

The Three-Step Test-Interview (TSTI): An observational instrument for pre-testing self-completion questionnaires

Kees van der Veer, Department of Social Research Methodology, Vrije Universiteit De Boelelaan 1081c, 1081 HV Amsterdam, the Netherlands
Tel. +31 20 4446866, e-mail: vdveer@staff.scw.vu.nl

Tony Hak, U.S. Census Bureau, 4700 Silver Hill Road, Stop 6200, Washington, D.C. 20233-6200, USA
Tel. +1 301 763 7614. E-mail: thak@fbk.eur.nl

Harrie Jansen, Addiction Research Institute, Heemraadssingel 194, 3021 DM Rotterdam, the Netherlands
Tel +31.10.4253366, e-mail: jansen@ivo.nl

Paper for the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), Charleston, South Carolina, November 14-17, 2002

Abstract

The Three-Step Test-Interview (TSTI) is an instrument for pre-testing a self-completion questionnaire by *observing actual* instances of interaction between the instrument and respondents (the response process). Because this process mainly consists of cognitive processing ('thinking') and is therefore hidden from the observer, (*concurrent*) *thinking aloud* is used as a technique for making the thought process observable. The TSTI consists of the following three steps:

- (a) Concurrent thinking aloud aimed at collecting observational data.
- (b) Focused interview aimed at remedying gaps in observational data.
- (c) Semi-structured interview aimed at eliciting experiences and opinions.

As yet, the TSTI has been 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 and features of a respondent's actual behavior and biography. In the second pilot study, Dutch and Norwegian versions of an attitude scale, the 20-item Illegal Aliens Scale, were validated. The TSTI appeared to be uniquely productive in identifying problems resulting from different 'response strategies'. In the third pilot study, the TSTI appeared to be an effective instrument for producing data that document processes of 'response shift' in the measurement of health-related Quality of Life (QoL).

A manual for the application of the TSTI is appended to this article.

Introduction

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.

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 subject 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 been developed as an instrument for identifying such problems in the response process, their localization (both in the response process model and in the questionnaire), their effects (in terms of data error), and their causes.

In current pre-testing practice the term 'cognitive interviewing' refers to two main techniques, *think aloud* and *probing* (see, e.g., Willis, 1999: 3). It must be stressed that these two techniques are very different in terms of their aims and of their methodological status. Think aloud was developed and is used by (cognitive) psychologists as a technique for producing data about the process of thinking. Its aim is to make this process, that normally is hidden, observable by asking subjects 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 of the utmost importance to recognize that it is the aim of the think aloud technique to make the thinking process *itself* observable. 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 observations to self-reports. This is an important difference, which is the more pertinent when "probing is the basic technique that has increasingly come into favor by cognitive researchers" (Willis, 1999: 6).

In the pre-testing literature the distinction between observational data (i.e., the *actual thinking process* of the respondent, made observable through think aloud) and self-report data (i.e., the respondents' *accounts* of this process) is not (or insufficiently) acknowledged. This is the case even in most of the (rather rare) instances in which it is the explicit aim of the cognitive interview to collect data on the *actual* response process in specific instances. It is significant that reports on cognitive pre-testing research only rarely mention insights that are gained from observing what respondents in the research actually did when they responded to the test questionnaire, i.e. based on concurrent think aloud protocols. Instead, reported results of cognitive pre-testing

are usually based on data produced by respondents when probed to perform *other* tasks (than responding to the questionnaire) such as paraphrasing questions, explaining definitions of terms, and expressing preferences regarding wordings, layouts, etc. However, such reports *about* interpretations and preferences, produced in pre-testing, have a questionable relationship to the actual response process, which occurs when the same respondents complete the questionnaire.

