were not previously removed by statistical methods for item reduction.
They concluded that cognitive interviewing is valuable for questionnaire development and could complement statistical methods, but they also stated that comparing these two methods warrants further study. Therefore, we conducted a mixed-methods study to examine the merits of qualitative pre-testing, by means of cognitive interviewing, over field testing for optimizing a questionnaire.

Cognitive interviewing.

Cognitive interviewing has become a popular method for refining and validating questionnaires in the early stages of a questionnaire development process (Harris-Kojetin, Fowler, Brown, Schnaier, & Sweeny, 1999; Murtagh, Addington-Hall, & Higginson, 2007; Watt et al., 2008). Just like any other pre-survey evaluation method, cognitive interviewing identifies potential response problems that may compromise the response rate and the quality of the data (Ahmed, Bestall, Payne, Noble, & Ahmedzai, 2009; Beatty & Willis, 2007; Drennan, 2003; Knafl et al., 2007). The cognitive interview method entails administering a questionnaire and asking participants for additional verbal information (i.e. their thoughts and interpretations) to assess whether questions are comprehensible and are interpreted as intended (Beatty & Willis, 2007; Goldstein, Farquhar, Crofton, Darby, & Garfinkel, 2005; Willis, 2005). According to the cognitive processing model of (Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000), answering survey questions can be considered as a complex cognitive process which compromises four successive stages: (a) comprehension (i.e. comprehend questions or follow instructions as indented); (b) retrieval (i.e. remember relevant information); (c) judgment (i.e. final formulation of response, based on relevant memories); and (d) actual response (i.e. produce a response that is consistent with the personal experience) (Collins, 2003; Jobe, 2003; Watt et al., 2008; Willis, 1999). As such, cognitive interviewing provides insight into the type and cause of questionnaire problems as experienced by the study population and provides leads for revising problematic items (Knafl et al., 2007). Thus, cognitive interviewing may be an effective evaluation method to identify and solve questionnaire problems in early stages of questionnaire development, which might enhance questionnaires' reliability and validity (Ahmed et al., 2009; Beatty & Willis, 2007; Drennan, 2003; Knafl et al., 2007).
Unfortunately, there is no consensus on how to conduct, analyse or report cogni- tive interviews (Beatty & Willis, 2007; Boeije & Willis, 2011; Knafl et al., 2007; Levine, Fowler, & Brown, 2005; Presser et al., 2004). First of all, there are different techniques which can be used to conduct cognitive interviews: 'thinking aloud' and 'probing' (Beatty & Willis, 2007; Drennan, 2003; Willis, 2005). When using thinking aloud, participants are asked to read the questions out loud and to verbalize their thoughts as they fill in the questionnaire. When probing, the interviewer asks follow-up questions to comprehend a respondent's interpretation more precisely and clearly. Although, both techniques can be used simultaneously, most cognitive inter- view studies have solely used thinking aloud for refining and redrafting question- naires (Priede & Farral, 2011). Nevertheless, a combination of probing and thinking aloud is recommended to elicit as much information as possible on participants' interpretations and thoughts of a questionnaire (Priede & Farral, 2011).

Secondly, there are no uniform guidelines for identifying and coding questionnaire problems derived from cognitive interviews. To illustrate, difficulties with a specific term could be classified either as a lexical problem (Drennan, 2003), a clarity problem (Knafl et al., 2007; Levine et al., 2005) or as a technical term problem (Willis, 1999, 2009). These different coding systems may broadly cover the same categories of questionnaire problems, but it remains unclear whether the type of coding system used makes a difference for identifying and revising questionnaires' shortcomings.

Consequently, the large variety in cognitive interview methods and analysis hampers the comparison of research findings, and establishes the value of cognitive interviewing for questionnaire development. Specifically, the merits of cognitive interviewing using both thinking-aloud and probing techniques, next to quantitative testing, need further exploration.

Aim and research questions.

The current mixed-methods study aims to contribute to the existing body of literature on the value and practical implications of conducting cognitive interviewing in the questionnaire development process in two ways. Firstly, we aimed to determine the merits of cognitive interviewing (i.e. qualitative testing) over quantitative testing by comparing the findings of cognitive interviews with item non-response in a field test. Secondly, this study aims to provide insight into the usefulness of coding systems for identifying item flaws and optimizing a questionnaire by comparing two commonly used coding schemes for cognitive interviewing, i.e. those of Levine et al. (2005) and Willis (1999, 2009). The research questions of this study are: (1) what is the value of cognitive testing in addition to quantitative field testing in a questionnaire development trajectory? and (2) which coding system for cognitive interviewing is most useful for identifying questionnaire problems in order to optimize a questionnaire?

METHODS.

CQI cataract questionnaire.

