Survey Research Methods (2008)
Vol.2 , No.3 , pp. 143-150
ISSN 1864-3361
http://www.surveymethods.org

The Three-Step Test-Interview (TSTI): An observation-based method for pretesting self-completion questionnaires

Tony Hak
Erasmus University, The Netherlands

Kees van der Veer
VU University Amsterdam, The Netherlands

Harrie Jansen
Addiction Research Institute, The Netherlands

Three-Step Test-Interview (TSTI) is a method for pretesting a self-completion questionnaire by first observing actual instances of interaction between the instrument and respondents (the response process) before exploring the reasons for this behavior. The TSTI consists of the following three steps:

1. (Respondent-driven) observation of response behavior.
2. (Interviewer-driven) follow-up probing aimed at remedying gaps in observational data.
3. (Interviewer-driven) debriefing aimed at eliciting experiences and opinions.

We describe the aims and the techniques of these three steps, and then discuss pilot studies in which we tested the feasibility and the productivity of the TSTI by applying it in testing three rather different types of questionnaires. In the first study, the quality of a set of questions about alcohol consumption was assessed. The TSTI proved to be productive in identifying problems that resulted from a mismatch between the 'theory' underlying the questions on the one hand, and features of a respondent's actual behavior and biography on the other hand. In the second pilot study, Dutch and Norwegian versions of an attitude scale, the 20-item Illegal Aliens Scale, were tested. The TSTI appeared to be productive in identifying problems that resulted from different 'response strategies'. In the third pilot, a two-year longitudinal study, the TSTI appeared to be an effective method for documenting processes of 'response shift' in repeated measurements of health-related Quality of Life (QoL).

Keywords: cognitive pre-testing, self-administered questionnaire, think aloud, protocol analysis, observation

Cognitive process of answering questions

Increasingly, non-sampling data error in surveys is analyzed as resulting from problems that might occur in the response process, i. e. the process of interaction between the instrument (questionnaire) and the respondent. This response process has been described by Tourangeau (1984) as consisting of four main 'cognitive' steps, namely:

1. *Comprehension*: Understanding the meaning of the question.
2. *Retrieval*: Gathering relevant information, usually from memory.
3. *Judgment*: Assessing the adequacy of retrieved information relative to the meaning of the question.
4. *Communication*: Reporting the response to the question, e.g. , selecting the response category, editing the response for desirability, etc.

Contact information: Tony Hak, Rotterdam School of Management, Erasmus University P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: thak@rsm.nl

This model can be applied to the interaction between the respondent and the questionnaire as a whole or to parts of this process such as the respondent's response to specific sections of the instrument (such as multi-item scales) or to separate questions. When a respondent responds to a questionnaire, a problem may arise at any step in the process (as defined in this model) at any point in the completion of the questionnaire. When such a problem occurs, data error might or might not result. Cognitive interviewing has become increasingly important as a method for identifying such problems and for providing the surveyor with information about their likely causes and effects (in terms of data error).

Cognitive interviewing

In current pretesting practice the term 'cognitive interviewing' usually refers to an interview in which an interviewer asks the questions of a questionnaire and the respondent answers them, just like in a regular survey interview. However, because the primary goal of the cognitive interview is not to get these answers but to get a better idea of how the questions are working, two additional tasks might be required from the respondent. One of these is to *think aloud* ("Just tell me everything you are thinking about as you go about answering") and the other is to answer ques-

tions (*'probes'*) about the terms or phrases in the questionnaire as well as about the meaning of its questions. This is the format used in the example of a cognitive testing protocol given in Appendix 1 in Willis (2005). This protocol also clearly illustrates the main difference between the traditional cognitive interviews and respondent debriefings, which is the timing of the probes. In the cognitive interview usually one question or set of questions is asked and answered, followed by probes, before moving to a next question, whereas in debriefing usually the whole questionnaire is answered before the respondent is asked questions about the meaning of terms and about the response process.