The emphasis in the current practice of cognitive pre-testing on the exploration of ideas (definitions, etc.) through probing has also the effect that important differences between the two main modes of questionnaire administration (interview and self-completion) are neglected. What is tested by probing is the wording of *questions* (see, e.g., Willis, 1999: 3, 28) outside their context (a questionnaire), not the questionnaire as encountered and experienced in practice by the respondent. Thus, it is not tested how the question is understood in its actually intended form, i.e. when it is delivered (for hearing) by a specific interviewer in a specific interview context or provided (for reading) in a specific textual format for self-completion. By focusing on the response to context-free questions, such cognitive interviews do *not* test the actual performance of the instrument in the field, either in an interview or on self-completion. In many respects, reading and hearing are very different processes, and comprehension of questions and their terminology may differ accordingly. If it is our aim to assess and improve the *actual* performance of the instrument in the field (i.e. in the chosen mode), we must observe the response process in action.

Because responding in an interview is very different from self-completing a written questionnaire, the appropriate techniques for observation should be different as well. The appropriate technique of observing the response process *in an interview* is observing (i.e. audio or videotaping) the interaction in the interview. This type of observation is known under the term ‘behavior coding’ (see Fowler & Cannell, 1996; Dijkstra, 1998). However, coding (and counting) is only one way of producing and analyzing observational data, and it is not necessarily the most productive in terms of producing insights in the problems that respondents encounter. (See Maynard et al, 2002, for a qualitative approach.)

The methodological issues associated with the production of observational data regarding the respondents’ behavior *during self-completion* are different from those associated with observing the interaction between an interviewee and a questionnaire that is delivered by an interviewer. The former are very similar to the issues encountered by psychologists when they started to study the process of thinking, in response to which they developed the think aloud technique (Ericsson & Simon, 1980; Van Someren et al., 1994). They also resemble the issues that are encountered in usability testing. Therefore, the Three-Step Test-Interview is, in many respects, similar to the think aloud technique used in cognitive psychology and to the instruments used in usability testing. Nielsen, one of the pioneers on usability testing, calls thinking aloud “the single most valuable usability engineering method” (Nielsen, 1993: 195; see also Nielsen & Mack, 1994).

For testing interview questionnaires behavior coding procedures (followed by both interviewer and respondent debriefing) seem to be most adequate. Self-completion questionnaires however should be tested by means of a think aloud technique. This

implies that the questions for the think aloud session should be offered to the respondent in a written format (as would be the case in the real-life questionnaire). In this article we present the Three-Step Test-Interview (TSTI) as a technique for the pre-testing of *self-completion questionnaires only*.

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 a self-completion questionnaire. Because much of this behavior consists of ‘thinking’ and is therefore hidden from the observer, the (*concurrent*) *thinking aloud* technique is used for making it observable. Therefore, the first and main step of the TSTI is:

1. *Concurrent thinking aloud aimed at collecting observational data.*

Two additional steps follow upon this think aloud step:

2. *Focused interview aimed at remedying gaps in observational data.*

3. *Semi-structured interview aimed at eliciting experiences and opinions.*

Steps 2 and 3 are not only additional in a chronological sense -- 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 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. Concurrent thinking aloud aimed at collecting observational data

The aim of the first step of the TSTI is to collect observational data regarding the respondent’s response behavior. These data consist of two types:

(1) Observations of respondent behavior (such as skipping questions; correction of the chosen response category; hesitation; distress; etc.).

(2) Think aloud data.

Obviously, respondents must ‘produce’ the required behavior for observation. For that purpose, respondents are instructed to complete the questionnaire as they would do at home or otherwise when they would be asked to complete the questionnaire, with the additional task to concurrently verbalize what they think. (See the appendix for details of the instructions given to respondents, particularly regarding think aloud.) Ideally, both types of observational data – actions and verbalizations – are recorded and kept on audio and videotape for later analysis. But the researcher also makes ‘real time’ notes of observed behaviors as well as of verbalized thoughts that seem to be indicative of problems in the response process. These real time notes are made for immediate use in the following steps of the interview.

The strictly observational nature of this first and essential step of the TSTI must not be compromised by any intervention – such as a question, comment, probe – by the researcher that might suggest that a self-report from the respondent be required.