This paper reports on the results obtained from cognitive interviewing and a large field test with a self-report questionnaire on patient experiences with cataract surgery, the consumer quality index (CQI) cataract. The CQI (consumer quality index or CQ-index) is a standard and validated instrument to measure patient experiences in various health care settings in the Netherlands (Delnoij, Rademakers, & Groenewegen, 2010; Koopman, Sixma, Hendriks, de Boer, & Delnoij, 2011). Developing a CQI consists of a standardized multistage approach with both qualitative research (e.g. focus groups) and quantitative testing phases (e.g. field test and psychometric testing). However, cognitive interviewing is not a standard procedure in the CQI development trajectory yet.

The CQI cataract questionnaire is an instrument to measure patient experiences with the quality of care after a cataract operation (Brouwer, Sixma, Triemstra, & Delnoij, 2006; Stubbe, Brouwer, & Delnoij, 2007; Stubbe & van Dijk, 2007). This instrument has been derived from the QUOTE-Cataract and the Dutch H-CAHPS, and is further developed with qualitative research (e.g. focus groups) and a quantitative field test (N= 4635). The CQI cataract questionnaire consists of 75 items that cover several domains: communication with the ophthalmologist and nurses, information provision, pain management, medication, health insurance, global rating of the hospital and background characteristics of participants such as age, gender and health status.

The CQI cataract questionnaire was selected for this study for two reasons.

First, cognitive problems were expected to be more common in an elderly population such as cataract patients compared to younger patients. Thus, solving questionnaire problems for this population could also benefit other patient experiences questionnaires. Second, the CQI cataract questionnaire has already been tested quantitatively in a large-scale field test (Stubbe et al., 2007), but not with cognitive interviews yet. Therefore, the value of cognitive interviewing for optimizing this questionnaire can be effectively established by comparing results of this qualitative method with the response problems (i.e. item non-response) in the quantitative field test.

Procedure of data collection: cognitive interviews.

Cognitive interviewers were conducted with the CQI cataract questionnaire among people who underwent cataract surgery at the Rotterdam Eye Hospital in the Netherlands in October 2010. A purposive sample (N = 26) included adult patients with a sufficient comprehension level of Dutch and who had cataract surgery in the past six weeks. Eligible patients were contacted by telephone and asked to participate in the study. Of the 26 patients who were called, 12 patients could not be reached, two other patients did not meet the inclusion criteria due to language problems and two patients were unable to participate. Ten patients agreed to participate in the cognitive interviews. All participants received general information about the aim of the study and the location of the cognitive interviews by email.

Ten cognitive interviews were conducted in a private room after a follow-up visit to the ophthalmologist in the Rotterdam Eye Hospital. The interviews were face-to-face performed by an experienced researcher (EB). After a short introduction, a written consent was obtained from participants before starting the interview. During the cognitive interviews, both thinking-aloud and probing techniques were used (Priede & Farral, 2011). Participants were asked to think aloud and to verbalize their thoughts while answering questions and, if necessary, the interviewer used follow-up questions (i.e. probes) to get additional information about the understanding and interpretation of the questionnaire. Probing was used if participants were hesitating or unclear during their thinking aloud and if they had to make an estimation of time or asked the interviewer for help. The interviewer used both pre-scripted and spontaneous probes during the interview (concurrent) and directly after the interview (retrospective); see Table 1 for some examples. Furthermore, the interviewer took observation notes of respondents' behaviour such as mumbling, sighing and having problems with the routing of the questionnaire (e.g. skipping items).

All interviews were audio-taped and lasted 20–58 min (mean: 34 min). After the interview, the participants received a small gift certificate and a reimbursement of travel expenses.

[TABLE 1].

Procedure of data collection: field test.

A large-scale field test was conducted with the CQI cataract questionnaire among people who underwent cataract surgery in the Netherlands in 2004 or 2005 (Brouwer et al., 2006; Stubbe et al., 2007; Stubbe & van Dijk, 2007). The questionnaire was sent by email to 6468 cataract patients through four Dutch hospitals after cataract surgery (n = 1145) and four Dutch insurance companies who received claimed costs of cataract surgery (n = 5323). A total of 5436 patients returned the questionnaire but 801 cases were excluded from analysis, either because they were not willing or able to participate (n = 447), did not complete the questionnaires by themselves (n = 203), did not indicate to have underwent cataract surgery within the past 12 months (n = 119), or did not fill in core items (n = 32). Consequently, the final sample consisted of 4635 cataract patients (response rate 72%). These respondents seemed to represent the Dutch population of cataract patients fairly well as they did not significantly differ from the non-response group with respect to gender and education level, although elderly patients (_80 years) and those from a nonnative origin were slightly under-represented (Brouwer et al., 2006; Stubbe et al., 2007; Stubbe & van Dijk, 2007).