Think aloud (or, more precisely, saying aloud what you are thinking) was developed and is used by (cognitive) psychologists as a technique for producing data about the process of thinking (Ericsson and Simon 1980; Van Someren et al. 1994). Its aim is to make this process, which normally is hidden, observable by asking respondents to verbalize their thoughts *concurrently*, i. e. at the very moment they think them. It is debatable whether this can be done at all without changing the process of thinking and these thoughts themselves. But it is important to recognize that this is the aim of the think aloud technique and that, even if imperfect, it is the most proximate account of the cognitive response process attainable and the best approximation to an 'observation' of that process. *Verbal probing*, on the other hand, is a technique for eliciting *reports* from respondents *about* their thinking. As soon as we start probing, the nature of the data is changed from a respondent-driven to an interviewer-driven account.

In the pretesting literature the distinction between data from observation of response behavior and data from think aloud is not always clearly made. Neither is it noticed that observation of self-completion response behavior is impossible if the questionnaire is presented in an interview mode. In this article we present the Three-Step Test-Interview (TSTI) as a procedure for the pretesting of *self-completion questionnaires* (only) which starts by collecting respondent-driven information about behavior (observation) and then in two subsequent steps moves to a more interviewer-driven exploration of factors affecting this behavior. By adopting this approach we reduce the risks of respondent's inventing problems as a result of interviewer-probing.

The Three-Step Test-Interview (TSTI)

As explained above, the aim of the Three-Step Test-Interview (TSTI) is to produce observational data on actual response behavior of respondents who respond to self-completion questionnaires. Because much of this behavior consists of 'thinking' and is therefore hidden from the observer, the *(concurrent) think aloud* technique is used for making it observable. Therefore, the first and main step of the TSTI is:

1. *Observation of response behavior and concurrent think aloud verbalization, aimed at collecting primary data on problems encountered in responding to the questions to be tested.*

Two additional steps follow upon this obervation/think aloud step:

2. *Follow-up probing aimed at remedying gaps in observational data and in the think aloud account.*
3. *Debriefing aimed at eliciting experiences and opinions.*

Steps 2 and 3 are not only additional in a chronological sense - i. e. they follow the first step - but also in a methodological sense: these data illuminate, illustrate and explore the principal data, the observational ones that are collected in the first step. In the following section we will first describe in more detail the aims and the techniques of the three steps of the TSTI, and will then illustrate how we developed and standardized this technique in three pilot studies.

Step 1. Observation of response behavior and of concurrent verbalization.

The aim of the first step of the TSTI is twofold:

1. *Collecting observational data regarding the respondent's behavior (such as skipping questions; correction of the chosen response category; hesitation; distress; etc.) by concurrently taking notes.*
2. *Collecting observational think aloud data by listening to verbal expressions of respondents' thoughts while responding to the questionnaire.*

Obviously, respondents must 'produce' the required behavior for observation. For that purpose, respondents are instructed to complete the questionnaire as they would do in real-life, with the additional task to concurrently verbalize what they are thinking. Ideally, both types of data - actions and verbalizations - are recorded and kept on audio and videotape for later analysis. Because this first step of the TSTI must replicate the structure and format of the 'real' survey situation, it is important that its strictly observational nature is not compromised by any intervention - such as a question, comment, probe - by the researcher that might suggest that a self-report from the respondent is required.

Respondents differ considerably in the degree to which they are able to perform the think aloud (or rather "say aloud what you think") task. Some find it very difficult. Usually such difficulties are not the result of a lack of understanding of what is requested (just saying aloud thoughts) but rather of not being used to it. Therefore, it is useful to start the interview with a short exercise in thinking aloud and to encourage the respondent during moments of silence by statements such as "Please say aloud what you think". (These and other detailed instructions for conducting the TSTI are available from the Internet. See the endnote to this article.)

Step 2. Follow-up probing aimed at clarifying and completing primary data.

In this step the observer only considers those actions or thoughts that he has observed (in step one) about which he feels not fully informed, in order to fill in gaps in the observational data or to check information (e.g. "Did I hear you say…?" or "You stopped for a while there, what did you

think?"). The observer must rely on the observations (notes) made during the first step. The main methodological criterion (and also technically the most difficult aspect of this step for both respondents and 'test-interviewers') is that respondents should only report about what they did and thought in the first step, not about what they think now (in retrospect). It needs to be very clear that the aim of this step is not to elicit accounts, comments, etc., that were not already thought during step 1.