Step 2. Focused interview aimed at clarifying and completing observational 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 assessment of this (in)completeness must be made in real time based on the researcher’s 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 is *not* the aim of this step to elicit accounts, comments, etc.

Step 3. Semi-structured interview aimed at eliciting experiences and opinions.

Steps 1 and 2 of the TSTI result in two types of *observational* 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, observational ones. In our pilot studies, reported below, this third step took very different forms depending on the kind of questionnaire that was ‘tested’, 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, they could comment on what they thought the exact nature of the problem was and why they behaved as they did – as 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. The aims and form of this interview will be similar to those of ‘respondent debriefing’. It is important to acknowledge that this kind of comments constitute ‘opinions’, 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 pre-test).
- (b) Respondents might be asked to paraphrase questions and to comment on their definitions of terms. In other words, some form of ‘cognitive interviewing’ might be done in this stage of the TSTI.
- (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 interview, 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 may seem very similar to usual formats of ‘cognitive interviewing’. It must, however, be emphasized that data collected in this step 3 of the TSTI have a very different status from those that are collected in such usual cognitive interviews. The main difference is that they are collected additionally (and secondarily) to other (primary, observational) data. They are elicited as aids in the analysis of those primary data on *actual* response behavior (and on actual problems that occur) rather than as primary data about (by definition) *potential* problems. Therefore, it should be noted that, in the strict sense of the term, the TSTI is *not* an interview. Rather it is a sequence of (a) observation, (b) follow-up probing and (c) validation.

First pilot study: a test of questions on behavior

The object of this pilot study (Jansen & Hak, 2002) was a set of six questions 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 that, according to the literature on the measurement of alcohol consumption, are known to occur with these specific questions; *and additionally*
- (b) a number of unknown but relevant problems.

The second criterion was the most important to us, because we wanted to test our claim that TSTI discovers problems that have not yet been reported in the literature from ‘traditional’ cognitive research or encountered by experienced users of the QFV questionnaire. The first criterion was relevant for the issue whether the TSTI would incorporate the achievements of other techniques (and, thus, could replace them) or would not be able to do that (and, thus, could only complement them).

We used a form of theoretical sampling, aimed at the discovery of as many ‘problems’ with the questions as possible and, therefore, including as many different kinds of respondents as possible, until ‘saturation’ was achieved. This saturation was achieved after sixteen interviews (see Jansen & Hak, 2002, for details).

In this pilot study, the third step of the TSTI entailed an intensive semi-structured interview on drinking habits resulting in a computation of the respondent’s volume of alcohol intake independent of the measurement by the questionnaire (in step 1). In these step 3 interviews, respondents often remembered drinks and drinking occasions that they had forgotten to report in step 1. Furthermore, quantities as well as frequencies were specified in much more detail in these interviews; respondents consistently evaluated their step 3 report as more accurate than the primary report. Therefore we feel that in this case the self-report in step 3 can be used as a criterion, a ‘gold standard’, for the assessment of the quality of the primary report in step 1.

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 basis of the literature review (see Jansen & Hak, 2002, 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’ 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 bacchanal 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 about here]

In general, mismatches as illustrated in Box 1 were *discovered* (identified) in steps 1 and 2 of the TSTI but could only be *interpreted* by the exploration of these respondents' lifestyles and consumption patterns expressed by them in step 3 of the TSTI. We conclude that, regarding this specific set of questions, it is the combination of observation and exploration in the TSTI (in that order) that makes it 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 of 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. These results cannot automatically be generalized to the testing of other types of self-completion questionnaires, such as attitude measurements. Therefore, we conducted a second pilot study in which an attitude scale was tested.

