SELECTING PRETESTING TOOLS ACCORDING TO A MODEL OF QUESTIONNAIRE DEVELOPMENT, WITH ILLUSTRATIONS CONCERNING PATIENT SATISFACTION WITH MEDICAL CARE

Hans Akkerboom and Annemiek Luiten, Statistics Netherlands (CBS)[1]
Hans Akkerboom, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands

Keywords: Pretesting Model, Patient Satisfaction

1. Selecting tools for a pretesting program

At Statistics Netherlands (CBS), a questionnaire laboratory, more fully called ''Questionnaire Design Resource Center (QDRC)", offers research facilities and methods for questionnaire design and development. Such tools are used in *pretesting programs* as part of survey design and development. A pretesting program is considered to consist of *a series of interdependent empirical research steps for questionnaire design and development, meant to strike a balance between information supply and demand.* This balance refers to the quality of the fit between operationalizations of research aims and respondent-related data concepts and data handling (both content and procedure of the survey). We discuss the case of a continuous survey of patient satisfaction with medical care in a general hospital, as an illustration of how we use a (tentative) Model of Questionnaire Development as a guideline in composing pretesting programs.

1.1. A Model of Questionnaire Development: various tools for various phases. QDRC facilities comprise

- *Advice* on questionnaire drafting,
- *Review* of draft questionnaires,
- *General consultation* of volunteer respondents and/or users,
- *Observation* of ordinary (structured) interviews,
- *Specific consultation* of volunteer respondents (interviewers, proxies) *about questionnaires,*
- Design and development of *monitoring tools for pilot studies* (field trials, or, perhaps, experiments),
- Design and development of *monitoring tools for surveys* in progress.

The QDRC is losely called 'questionnaire laboratory'. This term focuses on consultation and observation, which usually require test sessions during which volunteer respondents are asked to take part in 'ordinary interviews', 'in-depth interviews', 'cognitive interviews', 'focus groups', or combinations thereof. As a rule, test sessions are held either at the questionnaire laboratory or at the 'natural response location', i.e. the place where respondent and survey are supposed 'to meet each other'.

At the QDRC, in-depth interviews are considered to be a sort of 'one-to-one focus group', whereas cognitive interviews comprise thinkaloud protocols and other 'cognitive stimuli', like paraphrasing and sorting tasks. Focus groups and in-depth interviews are often used to research key issues of survey content and procedure. Cognitive interviews are often used to spot potential 'response errors' in a (prototype) questionnaire. Focus groups and in-depth interviews can also be used to the latter purpose, provided the participants get advance information about the questionnaire under study. It is quite common at the QDRC to have ordinary interviews immediately precede two types of focus group, one with the respondents and one with the interviewers.

The QDRC facilities are in line with the five-step Model of Questionnaire Development indicated in Scheme 1 below. Step 1 checks how volunteer respondents view the topic and procedures of the survey: 'What is relevant and measurable from their viewpoint?' These issues may return in Step 2, which is primarily meant to check how respondents fare while answering the questions and transmitting the answers: 'What conflicts of interpretation can arise? Under what circumstances is the respondent willing and able to provide data with reasonable effort and cost'? Step 3 has the same flavor as Step 2, but the focus is on data collection rather than the questionnaire. Step 3 begins to adress the issue 'Can information exchange be organized in an easier and cheaper way, with higher quality?', which is the proper topic of Step 4. By then, emphasis has shifted from qualitative exploratory tools to quantitative confirmatory ones. Steps 1-4 comprise a *Pretesting Model,* which may guide in decisions on *what to test when and by what means.* Step 5 reflects the transition from pretest to survey.

[1] The views and opinions expressed in this paper are those of the authors and do not necessarily reflect official policies of Statistics Netherlands (CBS).