Step 3. Debriefing aimed at eliciting experiences and opinions.

Steps 1 and 2 of the TSTI result in two types of primary data, regarding actions and thoughts, which have been recorded in two ways, on tape for later analysis and in the form of real time notes by the researcher for use in the interview itself. The final step, which now follows, is the only one in the TSTI in which the respondent is 'allowed' and even stimulated to add secondary data - accounts and reports of feelings, explanations, preferences, etc. - to the primary ones. In our pilot studies, reported below, this third step took very different forms depending on the kind of questionnaire that was pretested, but three main forms (and corresponding aims) can be distinguished:

a) Respondents might (be requested to) 'explain' their response behavior. Particularly when specific problems were encountered in responding to the questionnaire, respondents could comment on what they thought the exact nature of the problem was and why they behaved as they did - which was recorded in steps 1 and 2 of the 'interview'. Also, respondents might suggest improvements in terms of wording of questions, layout of the questionnaire, instructions, etc. It is important to acknowledge that these kinds of comments constitute 'opinions' or informal 'hypotheses', not facts, regarding the causes of problems detected in steps 1 and 2. Researchers must make their own analysis of problems associated with the questionnaire (based on observations in all interviews of the pretest).

b) Respondents might be asked to paraphrase questions and to formulate definitions that they (implicitly or explicitly) used when answering the questions.

c) Respondents might be probed about the substantive issues that are covered by the questionnaire that is tested. For instance, if an alcohol consumption questionnaire is tested, respondents might be invited to describe their alcohol consumption in their own words. Or, if a scale for the measurement of attitudes towards 'illegal aliens' is tested, respondents might be asked to explain these attitudes in their own words. In our pilot studies (see below) it appeared that such data from this part of the TSTI, when compared to respondents' responses to the questionnaire in step 1, were useful as indicators of the validity of the data collected by the instrument.

The forms (a) and (b) of this third step are very similar to usual interviewer-driven techniques of 'cognitive interviewing'. In the TSTI these techniques are used to elicit information that is additional to previously collected respondent-driven information (including think aloud). The Three-Step Test-Interview, thus, is a sequence of (a) respondent-driven observation, (b) interviewer-driven retrieval of additional data by follow-up probing, and (c) validation by interviewer-driven debriefing.

First pilot study: a test of questions on behavior

The objective of this pilot study (Jansen and Hak 2005) was to test a set of six questions, used in self-completion questionnaires on alcohol consumption, collectively known as a Quantity-Frequency-Variability measurement (QFV). This specific set of questions was not new but had been used in the Netherlands since the beginning of the 1980s in several surveys on health related behavior and it was planned to be applied in future studies as well. Internationally, research on alcohol consumption has an established record of discussions on methodological aspects of different ways of measuring this consumption, which is partially grounded in cognitive research. Informed by that literature, we started our study with a close reading of the latest version of this QFV questionnaire and, then, discussed our findings with the authors of this particular version. We concluded this desk *expert review* with a set of expectations or hypotheses regarding the problems that might be encountered by respondents when answering these questions. We thought that this list of predictions based on previous research would be a strong test for the TSTI. Our criteria for a successful test of the TSTI in this pilot study were that it should detect

a) all problems with these specific questions that are described in the literature on the measurement of alcohol consumption or were mentioned by experts; and *additionally*

b) a number of relevant problems not mentioned in the literature or by experts.

In the analysis of the protocols (transcripts) from steps 1 and 2 in our interviews it appeared that the TSTI identified almost all problems that could be expected on the basis of the literature review (see Jansen and Hak 2005, for details). These were the ones that appeared to originate from the complexity of the tasks implied by specific question formats, such as problems related to interpretation or computation, or by inconsistencies between questions. But we found a number of other problems that were not predicted. Most of these problems seemed to arise from a mismatch between the 'theory' (on 'normal' or 'standard' patterns of alcohol consumption) that underlies the questions and the ('non-standard') lifestyles, biographies or other peculiarities of respondents.