The item non-response of this field test was used as an indicator for cognitive problems expressed by the percentage of respondents that did not understand or failed to answer a question. Following the CQI Manual for questionnaire development (Koopman et al., 2011), we used the directive that item non-response should not exceed 5%; otherwise, the question should be reformulated or might be deleted.

Data processing and analysis.

The 10 cognitive interviews were transcribed verbatim, including elaborations on: (a) how respondents constructed their answers; (b) interpretations of questions; and (c) any difficulties in answering questions (Beatty & Willis, 2007). Furthermore, the completed questionnaires were compared to the audio and verbatim transcripts of the cognitive interviews to assess the match between verbal and written answers.

Data analysis started with reading the interview transcripts and notes of the interviewer thoroughly, and potential problems (e.g. hesitation, ambiguous interpretation and skip problems) were marked independently by two researchers (CB, NZ).

Secondly, both researchers coded the marked problems independently by using both the coding system of Levine et al. (2005) and of Willis (1999, 2009). The former distinguish six broad categories to classify questionnaire problems, whereas the latter identifies seven broad categories with specific subcategories (see Table 2).

Although it was aimed to assign only one code per coding system to a problematic item, in order to identify predominant problems, sometimes multiple codes were used. For example, some participants experienced both recall and clarity problems with the question: ‘How long did you have to wait for eye surgery?’ Problems that were not clearly related to either one category of the coding systems were content analysed and categorized accordingly. Subsequently, the coded sections of the interviews were compared and discussed with a third researcher (MT), until consensus was reached on the coding. Finally, all data were entered in a spreadsheet to get a systematic overview of problems per item of the CQI cataract questionnaire; including both qualitative data (i.e. verbatim quotations, comments of researchers and problem codes) and quantitative data (i.e. number of problem codes and percentage of item non-response from the previously conducted field test).

[TABLE 2.].

RESULTS.

Participants.

The 10 participants to the cognitive interviews, including three men and seven women, were all Dutch and at least 55 years old (see Table 3). They represented the cataract population well regarding gender and education level, but they were relatively young and all of them were native speakers with a good health status. Three participants had their second cataract surgery.

Problematic items and problem codes.

A total of 55 (73%) of the 75-items of the CQI cataract questionnaire were considered to be problematic according to at least one participant (see Table 3). For each participant, between seven to 25 problematic items were identified. No relation between patient characteristics (age, gender, education) and the number of questionnaire problems was observed. For the 55 problematic items, a total number of 174 problems were identified. This resulted in 189 problem codes according to the two coding systems and 10 ‘other problems’ that were not predefined. The difference between the total number of questionnaire problems (174 problems) and the total number of problem codes (199 codes) was caused by sub-questions or items with multiple problems. There were no differences in the number of coded problems between the coding system of Levine et al. and Willis.

[TABLE 3].

Identified problems.

According to the coding system of Levine et al. 189 problems could be identified, including 78 comprehension problems, 50 knowledge problems, 37 general problems, 20 inapplicable problems and four answering problems. The 189 problems encoded according to Willis’ system included: 78 clarity problems, 50 knowledge problems, 30 response categories, 18 instructions, 12 assumptions and one formatting problem.

Although, almost all problems could be coded with the existing coding systems, two other types of cognitive problems emerged. First, a mismatch problem occurred for four items with five different participants, that is to say a difference was found between participants’ verbal thinking-aloud comments and their written answers to these questions. For example, one participant commented that the ophthalmologist did not inform her about the risks of a cataract surgery, but ticked the answer ‘yes’ in response to ‘Did the ophthalmologist inform you about the risks of a cataract surgery?’ Secondly, two participants misread a certain word of a question. For example, one participant read the word ‘contact’ instead of ‘contract’ in the question: ‘Did you visit an ophthalmologist for your cataract surgery that has no contract with your health care insurance?’ Both coding systems showed that the majority of the detected questionnaire problems concerned comprehension followed by knowledge problems.

Most problematic items.

Table 4 provides a detailed overview of the three most problematic items of the CQI cataract questionnaire. The type of cognitive problems, the number of participants who experienced difficulties in answering and the item-non response in the field test are displayed.

In the first example, almost all participants experienced comprehension and knowledge problems with the question: ‘How long did a consultation with the ophthalmologist approximately take?’, although the item non-response for this question was rather low. Participants struggled with this question because some participants were unfamiliar with the meaning of ‘consultation’ and included ‘time in the waiting room’ to construct their answers, as shown by the following statement:

- Well, more than fifteen minutes I suppose. Not only the time spent with the ophthalmologist himself, but you have to wait as well, and Is the time you have to wait included? Or do you mean the actual time I spend with the ophthalmologist?

Others faced difficulties in estimating the consultation time regardless of the predefined response categories. Participants then had to reread the question, were mumbling or sighing:

- Pfff, I think between ten and fifteen minutes. Or ten … Hmm … Let’s see … Ten minutes. I will write down five to ten minutes (…) It’s difficult because it could have been five to ten minutes, or it may have been ten to fifteen minutes. I never checked my watch.