2. Patient satisfaction: 'The survey to be designed'

The following case was the result of a general hospital in the Netherlands seeking methodological support from the QDRC. Survey instruments based on ex-patient reporting had to be developed that would yield

- quick and global indicators of patient satisfaction on a continuous basis, say once in a quarter,
- in-depth exploration of , and possible solutions for, any suspected problem area on an incidental basis.

On the basis of these goals, it was decided to develop (1) a *global, self-administered questionnaire* covering a limited number of broadly varying satisfaction items, and (2) *a detailed questionnaire* for *structured and semi-structured face-to-face interviews,* meant to elaborate on those subjects or items that might show up as 'alarming' at anyone time, on the basis of (1). The items should appear *relevant to patient satisfaction with medical care.* The questionnaires and their administration should conform to the *needs and capabilities of ex-patients.* Data collection should be easy to incorporate into *daily hospital practice.*

The QDRC's role focused on the following projected pretesting steps (*G=global, D=detailed*):

Step 1: Definition and feasibility study:

- consultation of hospital personnel (e≈10),
- consultation (at CBS) of volunteer respondents (v≈2x5) having recent experience with care in an arbitrary general hospital,
- (from the above and a literature review:) design of prototypes of the global and detailed questionnaire by 'QDRC experts' (e=2), review by hospital staff (e≈7), and redesign by the QDRC (e=1).

SCHEME 1. A Model of Questionnaire Development *(Steps 1-4: a Pretesting Model)*

Step (: goal)	**Topics**	**Tools**	**Test size***

1. Definition/feasibility study: prototype questionnaire and data collection procedure (go/no go)	Survey *topic*, key concepts, key procedures
	Key respondent *attitude*
	Data *accessibility*
		Review
		Expert** appraisal
		Focus groups
		In-depth interviews
			
			e=3...25
			v (i,u)=1×5...3×10
			v=1...10

2. Qualitative content test: 'less error-prone' questionnaire draft	Survey topic, key *concepts*, key *procedures*
	Key attitude/motivation in *social encounter*
	Interpretation of 'items and answers', *retrieval, judgment, answer* selection
	Data *disclosure*
		Focus groups
		Observation (ordinary interviews)
		In-depth interviews
		Cognitive interviews
		Expert reappraisal
			v (i,u)=1×5...3×10
			v or n =5...50
			v=1...10
			v=5...50
			e=3

3. Qualitative operational test: 'less error-prone' data collection procedure	*Response* (unit, item)
	Key *procedures*
	Social encounter
		Observation (trials)
		Focus groups
		Evaluation questions
		Expert reappraisal
			n=5...50, i=5...30
			i=1×5...3×10
			n=5...50,i=5...30
			e=3

4. Quantitative pilot study: final questionnaire and data collection design	*Response* (unit, item)
	Key *outcomes*
	Key procedures
	Data *processing*
		Analysis outcomes
		Experiments
		Evaluation questions
		'Other monitoring tools', e.g., focus groups
			n=25...250
			n=25...250
			n(i)=25...250
			n=sample size, i=1×5...3×10

5. Implementation: quality/efficiency	Get survey going, with required cost/benefit	'Monitoring tools'	n=sample size

* Approximate test size range, *if a tool is used*, for **v =volunteer respondents** or **n=sample respondents**; **e=expert** (cf. note 3); **i=interviewers**; **u=users, e.g. statisticians or researchers**.

** One may need **experts** in topical background, questionnaire design (cognition, social encounter), fieldwork.

- *Step 2G: Qualitative content test of the global questionnaire:* consultation of ex-patients of the sponsor hospital (5 focus groups; v≈5×5) and of hospital personnel (1 focus group; e=1×7).
- *Step 3G: Qualitative operational test of the global questionnaire:* field trials in three hospital departments (n≈100, n≈50, n≈50).
- *Steps 2D&3D: Qualitative operational and content test of the detailed questionnaire:* in-depth interviews with ex-patients at their home (v=7).
- *Step 4G: Quantitative pilot study for the global questionnaire* (n=704).