Examples are respondents with shift work whose drinking pattern follows the rhythm of their shifts, respondents who get tipsy when drinking small amounts, respondents who recently changed their drinking habits, or respondents who have just returned from a binge-drinking holiday. For all of such respondents, the tasks imposed on them by the questions did not allow them to account for their specific (changes in) circumstances, resulting in invalid responses. An example is given in Box 1.

Box 1. Example from alcohol consumption questionnaire

Question:
How often did you drink six or more glasses on one day, during the last six months?
Response categories ranging from (1) 'every day' until (8) 'never' and (9) 'don't know' (one answer permitted).
The expert review did not predict problems with the response categories, but some appeared during the TSTI.

Step 1: Observation of response behavior
R9 marks two response categories: 3 (3 or 4 *times a week*) and 9 (*don't know*)

Step 2: Follow-up probing
I: So it is about three or four times a week you drink six glasses or more?
R9: [There] May also [be] a week that I don't drink ... [you] can take also a week that six [times]...
I: You also marked "don't know"
R9: Well, the one time three and the other time nothing

Step 3: Debriefing
It appears that this respondent is a shift worker at Heineken brewery (!). He only drinks alcohol in weeks (one in four) in which he does not work. In such weeks he often drinks more than 6 glasses of beer a day. But that varies a lot too. In some weeks he might drink alcohol on 3 or 4 days, in other weeks 5 or more.

Conclusion
The respondent wants to express the variability of his drinking behavior in his response.

Mismatches as illustrated in Box 1 were *discovered* (identified) in steps 1 and 2 of the TSTI and could be *interpreted* by the exploration of these respondents' lifestyles and consumption patterns expressed by them in step 3 of the TSTI. We summarized our findings in tables such as Table 1.

Table 1 shows for one question which problems were identified in the expert review and which were identified in TSTI, as well as their overlap. The TSTI identified almost all (6 out of 7) problems that were identified in the expert review for this specific question, and additionally found three other problems. These three problems were first observed in step 1 of the TSTI (as illustrated with the example in Box 1). We found similar patterns for the other questions. We conclude that, regarding this specific set of questions, it is the combination of (a) observation of response behavior and think aloud with (b) biographical exploration that made TSTI productive.

The results of this pilot study were specific for the questionnaire that was tested. This questionnaire was aimed at measuring behavior and involved complex tasks such as the identification of pertinent information in memory and computation. Whereas problems regarding retrieval and computation already had been identified in the expert review, the TSTI was productive in detecting additional complications resulting from unusual drinking patterns - or rather patterns unforeseen by researchers.

Second pilot study: a test of an attitude scale

In the second pilot study, Dutch and Norwegian versions of an attitude scale, the Illegal Aliens[1] Scale, were tested. This scale consists of 20-items and is meant to be used in a self-completion mode (Hak et al. 2006; Van der Veer et al. 2002). Responding to an attitude scale is an activity that differs considerably from answering questions about behavior, such as one's alcohol consumption. We were interested in finding out whether the TSTI would be equally productive in discovering problems with respect to an attitude scale and which kind of problems would be found. As in the first pilot study, TSTI results were evaluated against an *expert review*. Our criterion for success was, as in the other pilot, that the TSTI

a) would detect the problems predicted by the expert review, and *additionally*
b) would identify other problems.

The Illegal Aliens (IA) Scale (Ommundsen and Larsen 1997) is an attitude scale, consisting of 20 parallel interval items. Each item consists of a statement about, or related to 'illegal aliens' (e.g. "Illegal aliens cost The Netherlands/Norway millions of [currency] each year" and "Illegal aliens provide The Netherlands/Norway with a valuable human resource"), followed by five response categories:

Agree strongly	1
Agree	2
Uncertain	3
Disagree	4
Disagree strongly	5

Although the IA Scale was developed for use in large sample comparative studies of political and ideological attitudes, e.g., between several groups within populations or between countries, it was initially only used in student samples. For the purpose of comparative studies between countries, the IA Scale was translated into Norwegian, Danish and Dutch, and subjected to a series of validation studies (see Ommundsen et al. 2002). The specific aims of our study were, first, to describe the range of possible interpretations of the items of the scale by Norwegian and Dutch students and, secondly, to explore possible reasons for observed differences in interpretation. Two convenience samples were recruited, one consisting of six undergraduate students in the social sciences at the VU University Amsterdam and the other of eight students in psychology at the University of Oslo.