Second pilot study: validation of an attitude scale

In the second pilot study, Dutch and Norwegian versions of an attitude scale, the 20-item Illegal Aliens Scale, were validated (Hak et al., 2001; Van der Veer et al., 2002). Responding to an attitude scale is an activity that differs considerably from answering questions about, e.g., 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*. As in the first pilot study, predictions regarding problems with the questions were based on this other type of evaluation, and again our criterion for success was 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 & Larsen, 1997) is a Likert-type 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

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. 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 Van der Veer 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 respondents and, second, to explore possible reasons for found differences in interpretation. Two convenience samples were recruited, one consisting of six undergraduate students in the social sciences at the Vrije 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 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., 2001). Additionally, the TSTI identified other, unpredicted problems related to the interpretation of items. Our main finding was that several respondents, regarding a number of items, felt as if ‘forced’ to make a choice between two possible ‘readings’. Take the following example in Box 2.

[Box 2 about here]

In step 1 (concurrent think aloud), the respondent recognizes the item as one in which a difference is constructed between ‘legal’ and ‘illegal’ immigrants. He disagrees with this distinction. His selection of the response category *uncertain* can be seen as expressing avoidance to take sides for or against ‘illegal’ immigrants. In step 2 (focused interview), the respondent gives another reason for choosing *uncertain*, namely that he is genuinely uncertain whether illegal aliens actually are a valuable human resource in the economy. In this reasoning, the respondent interprets the statement ‘literally’, not as the expression of a hostile or friendly attitude towards immigrants. In the third step of the TSTI (semi-structured interview; fragment not shown in Box 2), this respondent confirmed that he was aware of the fact that the resulting IA score was less ‘friendly’ towards illegal immigrants than would have been the case if he had followed the expectations of the authors of the questionnaire. He had clearly recognized that the authors would expect him to demonstrate his friendly attitude to illegal immigrants wherever possible, i.e. by reading items as invitations to position themselves politically rather than as questions about economic or social facts. 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). 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 TSTI's, both in the Netherlands and in Norway. Our conclusion is that there *might* be a problem with the IA Scale in the sense that, due to this phenomenon, 'friendly' attitudes to immigrants might be underrepresented in IA Scale results. This possibility deserves further research. Our conclusion regarding the TSTI is that it appeared to be productive in having detected this problem that obviously might compromise measurements with the IA Scale.

In sum, as in the first pilot study, the TSTI both replicated the results from an expert review, and 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. The results as found in this study could not have been produced with traditional 'cognitive' interviews, which do not focus on *observation* of actual response behavior. If we compare the results of this study with the first pilot study, we notice an important similarity and an important difference. The similarity is that in both cases the TSTI appears to be productive in identifying problems that arise from the (biographical, cultural, 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 (semi-structured interview). Interviewers and respondents in the study of alcohol consumption questions used this third step for an exploration of the 'facts' of the respondents' drinking behavior. In the study of the IA scale it appeared that, in general, there was not much left to explore with respect to the respondents' attitudes after they had completed the second step. The only issue that could be explored in the third step, which is reported above, was the respondent's attitude to the questionnaire (rather than to illegal immigrants). 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 our first series of pilot studies with an application of the TSTI to questions about health related Quality of Life (QoL). QoL is neither a behavior about which can be reported 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 reported QoL. We assumed that think aloud protocols would demonstrate us 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 pre-testing, namely the aim to ascertain whether the question measures the intended ‘concept’.

In this study (Westerman et al., 2002), 30 lung cancer patients are sampled from different Dutch hospitals for a two-year longitudinal study in which these patients will complete QoL questionnaires at least five times. This enables us to assess response shift over the duration of an entire illness trajectory (or at least a considerable part of it). Each time a patient in this study completes a QoL questionnaire, the TSTI format is applied. At the moment of the writing of this article (September 2002), ten patients have been included, and some of them have completed the QoL scales several times. We consider the data collected so far sufficient for the purposes of this article.

One of our concerns was whether old, rather sick people would be able to adhere to the think aloud technique. This proved difficult indeed. We will report elsewhere about restrictions that apply to the think aloud technique (and, therefore, to the TSTI) in terms of abilities that respondents must have. Important for the present discussion is that many patients were able to endure the TSTI and that the resulting protocols (transcripts) were useful for our purposes. Interviews were conducted at the patients’ homes, which is in line 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 about here]

In Box 3, the think aloud protocols make the respondent’s reasoning observable, for each measurement point (T1, T2, T3) separately. 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, which can be supported by data collected in the steps 2 and 3 of the ‘interview’, 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).