3. Patient satisfaction: 'What is relevant and measurable?'

Three pretesting steps were taken before prototype questionnaires were constructed in Step 1.

3.1. Commitment of hospital personnel (Step 1a). Even the best questionnaire doesn't stand a chance if the hospital personnel does not fully support the idea of measuring their (ex-)patients' satisfaction with the care received. One way to secure this commitment is to give people a say in the development of the instrument. A number of medical specialists, representatives of nurses and other caregivers were asked about their opinion on measuring patient satisfaction, what they considered relevant for the purpose, and how they thought measurement could best be carried out. For reasons of privacy, we chose to conduct in-depth interviews, with one or two persons at a time.

3.2. Commitment of patients (Step 1b). Step 1a provoked everybody to come up with his or her own favorite problem area, the link to patient satisfaction sometimes being far-fetched. The two focus groups with ex-patients, step 1b, helped to narrow down the list of about 100 problem areas resulting from 1a. Respondents were also asked about their willingness to fill in a satisfaction questionnaire in future, should they have new hospital experience. Several of them would definitely not do so, because they doubted their hospital's willingness and capacity for change. Some feared that, on the contrary, complaints would deteriorate medical care given to them in the future. These findings were confirmed by subsequent desk research (Spangenberg, 1983; Wendte, 1981), but it was not until Step 4G (the quantitative pilot) that we fully came to realize their impact. Had we done so, then we would have interrupted the projected pretesting scheme by a few operational trials (Step 3G) right here, to find out about optimal conditions for eliciting response.

3.3. Desk research (Step 1c). The Dutch literature was searched for good ideas and for pitfalls to avoid. Major concerns turned out to be:

- Extremely *low response*: 65% is attainable (Meijer & Nieman, 1992) but 5% is mentioned too (Spangenberg, 1982; Meijer & Nieman, 1992). A common figure is about 20% (HZH, 1983).
- Gratitude and *social desirability* threaten honest and critical evaluation of the care received (Visser, 1988).
- Most patients have no idea by which *reference set* to judge the quality of care (Bon et al., 1992).

Literature on theories of visual presentation was studied to optimize lay-out of the self-administered questionnaire (e.g., Jenkins & Dillman, 1995).

3.4. Drafting the two questionnaires. In the false hope that we would 'solve the response issue' in Step 3G to come, the results of Steps 1a-1c were used primarily for determining survey content. The following key satisfaction dimensions were identified: A) Accommodation ('hotel' function of the hospital), B) Information, C) Emotional Support, and D) Human Interaction. Other areas, like quality of professional medical treatment, and quality of hospital buildings and compound, were considered difficult and/or less relevant to be judged by ex-patients. Some of the technical issues we adressed were as follows.

Global questionnaire length. Patients can hardly be expected to fill in questionnaires consisting of over a 100 items. We reduced the number of items to about 30. To this purpose, we deliberately formulated some possibly double-barreled questions in the global questionnaire, like 'How was the information you received about your examination and treatment?" Any ambiguity could be solved once the detailed questionnaire would be used for further investigation. This questionnaire covered the same dimensions as the global one, but went into much more detail. For example, the single question in the global list about information on medical treatment and examination was expanded to 51 separate questions in the detailed list, which also contained a number of open follow-up questions.

Specifying the reference set. A typical example of a 'global question' is:

> 'Doctors and nurses should treat patients with respect. This means that they should be friendly and honest, and that they should pay attention

to you, listen to you, and respect your point of view. In this sense, how did you fare with your *doctor* this time?
0 All right 0 Should be slightly better 0 Should be rather better 0 Should be much better
0 Should be very much better".

This formulation was preferred over a simple one like 'Did the *doctor* treat you with respect this time?" , because it guides judgment by stating a standard of good practice and giving a detailed description of 'respect', the target of judgment. The unipolar scale was an attempt to reduce the impact of social desirablity in judgments on the quality of hospital care.