As in the first pilot study, the TSTI study replicated almost all problems that were identified by the expert review, such as problems regarding the meaning of concepts in the questions, the ambiguous wording of some questions, and

[1] The term 'Illegal Aliens' is an everyday phrase in american English. In Europe the term 'Illegal Immigrant' is used.

Table 1: A comparison of results from the expert review and TSTI for Question 2 on alcohol consumption (from Jansen en Hak, 2005)

Question2.	expert review	field TSTI
Research has shown that a considerable part of the population drink six glasses or more of alcoholic beverages on one day, more or less regularly. ①		− ① this introduction works adversely with R7
In the past six months ②, did you ever drink six or more ③ glasses ④ of alcoholic beverages ⑤ on one day ⑥ ?	+ ② R's probably will not assess the period of 'six months' accurately + ② especially R's with high drinking frequencies probably will extrapolate from last week to six months	+ ② R4 and R7 assess the past half year accurately; others don't. − ② R4, R8, R9 and R13 experience difficulties in calculating a mean because their drinking does not fit into this time schedule
1 () Yes, every day 2 () Yes, 5 or 6 times a week 3 () Yes, 3 or 4 times a week 4 () Yes, 1 or 2 times a week ⑦ 5 () Yes, 1 to 3 times a month 6 () Yes, 3 to 5 times this half year 7 () Yes, 1 or 2 times this half year 8 () No, never this half year 9 () Don't know	+ ③ not every R counts glasses when drinking + ③ 'six glasses' may be taken as drinking a lot + ④ it is known that people at home often drink beer from bottles and other drinks from non-standard glasses + ⑤ as a selection effect from question 1 some R's will not count all types of alcoholic drink − ⑥ some R's probably count drinks by occasion rather than by day	+ ③ R7 takes this as a question for occasions of drinking sigificantly more than average – in her case still far below six glasses +④ R8, R15 and R16 count bottles or big glasses + ⑤ R8 asks whether this concerns beer only or other drinks as well − ⑦ leading categories: R7 prefers '2 to 3' which is missing. R16 unjustly marks the first 'yes' which means every day

An identified problem is marked with '+' if it also was identified in the expert review and with '−' if it was not revealed in the other study

the meaning of the response category uncertain (which could mean both uncertainty about the meaning of the item and 'no opinion') (see Hak et al. 2006). Additionally, the TSTI identified other, unpredicted problems related to the interpretation of items. Our main finding was that several respondents felt as if 'forced' to make a choice between two possible 'readings' of a number of items. Take the following example in Box 2.

In step 1 (concurrent think aloud), the respondent recognizes the item as one in which a difference is constructed between 'legal' and 'illegal' aliens. He disagrees with this distinction. His selection of the response category *uncertain* can be seen as expressing an avoidance to make this distinction here and, thus, to take sides for or against 'illegal' aliens. In step 2 the respondent states that he is genuinely uncertain whether illegal aliens actually are a valuable human resource in the economy. In step 3 (debriefing) this respondent stated that he was aware of the fact that the authors of the IA scale expected him to express his hostile or friendly attitude towards aliens by taking sides rather than to evaluate the 'literal' or 'factual' contents of the item. He confirmed that he knew that the IA score resulting from his 'literal' response was less 'friendly' towards illegal aliens than it would have been if he had behaved according to the expectations of the authors of the questionnaire. He described himself as someone who tends to "interpret everything always very literally". This self-description explains how the wording of the items of this questionnaire had made it possible for this respondent to find a lack of clarity in many items and to justify a 'literal' reading of them.