So, participants tried to make sense of a question which they found difficult to answer and, as a result, various interpretations and answers could emerge. For this item, the codes of Willis (i.e. ‘reference period’ and ‘technical term’) clearly provide specific directions for revision, unlike the broad categories of Levine et al.

The subcategories of Willis immediately indicate that this question could be reformulated and simplified by defining the reference period and by clarifying or replacing the term ‘consultation’.

[TABLE 4].

In the second example (‘Have you talked with anyone about whether you would have the necessary help at home after your cataract surgery?’), participants also misunderstood the question or thought it did not apply to their situation. Participants for example stated: I already have someone who helps me out, so this does not apply for me or It was not necessary, so I did not speak about it. As illustrated, participants lacked contextual information (e.g. anyone from the hospital staff or family), struggled with the concept ‘help at home’ (e.g. professional assistance or family) or missed an answering category. This example showed a clear difference between the two coding systems. According to the system of Levine et al. this item showed ‘general’ and ‘comprehension’ problems, whereas Willis’ system clearly indicated that the question is vaguely formulated and that a response category is missing. Thus, two recommendations for item revision directly followed from the sub-codes of Willis: (a) focus on concrete actions (e.g. ‘did the hospital staff ask you about …?’) and (b) add an extra response category (e.g. ‘not applicable’ or ‘I already have someone who helps me at home’).

Finally, a lack of information may also produce unintended interpretations of a question as shown by the third example (‘was your ophthalmologist aware of your general health status?’). Six people did not know whether their ophthalmologist was well informed about their general health status. Surprisingly, almost all participants ticked the box ‘yes’, and stated they assumed that the ophthalmologist was aware of their status. One participant told: Yeah, yeah, they asked about my health status during the intake at the hospital. I suppose that it is written in my file, but the ophthalmologist did not ask about that (…) Well, I feel … I assume that such an intake list has a reason and is used by the ophthalmologist.

This indicates that participants may provide answers on the basis of their assumptions instead of the actual behaviour. Therefore, it is recommended to focus questions on concrete and observable behaviour (e.g. ‘did your ophthalmologist ask about your health status?’).

To summarize, these examples showed that cognitive interviews provided insight into the type of problems participants face when completing a questionnaire.

The majority of the comprehension problems occurred due to lack of a clear reference period and multiple ways to interpret questions, whereas most knowledge problems emerged due to ignorance or recall problems of patients, or response categories being too specific. Therefore, most questionnaire problems can be solved by reformulating questions in a more clear, comprehensible, specific and straightforward manner, by focusing on actual behaviour, or by adjusting or adding a response category. Furthermore, the item non-response of these three most problematic questions showed that cognitive problems are not by definition related to a high item-non response as two of these items met the directive of less than 5% item non-response.

[TABLE 5].

Item non-response and identified problems.

Table 5 shows the association between the item non-response in a previous field test and the number of problems experienced with certain items during the cognitive interviews. Since there were no differences between the number of problem codes encoded according to Levine et al. or Willis, the results represent the findings for both coding systems.

Three quarter (57 items) of the 75 items of the CQI cataract questionnaire were filled out by more than 95% of the respondents to the field test and thus met the directive ‘less than 5% item non-response’ of the CQI manual. However, 38 of these 57 items (67%) did show problems in the cognitive interviews but there were no candidates for revision according to the field test results because these items met the 5%-directive. Many of these items even showed multiple problems. Another 18 items exceeded the directive of the field test and almost all of these items (17/18) also appeared to be problematic in the cognitive interviews. Only 19 items showed no problem at all.

Altogether, in the case of the CQI cataract questionnaire, cognitive interviewing was about three times more sensitive for identifying problematic items (55/75 items, 73%) than the indicator ‘item non-response’ derived from field testing (18/75 items, 24%).

DISCUSSION.

This study aimed to assess the value of cognitive interviewing in addition to field testing and to determine which commonly used coding system for cognitive interviewing (i.e. the system of Levine et al. or Willis) is most useful for optimizing a questionnaire. The study clearly shows merits of cognitive interviewing over quantitative pretesting of a self-report questionnaire on patient experiences with cataract surgery. Cognitive interviewing yielded about three times more problematic items than those could be identified with high item non-response in a field test. The number of problems identified by cognitive interviewing was only partly related to the item non-response in the field test. Although items with a high non-response were also likely to show problems in the cognitive interviews, most of the problematic items found (67%) did have acceptable non-response rates. This suggests that participants who do provide a written answer may not always understand the question as intended by the researcher. Thus, cognitive interviews could effectively add to the validity and reliability of questionnaires by identifying problems that would have remained unnoticed from a quantitative pre-test. Furthermore, results indicate that the type of coding system does not play a significant role in detecting questionnaire problems as they revealed similar numbers and generally the same types of problems. Nevertheless, the system of Willis (1999, 2009) appeared to be more useful for questionnaire optimization than the coding system of Levine et al. (2005), as it provides more detailed codes and specific directions for revisions.