4. Patient satisfaction: 'What data are patients willing and able to provide?'

For the qualitative content test (Step 2G) of the global questionnaire prototype, it was considered imperative to recruit volunteer respondents known as recent ex-patients of the sponsor hospital, and to consult them there. The prototype global questionnaire varied according to the type of hospital care received (outpatients, inpatients at shortstay or longstay) and according to age: under a certain age, hospitalized children's satisfaction had to be measured through proxy-reporting by their parents. Focus groups appeared to be a more practical option for step 2G than one-to-one cognitive interviews. By contrast, the projected nature of the face-to-face interviews using the detailed questionnaire seemed to require in-depth interviews at ex-patients' homes (Steps 2D&3D combined). In this questionnaire, variations in hospital experience required a complex routing and detailed questions, to such an extent that it would be hard to find a group of people able and willing to discuss one single list of questions.

4.1. Qualitative content test of the global questionnaire (Step 2G). The global questionnaire was tested in 5 focus groups with various patient groups, and one focusgroup with hospital personnel. Each focus group started with every participant filling in a questionnaire. The discussion topics concerned question comprehension and relevance, as well as question wording and answer format, the aim being to get hints for improving 'cognitive and emotional respondent-friendliness' of the questionnaire.

As a result of the focus groups, lay-out was changed, some questions were added, some were eliminated, wording was changed (e.g., 'your doctor' rather than 'your specialist'), and everywhere generic N.A. answer categories ('does Not Apply') were replaced by one or more explicit ones (e.g., 'N.A, I only had one doctor'). This was because generic N.A.'s were confused with don't know's.

4.2. Qualitative operational test of the global questionnaire (Step 3G). The goal of the operational test was to to gain insight into

- *the logistic requirements* of introducing this instrument in the hospital on a continuous basis,
- *response willingness* of recent ex-patients on the basis of 'face-to-face distribution' and 'mail return',
- *plausibility of test outcomes*, e.g., in the sense of meaningful variation over doctors or departments.

The pilot was to run on three wards during one month. Longstay patients were to receive their form the night before discharge and could return it in a box on the ward. Shortstay patients and outpatients received the form at the end of their visit, from the specialists' secretary or from a shortstay nurse, and could return it in a post-free envelope. The only sampling criterion was that the hospital stay or visit was the first one for the disease or complaint involved. The expected sample consisted of 200 patients. All those responsible for distributing the questionnaires were trained by the information officer in what to say and do. Questionnaires were coded, so as to be able to trace and remind non-respondents. The hospital decided, however, not to remind non-respondents out of fear that patients would feel anonymity was threatened.

The operational test was rather succesfull as far as outpatients were concerned, with a response of 71%. For another ward, however, response was only 30%. On the third ward, only six of the expected 60 forms were distributed. Nurses found patients too sick to be bothered. Generally checking the sampling criterion took far too much time. In some cases, much effort was needed to convince patients to accept the form. Some of the hospital personnel developed a hostile attitude towards the test and its aims. It was decided that the sampling criterion had to be simplified, and that the quantitative pilot (Step 4G) had to be used also for trying out better procedures to improve response rates, especially on the inpatient wards.

Although the main goal of Step 3 was to try out procedure, it had considerable impact on content as well. Some outcomes for the outpatients urged redesign of certain questions. For instance, satisfaction about 'privacy of conversations with the doctor' was almost exclusively rated as 'good', in contradiction with strong opinions put forward in the focus groups (Step 2G) and in-depth interviews (Step 2D).

4.3. Qualitative operational and content test of the detailed questionnaire (Steps 2D&3D). The format for the detailed part of the survey turned out to be wholly inappropriate. In the in-depth interviews, people were willing to tell about their experiences in great detail, but only for a few items relevant to them. The rest was a burden to the ex-patients, who were certainly not willing to answer 51 questions about their satisfaction with Information. The questions themselves could, of course, not be tested at all at this stage.