At the end of the third TSTI step, the respondent's strategy and its implications were explicitly discussed with him (see Box 2; step 3). The phenomenon described here, regarding the availability of two different 'readings' of the items and the resulting arbitrariness, as experienced by respondents, of having to make a choice between them, occurred in several TSTIs, both in the Netherlands and in Norway. Our conclusion was that there might be a problem with the IA Scale in the sense that, due to this phenomenon, 'friendly' attitudes to aliens might be underrepresented in IA Scale results.

In sum, as in the first pilot study, the TSTI both replicated the results from an expert review and, additionally, detected other problems that were not predicted by the expert review. TSTI results showed more exactly what different respondents actually do when they complete the IA Scale and, therefore, it offered a more comprehensive diagnosis of the questionnaire as a whole.

If we compare the results of this study with the first pilot study, we notice a similarity and a difference. The similarity is that in both cases the TSTI appears to be productive in identifying problems that arise from the biographical, cultural, or political context in which the questionnaire is completed. The main difference between the two studies regards the function of the third step of the TSTI (debriefing). Interviewers and respondents in the study of alcohol consumption questions used this third step for an exploration of the 'facts' of a respondent's drinking behavior. In the study of the IA

Box 2. Example from Illegal Alien scale

Item:
Illegal aliens provide the Netherlands with a valuable human resource

Step 1: Observation of response behavior
R: illegal aliens provide the Netherlands with a valuable human resource . . . I immediately think, in my opinion it doesn't make a difference whether you are legal or illegal, to be a valuable human resource, so, well, I have not, a straightforward opinion . . . so, it is uncertain . . . because one can be valuable also if one is legal.

Step 2. Follow-up probing
I: okay . . . erm . . . and why . . . why have you
R: chosen the 2 (*agree*) rather than 1 (*agree strongly*)?
I: yes . . .
R: well . . . I think, I lack knowledge a bit . . . I have uncertain . . . a valuable human resource, well, what can I say about it?

Step 3. Debriefing
In an extensive discussion of how he had answered this item, the respondent stated that he thought that the authors would expect him to demonstrate his friendly attitude to illegal aliens wherever possible, i.e. by reading items as invitations to position himself politically or ideologically rather than as questions about economic or social facts. He described himself as someone who tends to "interpret everything always very literally". This discussion then moved to the following fragment.

Item:
Illegal aliens cost the Netherlands millions of guilders each year

I: take the item illegal aliens cost the Netherlands millions of guilders each year . . . such an item . . . you take it literally . . . you said that a couple of millions is not much so it's very likely that illegal aliens cost us millions
R: yes
I: do you think . . . so you interpret the item as a statement about facts but is it right that you assume that the designers of the questionnaire have something else in mind?
R: yes
I: now if this questionnaire is a test in logic . . . you have performed very well on the test
R: yes (laughs)
I: but if it is true that . . . if this questionnaire aims at measuring . . . say your benevolence regarding illegal aliens . . . in that case you have been almost deliberately, deliberately
R: yes on the wrong side
I: so you mislead the researchers . . . so one can say that for that reason alone the item does not measure your . . . what your real opinion is about illegal aliens
R: no . . . not at all
I: you say . . . well it is their responsibility to . . . how they interpret the responses . . . now it is possible that . . . if they want to draw political conclusions . . . that they as you say will interpret your response incorrectly
R: yes . . . but I try to attend to what . . . as much as possible . . . to what is printed here . . . that's in principle the only thing I have.

scale it appeared that, in general, there was not much left to explore with respect to the respondents' attitudes after they had completed step 2 of the TSTI. An issue that could be explored in step 3 was, as shown above, the respondent's attitude to the questionnaire (rather than to illegal aliens). In this case the TSTI proved to be a useful tool for testing attitude questions.