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 an instrument for assessing the quality of a self-administered questionnaire by *observing actual* instances of interaction between the instrument and a respondent (the response process). *Concurrent thinking aloud* is used as a technique for making the thought process observable. The TSTI has been 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 and features of a respondent’s actual behavior and biography. In the second pilot study, Dutch and Norwegian versions of an attitude scale, the 20-item Illegal Aliens Scale, were validated. The TSTI appeared to be uniquely productive in identifying problems resulting from different ‘response strategies’. In the third pilot study, the TSTI appeared to be an effective instrument for producing data that document processes of ‘response shift’ in the measurement of health related Quality of Life (QoL). While producing this kind of ‘new’ data on the performance of specific instruments (data that are not produced with other methods for the pre-testing of questionnaires), the TSTI produces at the same time also data that are produced with other extant 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 & Blair, 1994; Willis et al., 1999; and Rothgeb et al., 2001).

References

- Dijkstra, W. (1998) A new method for studying verbal interactions in survey interviews. *Journal of Official Statistics*, 15: 67-85.
- Ericsson, K.A., & H.A. Simon (1980) Verbal reports as data. *Psychological Review*, 87: 215-250.
- Fowler, F.J., & C.F. Cannell (1996) Using behavioral coding to identify problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research*, (pp. 15-36). San Francisco: Jossey-Bass.
- Hak, T, K. van der Veer, & R. Ommundsen (2001) An application of the Three-Step Test-Interview (TSTI): *A validation study of the Dutch and Norwegian versions of the Illegal Aliens Scale*. Paper presented at the Congress of the European Sociological Association, August 2001, Helsinki.
- Jansen, H., & T. Hak (2002) *Assessing the quality of quantity-frequency-variability questioning in a mail alcohol survey. Results from a field study with three-step test-interviews (TSTI) compared to a desk expert analysis*. Paper presented at the International Conference of Improving Surveys (ICIS 2002), Copenhagen.
- Maynard, D.W., et al. (eds.) (2002) *Standardization and tacit knowledge: interaction and practice in the survey interview*. New York: Wiley.
- Nielsen, J. (1993), *Usability engineering*, San Diego: Morgan Kaufmann.
- Nielsen, J., & R. Mack (1994), *Usability inspection methods*. New York: Wiley.

- Ommundsen, R., & K.S. Larsen (1997) Attitudes toward illegal aliens: the reliability and validity of a Likert-type scale. *The Journal of Social Psychology*, 137: 665-667.
- Presser, J., & J. Blair (1994). Survey Pretesting: Do Different Methods Produce Different results? In P.V. Marsden (Ed.), *Sociological Methodology*, Vol 24, (pp. 73-104). Washington, DC: American Sociological Association.
- Rothgeb, J., G. Willis & B. Forsyth (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, Montreal, May 2001.
- 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*, (pp.73-100). Washington, DC: National Academy Press.
- Van der Veer, K., R. Ommundsen, T. Hak, & K.S. Larsen (2002) Meaning shift of items in different language versions. A cross-national validation study of the Illegal Aliens Scale, In *Quality & Quantity* (in press).
- Van Someren, M.W., Y. Barnard, & J.A.C. Sandberg (1994) *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.
- Westerman, M., T. Hak, A.M. The, G. van der Wal & H. Groen (2002) *Response shift in Quality of Life in palliative treatment of small cell lung cancer patients*. Internal memorandum, Department of Social Medicine, VU Medical Center, Amsterdam.
- Willis, G. (1999) *Cognitive interviewing. A "how to" guide. Manual for the short course 'Reducing survey error through research on the cognitive and decision process in surveys'*. Presented at the 1999 meeting of the ASA. Rachel A. Caspar, Judith T. Lessler and Gordon B. Willis, Triangle Institute.
- Willis, G.B, S. Schechter, & K. Whitaker (1999) A comparison of cognitive interviewing, expert review, and behavior coding: what do they tell us? Paper presented at the Annual Meeting of the American Statistical Association, Baltimore, August 1999.