Contribution to literature

This study contributes to the literature in two ways. First, this study highlighted the importance of combining cognitive interviewing and quantitative field testing into a multistage process of questionnaire development in order to enhance the validity of survey questionnaires. In line with the study of Horwood et al. (2010), who used think-aloud techniques and psychometric testing to detect questionnaire problems, almost all of the problematic items in the field test also appeared to be problematic in cognitive interviewing. However, in contrast to Horwood et al. we demonstrated the sensitive nature of cognitive interviewing with both thinking-aloud and probing techniques to identify questionnaire problems that otherwise would have remained unnoticed in field testing. Specifically, probing could have identified an additional number of item flaws such as misinterpretations. These results raise concerns about the validity of survey questionnaires that have not been cognitively tested. Therefore, we recommend to integrate cognitive interviewing, using both the thinkingaloud and probing techniques, in an early stage of questionnaire development to detect problematic items and to optimize a questionnaire before conducting quantitative research. Nevertheless, further study is warranted to show whether revisions of problematic items following cognitive interviewing indeed improve the response, validity and reliability of the questionnaire.

Secondly, this study contributes to uniform guidelines regarding coding and analysing cognitive interviews as it provides insight into the usefulness of the commonly used coding systems of Levine et al. (2005) and Willis (1999, 2009). Despite differences in their classification of coding, both systems identified as the same number and types of questionnaire problems. However, the coding system of Willis seems to be most helpful for optimizing a questionnaire as it provided more detailed information about the type of problems, thus indicating specific directions for revisions.

Willis’ system was especially valuable for items with multiple comprehension problems due to its specific subcategories (e.g. wording, technical term, vague and lack of reference periods) in contrast to the broad 'comprehension' category of Levine et al. Furthermore, the broad category of Levine et al. on 'general problems' could suggest different options for revision, because they include both 'referral problems' and 'lack of response categories' whereas Willis identifies these problems separately. Consequently, we recommend the coding system of Willis for cognitive interviewing in order to identify and solve questionnaire problems.

Limitations There are some drawbacks of this study that should be noted. In recent years, there has been an extensive debate about the benefits and limitations of cognitive testing (Beatty & Willis, 2007; Drennan, 2003). According to several researchers, the main flaw of cognitive interviewing is that cognitive processes are complex and are performed so rapidly, that it is difficult to accurately explore thoughts and interpretations of participants (Collins, 2003; Drennan, 2003). Although, the think-aloud technique is used to elicit information processing, it is debatable to what extent insight into the question-answer-process can be obtained. The mismatch found between the think aloud and written answers in our study supports this notion of human information processing as a 'black box'. Nevertheless, by using thinking aloud as well as probing techniques, the interviewer aimed to gain as much verbal information as possible on participants' question-and-answer process. Study limitations must also be considered when interpreting the findings of this study. First, the findings of the cognitive interviews are restricted to a small sub-sample of cataract patients who were all native speakers and relatively young compared to the total cataract population. In addition, cataract patients are relatively old compared to other patient populations. Therefore, the question remains whether the type and number of identified problems would also emerge in the total cataract population and in other patient populations. As we selected this particular patient population on purpose because we expected the elderly to reveal more problems in cognitive interviewing, the merits of cognitive interviewing over quantitative testing might be somewhat overestimated in this study. Nevertheless, the findings could benefit other patient survey questionnaires which use similar sets of items and response categories.

Another limitation is that the revisions of problematic items are not subsequently tested in additional rounds of cognitive testing or a field test. Thus, it remains unknown whether revisions based on the cognitive interviews do indeed solve the questionnaires' difficulties and will actually improve the quality of the data.

Conclusions Face-to-face cognitive interviewing using both thinking-aloud and probing techniques is a sensitive qualitative method which can be used to optimize a patient experience questionnaire in addition to quantitative field testing. Revising instruments following cognitive interviewing is expected to increase the response, and the validity and reliability of data. Therefore, it is recommended to incorporate face-to-face cognitive interviewing as part of a multistage questionnaire development trajectory. Furthermore, we recommend the coding systems of Willis to analyse questionnaire problems and to optimize it as it provides more detailed codes that indicate specific directions for revisions. Nevertheless, the most effective and efficient way to conduct and analyse cognitive interviews remains a thorny one. Future research should explore cognitive interview procedures in greater detail and further research is needed to scrutinize the value of cognitive interviews by empirically testing the revised questionnaire.