Third pilot study: assessment of response shift in health related Quality of Life

After having conducted a pilot study in which the TSTI was applied to questions on (drinking) behavior, and another one applied to an attitude scale, we concluded this series of pilot studies with an application of the TSTI to questions about health related Quality of Life (QoL). Qol is neither a behavior nor an attitude. It is an evaluation of an experience (such as pain or depression) or of a situation (such as a bad prognosis). Because QoL is an *evaluation*, its measurement is dependent on the criteria that are (either implicitly or explicitly) used. These criteria tend to change over time ('response shift'). Some researchers claim that the occurrence of response shift makes measurements invalid because, at different times, a different 'concept' is measured. Other researchers claim that only the resulting (measured) QoL matters, because according to them QoL is the patient's evaluation, not the state that is evaluated. It is clear that these researchers use different concepts of what QoL is. For us, this debate became relevant when we realized that the TSTI might be able to make observable the evaluation process that results in a QoL score. We assumed that think aloud protocols would demonstrate how respondents evaluate a situation or experience, and what (shifting) criteria they apply. If this would actually be the case, this would not only be informative about the phenomenon of 'response shift', but would also give us detailed information about how QoL questions 'work'. This kind of information would contribute to at least one aim of pretesting, namely the aim to ascertain whether the question measures the intended 'concept'.

In this study (see Westerman et al. 2008), 30 lung cancer patients were sampled from different Dutch hospitals for a two-year longitudinal study in which these patients completed QoL self-completion questionnaires between two and five times. This enabled us to assess response shift over the duration of an entire illness trajectory (or at least a considerable part of it). The TSTI format was applied each time a patient in this study completed a QoL questionnaire. One of our concerns was whether old, rather sick people would be able to adhere to the think aloud technique. This proved difficult indeed, but many patients were able to endure the TSTI and useful protocols (transcripts) were generated. Interviews were conducted at the patients' homes, which is consistent with the aim of the TSTI to come as close as possible to the real-life situation in which the instrument is completed by a respondent.

It appeared that the think aloud technique is rather appropriate for QoL questions, because the evaluation of a situation, which is implied by the different QoL items, requires respondents to think, for each item again, what the relevant events and criteria are. Usually, the answer is not spontaneous but it must be constructed. In terms of Tourangeau's response model (see above): the *judgment* and *communication* steps require effort. This judgment and communication work can relatively easily be said aloud. Take the example in Box 3.

Box 3. Example from Quality of Life measurement

Questions:
❑ *Do you have any trouble taking a **long** walk?*
❑ *Do you have any trouble taking a **short** walk outside of the house?*
Response categories: (1) 'not at all'; (2) 'a little'; (3) 'quite a bit'; and (4) 'very much'.
These two questions are items from the EORTC QLQ-C30, a questionnaire that we tested at three different points in time (T1, T2, and T3) using a TSTI format.

T1. Step 1: Observation of response behavior
R: *Do you have any trouble taking a long walk?* Yeah, that must be very much . . . yeah at this moment . . . the shopping center . . . it's 450 meter . . . I cannot make it.

T2. Step 1: Observation of response behavior
R: *Do you have any trouble taking a long walk?* Yes, with a long one . . . I cannot walk kilometers.
(data omitted)
R: *Do you have any trouble taking a short walk outside of the house?* No, a short walk . . . I mean . . . 500 meter that way . . . that's nothing . . . I do it without any problem.

T3. Step 1: Observation of response behavior
R: *Do you have any trouble taking a long walk?* Yes, with a long one . . . I cannot do it . . . but I walk too fast, it's my own fault . . . I haven't tried it yet but I think quite a bit . . . I can make it to the shopping center though
(data omitted)
R: *Do you have any trouble taking a short walk outside of the house?* Well, a little.

Conclusion
The respondent's definition of a 'long' walk is a walk that is difficult to make. The response to the question about the long walk is, therefore, always 'quite a bit' or 'very much'. But at T1 this answer refers to a walk to the shopping center (450 meter), whereas at T2 it refers to a walk of kilometers. At T2 and T3 the 450-500 meter walk is considered a 'short' walk (because, at that moment, it can be walked without much trouble).

In Box 3, the think aloud protocols present a single respondent's concurrent accounts of reasoning at three different points in time (T1, T2, and T3). This respondent is a cancer patient who has received treatment. At T1 he is suffering from the side effects of treatment, whereas at T2 and T3 his condition has improved considerably. By comparing the three protocols, it is clear that a 'response shift' has occurred: different standards for what a 'long' walk is have been used,

and therefore the resulting scores refer to different kinds of 'walks'.