Appendix

Manual for the Three-Step Test-Interview (TSTI)

Respondent selection

The main aim of the TSTI is to produce data on how respondents (would) complete a questionnaire ‘in real life’. Respondents, therefore, should be members of the intended respondent population, i.e., patients of hospital X if the questionnaire will be used in a survey of patients of hospital X, elderly women in state Y if the questionnaire will be used in a survey of elderly women in state Y, etc. Because every study has its own objectives and design, which determine the sampling principles and procedures that should be used, sampling as such will not be discussed in this manual. As with all pre-test research, it is important to explain to potential respondents that the object of study is the quality of the questionnaire, not their skill in dealing with it. (This will also be stressed below, in the discussion of the introduction to the interview.)

Setting

Because the main aim of the TSTI is to produce data on how respondents (would) complete a questionnaire ‘in real life’, respondents should complete the questionnaire in a setting that is as similar as possible to ‘real life’ unobserved completion. There are two main options:

- (a) *‘Field interview’*: The TSTI is conducted in the setting in which, according to the respondent, ‘normal’ (unobserved) completion by the respondent would take place. This will be in the respondent’s home (at the dining table, at a desk, or on the couch) if a questionnaire is tested that will be mailed to the respondent’s home address. Or, this will be on a hospital ward if a questionnaire is tested that patients must complete during their stay in the hospital, etc. The main advantage of the ‘field interview’ is its ‘ecological validity’. Its main disadvantages are that the interview is disturbed by contingencies (telephone calls, etc.), that video recording of the interview is difficult or impossible, and that the technical quality of recorded data (e.g., the quality of the audio recording) can be compromised.
- (b) *‘Laboratory interview’*. The TSTI is conducted in a ‘cognitive lab’ or a ‘pre-test lab’. The obvious advantages of the lab setting, both in technical and in organizational terms, have to be weighed against the loss of ‘ecological validity’.

Informed consent

Before the interview begins, informed consent should be obtained for its recording. The usual safeguards should be given regarding issues such as confidentiality of everything that is said or seen in the interview, the way quotes from transcripts will be used in publications of the research, and the security of the collected data. These safeguards apply to the responses to answers in the questionnaire that is tested as well as to the data that are specifically produced in the TSTI procedure.

Introduction to the main aim of the interview: observation

The aims and procedures of the interview must be explained to the respondent. It must be emphasized that the interview's aim is to test the questionnaire, not the respondent. It must be made very clear to the respondent that the researcher (interviewer) wants to *see for her/himself* how good or problematic the questionnaire is by *observing* how actual respondents proceed. In this context it might be mentioned that there will be two distinct phases in the interview, first, the completion of the questionnaire by the respondent and, second, an evaluation of what happened. Respondents can be advised that there will be ample time for them to express their opinions, ideas, and proposals in the second stage of the interview. But "*for the sake of the test, could you please abstain in the first part of the interview from any commenting on what you do?*"

Introduction to thinking aloud

It is useful that the respondent knows why the think aloud technique is used. Most respondents understand that their response to the questionnaire mainly occurs 'in their head' and that, therefore, an observer needs additional information. It can, however, be more difficult to instruct the respondents in performing the think aloud technique. It is essential for the TSTI to be successful (as a procedure) as well as productive (for analysis) that the instruction to the think aloud task is done well. Minor deviations in instruction and execution might invalidate the entire interview. First, the conditions should be right. Has a good relationship between interviewer and respondent been established? Should anything been done to make the respondent feel more comfortable? The instruction itself is rather simple: respondents are asked to *say aloud what they think as they think it*. Note that the instruction is not to 'think aloud' because this is usually interpreted by respondents as a request to *think* in a specific way (which is confusing), whereas the request is (only) to *say aloud* what one is thinking anyway. Note also that this implies that no explanations of these thoughts are required, but just the verbalization of the thoughts themselves. Interviewers must explain this to respondents and they must also instruct them that they should not invent thoughts just to avoid silences at all cost. Respondents must be instructed to say aloud only those thoughts that come 'naturally' as part of the task of completing the questionnaire. Respondents differ considerably in the degree to which they are able to perform such a think aloud (or say aloud) 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 an inability to do it. Therefore, it is usually useful to do some exercises in thinking aloud before starting with testing the questionnaire.