Authors' contributions EB, MH and DD collaborated in designing the study. EB conducted the data collection and interviews. CB and NZ read the transcripts and coded the data with the assistance of MT. CB and MT analysed the data and drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank those who participated in the cognitive interviews and the staff of the Rotterdam Eye Hospital who helped to recruit the participants and who facilitated the interviews by providing an interview room. Financial support for this study was provided by the Dutch Centre for consumer experience in health care.

Notes on contributors

Corine Buers, MSc, is a sociologist who has interest in questionnaire development process. She has worked on several projects with the consumer quality index at NIVEL and now she is working as a PhD student at the School of Governance, Utrecht University.

Mattanja Triemstra, PhD, is a biomedical scientist, working as a senior researcher at NIVEL. She has ample expertise in developing and conducting patient experience surveys to evaluate the quality of care from the patients' perspective.

Evelien Bloemendal, MA, is a psychologist, trained in methodology and statistics. She worked at NIVEL as a researcher, conducting several projects including studies with the consumer quality index.

Nicolien C. Zwijnenberg, MSc, is a health scientist working as a researcher at NIVEL. She has worked on several projects and guideline development concerning the consumer quality index.

Michelle Hendriks, PhD, is a health psychologist, working as a senior researcher at NIVEL. She has expertise in measuring patient experiences and evaluating ways to present comparative healthcare information on the Internet.

Diana M.J. Delnoij, PhD, is a political scientist. She is the director of the Dutch Centre for Consumer Experience in Health Care (Centrum Klantervaring Zorg) and professor of transparency in Health Care at Tranzo, Tilburg University.

References

Ahmed, N., Bestall, J., Payne, S., Noble, B., & Ahmedzai, S. (2009). The use of cognitive interviewing methodology in the design and testing of a screening tool for supportive and palliative care needs. Supportive Care in Cancer, 17, 665–673.

Beatty, P., & Willis, G. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71, 287–311.

Boeije, H. R., & Willis, G. (2011). Introducing the CIRF: A framework for reporting cognitive interviewing studies. Paper presented at the European Survey Research Association, Lausanne, Swiss.

Brouwer, W., Sixma, H., Triemstra, M., & Delnoij, D. (2006). Kwaliteit van zorg rondom een staaroperatie vanuit het perspectief van patiënten: meetinstrumentontwikkeling [Quality of care of a cataract surgery from the perspective of patients: Instrument development]. Utrecht: NIVEL.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. Quality of Life Research, 12, 229–238.

Delnoij, D. M. J., Rademakers, J., & Groenewegen, P. P. (2010). The Dutch consumer quality index an example of stakeholder involvement in indicator development. BMC Health Service Research, 10, 88.

Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. Methodological Issues in Nursing Research, 42, 57–63.

Goldstein, E., Farquhar, M., Crofton, C., Darby, C., & Garfinkel, S. (2005). Measuring hospital care from the patients' perspective: An overview of the CAHPS hospital survey development process. Health Research and Educational Trust, 40, 1977–1995.

Harris-Kojetin, L., Fowler, F., Jr., Brown, J., Schnaier, J., & Sweeny, S. (1999). The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Consumer Assessment of Health Plans Study. Medical Care, 37, M10–M21.

Horwood, J., Pollard, B., Ayis, S., McIlvenna, T., & Johnston, M. (2010). Listening to patients: Using verbal data in the validation of the Aberdeen Measures of Impairment, Activity Limitation and Participation Restriction (Ab-IAP). BMC Musculoskelet Disord, 11(182), 182.

Jobe, J. (2003). Cognitive psychology and self-reports: Models and methods. Quality of Life Research, 12, 219–227.

Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., & Dixon, J. (2007). The analysis and interpretation of cognitive interviews for instrument development. Research in Nursing & Health, 30, 224–234.

Koopman, L., Sixma, H., Hendriks, M., de Boer, D., & Delnoij, D. (2011). Handboek CQI Metingen: Richtlijnen en voorschriften voor metingen met een CQI meetinstrument [Manual CQI Development; Guidelines for the development of a CQI measurement instrument]. Utrecht: NIVEL.

Levine, R., Fowler, F., Jr., & Brown, J. (2005). Role of cognitive testing in the development of the CAHPS hospital survey. Health Research and Educational Trust, 40, 2037–2056.

Murtagh, F., Addington-Hall, J., & Higginson, I. (2007). The value of cognitive interviewing techniques in palliative care research. Palliative Medicine, 21, 87–93.

Presser, S., Couper, M., Lessler, J., Martin, E., Rothgeb, J., & Singer, E. (2004). Methods for testing and evaluating survey questions. Public Opinion Quarterly, 68, 109–130.

Priede, C., & Farral, S. (2011). Comparing results from different styles of cognitive interviewing: 'Verbal probing' vs. 'thinking aloud'. International Journal of Social Research Methodology, 14, 271–287.