4.4. Redesign of survey instrument and remaining pretest scheme. The failure of the operational tests (step 3G partly and step 3D in full) meant we had to rethink the concept of the global and detailed questionnaires and to adapt key procedures accordingly. It was decided to design a short and a long module for every domain of patient satisfaction, and to have the global questionnaire contain precisely one long module, each time a different one. Detailed interviews would be open interviews, to be conducted occasionnally by specially trained hospital staff.

In reviewing the operational test 3G, the hospital expressed new concerns and wishes about questionnaire content. The hospital proposed that the global questions would be revised by

- leaving out the introduction and the examples to each question, so as to shorten the global list and enhance response rates,
- changing the answering format from qualitative ('should be improved') to quantitative (ratings 1...10).

The hospital agreed to have such changes tested in a series of cognitive interviews (v=15), rather than in the second operational test (*4G*), which was aimed to improve logistics and response rates and did not seem suitable for an experiment with question formulations.

Thus Step 2 (qualitative content test) was revisited. The revised questionnaire used in the cognitive interviews, held at CBS, consisted of 'long' and 'short' questions alternated. There were two complementary versions of the questionnaire: some volunteers had the long version of a question where others had the short one, which consisted of the core of the original question, cf. Subsection 3.4. The issues and findings were as follows:

- *Does leaving out the introductory text of a question make the reference set too vague? Do the examples given in the questions help or hinder?* Retrospective think-aloud protocols were used to determine the extent to which the examples given were used in people's ratings, and if other examples were thought of. By further probes ("What do you feel the word 'respect' means in this question?"), descriptions of key concepts were investigated. In most cases it turned out that respondents indeed needed specific guidance for making judgments. Without introduction and examples, some of them would focus on less relevant aspects of the issue at hand and ignore other important issues. In the 'respect' example, the short version prompted requests for clarification like "Do you mean if he addressed me formally ('U') or not ('jij')?" The examples given did not hinder people in thinking about other examples relevant to them.
- *What do people prefer, rating their satisfaction on an ordinal level or on a numerical scale*? By the simple technique of asking "what they preferred and why", the ordinal categories came out as the absolute winner: "We are not doing math exercises here, were are judging people!"; no one preferred numbers.
- *How do people cope with double-barreled questions?* 'Thinking aloud' and probing were used to determine the extent to which people noticed the double-barreledness and how they handled it. It turned out that no single respondent had problems, presumably because most double questions referred to related issues (e.g., information about medical treatment and examination). In some cases people did not really have to choose, because they had a common experience with both issues at hand. In other questions, people reflected on the issue most relevant to them. In a few cases issues within one question were considered unrelated. Such a question had to be split up into two separate ones.

4.5. A qualitative operational test repeated within a quantitative pilot. The second operational pilot was to run on 6 different wards; 3 for adults and 3 for children. The number of patients involved was 704. A number of changes was made vis-a-vis the first pilot:

- On each ward, one person became responsible for distribution of the questionnaires.
- Except for the outpatients, the claim that patients had to be first-time visiters was dropped.
- In the first pilot the specialists' secretaries had handled the distribution of the outpatients' questionnaires. A drawback was that the secretaries were reluctant to hand out the questionnaire to someone just after a possibly troublesome visit to the doctor. In the second pilot, reception counter personnel was asked to distribute the forms, before rather than after the appointment with the doctor.

The results of the second pilot were disastrous. Of the 704 lists that should have been distributed, only

178 (25%) were actually distributed. Of these, 93 were returned. Overall response: 13%. Response ranged from 67% for children shortstay to 4% for the outpatients (who, remarkably, scored 71% in the first pilot). These findings may reflect the observation that the succes of the distribution depends not so much upon the method used, but upon the enthousiasm of those responsible (Bon, Buis, et al ., 1992). As we are back at Step 1, it may be that the hospital will have to change the mode of data collection altogether (not so much its content)!