Think aloud exercises

The following are some exercises from the literature (see, e.g., Willis, 1999, and Van Someren et al., 1994).

1.
Try to visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about. (from Willis, 1999: 4)
2.
When was the last time you had dinner in a restaurant? Please say aloud everything you think in the process of finding that date.
3.
 - (a) Talk aloud while answering the following question: How often have you been to the grocery store during the last week?
 - (b) Please describe to me your first visit of last week to the grocery store. Tell me in a strictly chronological order what you did during that visit after having entered the store until you left it with the things you had bought.

Feedback by the interviewer

In order to be able to learn, in these exercises, what adequate ‘thinking aloud’ (or rather ‘saying aloud’) is, respondents must receive adequate feedback. This should reinforce good performance and discourage misinterpretations of what is expected from them. A distinction can be made between different forms of feedback according to their aims. (These do not only apply to feedback to the exercises but also to interviewer feedback during the consecutive ‘real’ think aloud task.)

- (a) Many respondents find it difficult to verbalize all thoughts they have, because they experience this verbalizing as interfering with the thinking itself. Therefore, one important aim of interviewer feedback is to support the respondent in saying aloud as many thoughts as is possible. This can be done by means of fairly simple remarks such as “Please continue talking” and “Please say aloud what you are thinking”. Note that these are requests (only) to say aloud what one is already thinking anyway. These are not requests to do and, thus, think something else.
- (b) Another aim of interviewer feedback is positively reinforce good performance. Interviewers might, therefore, use expressions such as “Good, you are doing this (very) well. Please continue in this way” when respondents have, either partially or completely, successfully verbalized a thought process.
- (c) Logically, a third aim of feedback is to alert respondents when they seem to deviate from what they are expected to do and, if possible, to correct them. The interviewer must ascertain continuously that the respondent only says aloud those thoughts that (apparently) belong to the response process. Respondents should not explain or justify these thoughts to the interviewer or comment on them. If necessary, the respondent should be reminded of the fact that ‘thinking (or saying) aloud’ is a technique for making observable ‘behavior’ (thinking) that is going on, not an ‘interview’ in which accounts, opinions or other reports about self are sought. The interviewer might, thus, make comments such as

“Please only say aloud what you think in order to respond to the question. Please do not comment on these thoughts just because I am here listening. Just ignore me, do as if I am not here. In the next phase of this interview you will have ample time to comment on what you have done and why you have done it the way you did”.

Interviewers might demonstrate explicitly to respondents that they are mere observers, at least during the think aloud task, by positioning themselves off the respondent’s view (e.g., by sitting behind the back of the respondent) or by adjusting their posture accordingly. (Incidentally, this is also a recommendation in the psychology literature on think aloud.)

Step 1. Concurrent thinking aloud

When both respondent and interviewer agree that enough exercises have been done and that the interviewer’s feedback is understood, the TSTI proper can begin with its first step, which is the ‘think aloud’ task concurrent with the completion of the questionnaire that is tested. The object of the TSTI is the way a respondent completes the task of responding to a questionnaire as a whole, not just to isolated parts of it. Therefore, step 1 (think aloud) should cover the entire response process, i.e. from its very beginning (e.g., opening the envelope in which the questionnaire will be mailed to respondents) until the end (e.g., putting the questionnaire in the pre-paid stamped mail-back envelope, if provided). Consequently, step 2 cannot begin until this entire process is completed. This requirement follows directly from the explicit aim of the TSTI to keep the procedure as close as possible to what ‘normally’, i.e. without the test situation, would happen. Note that this constitutes a difference with the common practice in pre-testing in which usually problems with one question are explored as widely and deeply as possible (e.g., by means of both concurrent and retrospective probes) before moving on to a following question. In contrast, no probing is allowed in step 1 of the TSTI. Talk by the interviewer, who really is an observer, should be confined to giving feedback, as discussed above.