Rothgeb, J., Willis, G., & Forsyth, B. (2001). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? Paper presented at the annual meeting of the American Association for Public Opinion Research (August 5–9), Montreal, QC. Proceedings of survey research methods section of the American Statistical Association. Washington, DC: American Statistical Association.

Stubbe, J., Brouwer, W., & Delnoij, D. (2007). Patients' experiences with quality of hospital care: The consumer quality index cataract questionnaire. BMC Ophthalmology, 7, 14.

Stubbe, J., & van Dijk, L. (2007). Het discriminerend vermogen van de CQ index Heup-/Knieoperatie [The discriminatory power of the CQ index cataract surgery]. Utrecht: NIVEL.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), Cognitive aspects of survey methodology: Building a bridge between disciplines (pp. 73–100). Washington, DC: National Academies Press.

Tourangeau, R., Rips, R. J., & Rasinski, K. A. (2000). The psychology of survey response. Cambridge: Cambridge University Press.

Watt, T., Rasmussen, A., Groenvold, M., Bjorner, J., Watt, S., & Bonnema, S. (2008). Improving a newly developed patient-reported outcome for thyroid patients, using cognitive interviewing. Quality of Life Research, 17, 1009–1017.

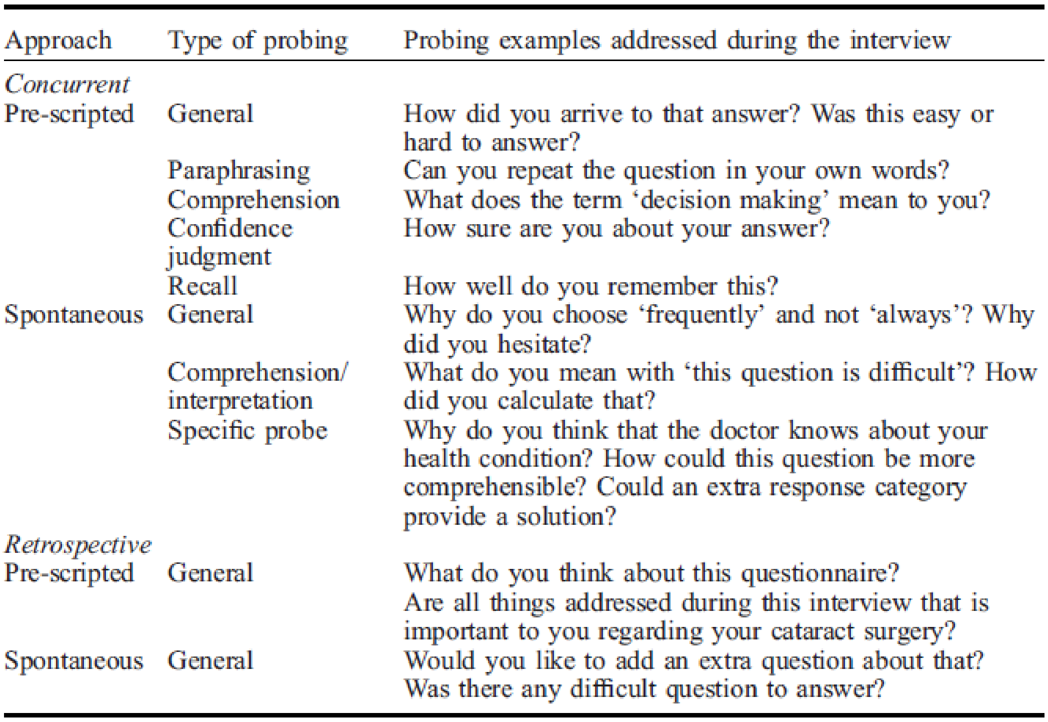
Willis, G. (1999). Cognitive interviewing a 'how to guide'. Reducing survey error through research on the cognitive and design processes in surveys. Short course presented at the 1999 meeting of the American Statistical Association. Rockville, MD: Research Triangle Institute. Retreived from: http://fog.its.uiowa.edu/~c07b209/interview.pdf Willis, G. (2005). Cognitive interviewing. Thousand Oaks, CA: Sage.

Willis, G. (2009, June 23). Question Appraisal System (QAS-2009) simplified coding form. Presented at Masterclass Pretesting Questionnaires. Utrecht: Utrecht University. Retreived from http://drie.moaweb.nl:8080/MOA/kenniscentrum/materiaal-bijeenkomsten/archief-materiaal-bijeenkomsten/2009/pretesten-van-vragenlijsten-23-juni/QAS%20Coding%20Form%20reduced.pdf/at_download/file Willis, G., Schechter, S., &

Whitaker, K. (1999). A comparison of cognitive interviewing, expert review, and behavior coding: What do they tell us? American Statistical Association.