5. Conclusions: the manageable but unpredictable dynamics of pretests

Sudman, Bradburn & Schwarz (1996) discuss the pros and cons of various "methods for determining cognitive processes and questionnaire problems in surveys". These authors make some comments about the situation where researchers apply such methods collaboratively in a so-called 'cognitive laboratory'. They consider "what distinguishes a cognitive laboratory" to be "the theoretical perspective adopted by the researchers and the use of the range of procedures to develop and pretest questionnaires". Knowledge about cognitive processes provides methodology 'from the inside out', so to say, giving guidance about *the how and why of questionnaire testing procedures* and *what contributions to questionnaire improvement to expect from these.*

The Model of Questionnaire Development represented by Scheme 1 considers pretesting programs 'from the outside in', that is in the context of (continuous) improvement of survey quality and efficiency. This means, among other things, that various options may compete with one another for limited time and budget, such as

- testing 'social encounter' aspects of questionnaires,
- testing cognitive aspects of questionnaires,
- testing data collection operations,
- problem exploration and problem prevention;
- evaluation of measurement quality and efficiency.

One of the problems of 'the cognitive testing paradigm' of Sudman et al. (1996) is that *the weight of various options for pretesting programs cannot be determined solely from the cognitive perspective*. This is simply because of the wider range of issues involved. The Pretesting Model that is the core of Scheme 1 can be considered part of some 'Survey Development Model'. The focus of the pretesting part is the collection of *respondent-related 'meta-information'* in addition to just the data needed for the survey in question. Generally, such meta-information is not of an exclusively cognitive nature: it is *information on what respondents think, do and feel while responding to the data requests imposed by the survey.*

Another important issue is that pretesting programs tend to develop dynamically, the next pretesting step being composed on the basis of both its 'inside qualities' and results from previous steps. *Each step of the Model (Scheme 1) may appear more than once in a pretesting program, in an earlier or later stage, or not at all. Overlap between steps is quite common,* e.g. because issues of procedure and content are strongly related. This asks for flexibility in allocating resources to the various phases of questionnaire design and development: one pretesting step, e.g., may be *post*poned for the benefit of repeating another one. All this appears to confirm the practical recommendation of Sudman, Bradburn and Schwarz (1996, p. 258) to "distrust any general recipes": testing steps can be planned ahead but both their order and content may require changes as need arises.

6. References

Bon, P.L.M., Buys, M.M.A., Heel, M.E.v., Kleingeld, P., Pas, F.G.E.M.v.d., Santen, W. v. & Tellegen, B. (1992). *Werken aan patiëntvriendelijkheid van algemene ziekenhuizen. Ontwikkelen van een kwaliteits-instrument.* Amsterdam : Bakkenist Management Consultants, Utrecht: Landelijk Patiënten/ Consumenten Platform.

Jenkins, C.R., & Dillman, D.A. (1995). Towards a theory of self-administered questionnaire design. In L. Lyberg et al (Eds.) *Survey Measurement and Process Quality.* New York: Wiley.

Meijer, W., & Nieman, F. (1992). *Overzicht van de reeds in Nederlandse ziekenhuizen gehouden patiënten-enquêtes.* Stafbureau Patiëntenzorg, azM.

NZI nieuws (1983). *Het ziekenhuis, 21,*902.

Spangenberg, F. (1983). Patiëntenenquêtes van ziekenhuizen geven te riant beeld. *Het Ziekenhuis,18,* 748-752.

Sudman, S., Bradburn, N.MN., & Schwarz. N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology.* San Fransisco: Jossey-Bass.

Visser, A. Ph. (1988). *Onderzoek naar de tevredenheid van ziekenhuispatiënten.* (Research into the satisfaction of hospital patients). De Tijdstroom, Lochem.

Wendte, J.F. (1981). Patiënten oordelen over ziekenhuizen. *Tijdschrift voor Sociale Geneeskunde, 59,* 102-106.