Such comparisons between think aloud statements (step 1), which can be supported by data collected in the steps 2 and 3, make the presence (or absence) of shifts in processes and criteria of evaluation observable. Apart from this specific use of these transcripts for the study of response shift, it allows developers and users of such instruments to assess what concept is actually measured (and how this is done in specific instances; see Westerman et al. 2008).

In sum, this third pilot study has demonstrated that the TSTI is a feasible and productive technique for producing data which are useful for the description and exploration of the manner(s) in which health related QoL questions are answered in actual instances. This allows an assessment of their validity with respect to their aims (which might differ between studies).

Conclusion

The Three-Step Test-Interview (TSTI) is a method for assessing the quality of a selfcompletion questionnaire by observing actual instances of interaction between the instrument and a respondent (the response process). Concurrent think aloud is used as an (imperfect) technique for making the thought process observable. This paper describes how the TSTI was tested in three pilot studies. In the first study, the quality of a set of questions about alcohol consumption was assessed. The TSTI proved to be particularly good at identifying problems that result from a mismatch between the 'theory' underlying the questions on the one hand, and features of a respondent's actual behavior and biography on the other hand. In the second pilot study, Dutch and Norwegian versions of an attitude scale, the 20-item Illegal Aliens Scale, were tested. The TSTI appeared to be productive in identifying problems resulting from different 'response strategies'. In the third pilot study, the TSTI appeared to be an effective method for documenting processes of 'response shift' in the measurement of health related Quality of Life (QoL). While producing this kind of additional data on the performance of specific instruments, the TSTI produces at the same time also data that are (or can be) produced with other methods. This suggests that, for self-completion questionnaires, the TSTI might replace the other methods without a significant loss of useful information. This should, however, be tested in experiments in which different methods are applied to the same instrument (as in Presser and Blair 1994; Willis et al. 1999; and Rothgeb et al. 2001).

Download material

A manual for the TSTI can be downloaded from `http://hdl.handle.net/1765/1265`.

References

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215-250.

Hak, T., van der Veer, K., & Ommundsen, R. (2006). An application of the Three-Step Test-Interview (TSTI): A validation study of the Dutch and Norwegian versions of the 'Illegal Alien Scale'. *International Journal of Social Research Methodology*, *9*, 215-227.

Jansen, H., & Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-administered Questionnaire on Alcohol Consumption. *Journal of Official Statistics*, *21*, 103-120.

Ommundsen, R., Hak, T., Mörch, S., Larsen, K. S., & van der Veer, K. (2002). Attitudes toward illegal immigration: a cross-national methodological comparison. *The Journal of Psychology*, *136*, 103-110.

Ommundsen, R., & Larsen, K. S. (1997). Attitudes toward illegal aliens: the reliability and validity of a Likert-type scale. *The Journal of Social Psychology*, *137*, 665-667.

Presser, J., & Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different results? In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 24, p. 73-104). Washington, DC: American Sociological Association.

Rothgeb, J., Willis, G., & Forsyth, B. (2001, May). *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, Montreal)

Tourangeau, R. (1984). Cognitive Science and Survey Methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive Aspects of Survey design: Building a Bridge between Disciplines* (p. 73-100). Washington, DC: National Academy Press.

van der Veer, K., Ommundsen, R., Hak, T., & Larsen, K. S. (2002). Meaning shift of items in different language versions. a cross-national validation study of the Illegal aliens Scale. *Quality and Quantity*, *37*, 193-206.

van Someren, M. W., Barnard, Y., & Sandberg, J. A. C. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.

Westerman, M., Hak, T., Sprangers, M., Groen, H., van der Wal, G., & The, A. M. (2008). Listen to their answers! Change in cancer patients' interpretations of quality of life questions. *Quality of Life Research*, *17*, 549-558.

Willis, G. B. (2005). *Cognitive interviewing. a tool for improving questionnaire design*. Thousand Oaks (CA): Sage.

Willis, G. B., Schechter, S., & Whitaker, K. (1999, August). *A comparison of cognitive interviewing, expert review, behavior coding: what do they tell us?* (Paper presented at the Annual Meeting of the American Statistical Association, Baltimore)