TABLES

Table 1. Probing examples addressed during the cognitive interviews classified according to Beatty and Willis (2007).



Approach	Type of probing	Probing examples addressed during the interview
Concurrent		
Pre-scripted	General	How did you arrive to that answer? Was this easy or hard to answer?
	Paraphrasing	Can you repeat the question in your own words?
	Comprehension	What does the term 'decision making' mean to you?
	Confidence judgment	How sure are you about your answer?
	Recall	How well do you remember this?
Spontaneous	General	Why do you choose 'frequently' and not 'always'? Why did you hesitate?
	Comprehension/ interpretation	What do you mean with 'this question is difficult'? How did you calculate that?
	Specific probe	Why do you think that the doctor knows about your health condition? How could this question be more comprehensible? Could an extra response category provide a solution?
Retrospective		
Pre-scripted	General	What do you think about this questionnaire? Are all things addressed during this interview that is important to you regarding your cataract surgery?
Spontaneous	General	Would you like to add an extra question about that? Was there any difficult question to answer?

Table 2. Coding systems for classifying questionnaire problems of Levine et al. (2005) and Willis (1999, 2009).

Levine et al. (2005)	Willis (1999, 2009)
Comprehension: Items with unclear or ambiguous terms, failed to understand the questions consistently	**Clarity**: Problems with the intent or meaning of a question *Subcategories*: wording, technical term, vague and lack of reference periods
Knowledge: Items for which respondents lacked information to answer a question	**Knowledge**: Likely to not know or have trouble remembering information *Subcategories*: knowledge, recall, computation
Inapplicable: Items measuring construct that are inapplicable for many respondents (e.g. made assumptions)	**Assumptions**: Problems with assumptions or underlying logic *Subcategories*: inappropriate assumptions, assuming constant behaviour and double-barrelled
Construct: Items failed to measure the intended construct	**Response categories**: Problems with the response categories *Subcategories:* missing, mismatch question-answer, vague, open-ended questions, overlapping and illogical order
Subtle: Items making discriminations that are too subtle for many respondents	**Sensitively**: Sensitive nature or wording/bias *Subcategories*: sensitive content (general), sensitive wording (specific) and socially acceptable
General: Several other general issues associated with the development of a questionnaire	**Instructions**: Problems with introductions, instructions or explanations **Formatting**: Problems with lay out or question ordering

Table 3. Descriptives of the 10 participants and questionnaire problems encountered in cognitive interviews with the 75-item CQI cataract questionnaire.

P	Sex	Age group	Education	Cataract surgery	Number of problematic questions	Number of problem codes cf. Levine et al. (2005)	Number of problem codes cf. Willis (1999, 2009)	Other problem codes
P1	F	75–79	Low	First	25	27	27	1
P2	M	75–79	Low	Second	21	24	24	1
P3	F	75–79	Medium	First	12	15	15	–
P4	F	65–74	Low	Second	14	14	14	–
P5	F	65–74	Low	First	7	6	6	2
P6	M	>80	Medium	First	15	18	18	1
P7	F	65–74	Higher	Second	18	18	18	–
P8	F	>80	Medium	First	22	23	23	1
P9	M	55–64	Higher	First	21	23	23	1
P10	F	65–74	Medium	First	19	21	21	3
Total problems					174	189	189	10

Table 4. Examples of most problematic items and their codes according to the two coding systems.

Original item	Response categories	Problem code of Levine et al. (2005)	Problem code of Willis (1999, 2009)	Number of participants who experienced problems	Item non-response (%)
1. How long did a consultation with the ophthalmologist on average take?	Less than 5 min 5–10 min 10–15 min More than 15 min	Comprehension	Clarity: –Reference period –Technical term	8	1.6
		Knowledge	Knowledge: –Computation –Recall		
2. Have you ever talked with anyone about whether you would have the necessary help at home after your cataract surgery?	No Yes	General	Response category: –Missing	7	5.4
		Comprehension	Clarity: –Vague		
3. Was your ophthalmologist aware of your general health status?	No Yes	Knowledge	Knowledge: –Knowledge	6	3.4

Table 5. Distribution of the 75 items of the CQI cataract questionnaire according to the item non-response in field testing ($n = 4635$) and the number of problems per item or the total number of problematic items in cognitive interviewing ($n = 10$).

Percentage item non-response in field test	Number of items	Number of problems per item in cognitive interviews				Total number of problematic items
		0 problem	1 problem	2–5 problems	>5 problems	(>1 problems)
0–5% item non-response	57 (76%)	19	13	17	8	38
5–10% item non-response	14 (19%)	1	2	10	1	13
>10% item non-response	4 (5%)	0	1	1	2	4
Total number of items (%)	75	20 (27%)	16 (21%)	28 (37%)	11 (15%)	55 (73%)