Apart from supporting the respondent in performing the think aloud task, the main other task of the interviewer is to make notes (in ‘real time’) of remarks and behaviors of the respondent. Specific notes might be made for different reasons and uses. Some might indicate specific problems with questions or, for that matter, other aspects of the response process that might be interesting for further exploration (in step 3). More importantly, notes must be made of instances in which the observer feels that a part of the thought process (and thus of the response process) has not been properly verbalized. Notes on such ‘missing’ data are important as ‘input’ to step 2, because the main function of step 2 in the TSTI is to provide for observational data that, for whatever reason, were missed in step 1. (Later analysis of the response process and of the quality of the questionnaire does not solely depend on these notes made in real time by the interviewer. The interview will be recorded as well. But these real time notes are important because steps 2 and 3 of the interview will, to a large extent, depend on these notes.)

Step 2. Focused interview

Step 1 (thinking aloud) can be concluded with a thank you for having completed this difficult and onerous task of completing the questionnaire while ‘thinking aloud’. If the respondent shows signs of fatigue, it is important to show understanding for this. However – though tempting – this is not a good time for a break. Step 2 is aimed at ‘filling in’ bits about the thought process that were not sufficiently captured in step 1, although the process itself cannot be observed for a second time. Because step 2 can only produce respondents’ self-reports about what has happened, rather strict criteria should apply to how the interview in step 2 is conducted in order to get ‘factual’ data about how the process occurred. An important condition is that step 2 immediately follows step 1, and that no time is lost. The most important element of the role of the interviewer in this step is that he – similar to his role in step 1, but now retrospectively – must help the respondent to focus on the reporting of actual thoughts rather than of interpretations of them. This restricts the repertoire of possible questions in this step quite rigorously to only one format: “And what happened next?” Note that this ‘probe’ is much more restricted than ‘retrospective probes’ in cognitive interviews, despite its similar location in the interview process. Memory and reporting of these missing bits of data on the actual thought process of the respondent could be invited and supported by quotes from the notes that the interviewer has made. For instance, he could say “You said/did ... [quote from the notes from step 1] ... and then you did not verbalize your thoughts for a while. What were you thinking at that moment?” However, in some cases, in order to reconstruct a complicated step in a thought process (e.g., a decision about whether one agrees or disagrees with a statement, which involves a series of different and incomparable criteria) it might be necessary to explore the respondent’s ideas and views about it. The result of step 2 is a more complete reconstruction of the thought process that resulted in the choice of a response category in this specific case. The guidelines in this manual for how to proceed in the steps 1 and 2 of the TSTI are meant to ascertain that exactly this kind of descriptions of actual response processes are produced. Questionnaire designers can inspect these results in order to get insight in how their questions function in actual practice.

Step 3. Semi-structured interview

Because the completion of steps 1 and 2 might have tested the endurance of the participants, this might be the right time for a break. Steps 1 and 2 of the TSTI are uniquely designed to produce data about actual behavior (thought processes). Both the (concurrent) thinking aloud procedure in step 1 and the focused interview in step 2 are described fairly extensively in this manual because they are not frequently found in this form in current practices of pre-testing. The following step 3 of the TSTI, a semi-structured interview, is much less specific than the other two steps but due to its placement after the two other steps it is productive in a different way than it usually is. Depending on the kind of questionnaire that is tested and on the specific aims of the pre-testing, this part might entail different forms of qualitative (semi-structured) interviewing. Examples are:

- (a) Respondents might (be requested to) ‘explain’ their response behavior. Particularly when specific problems were encountered in responding to the questionnaire, they could comment on what they think the exact nature of the

problem was and why they behaved the way they did – as 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. The aims and form of this interview will be similar to those of ‘respondent debriefing’. It is important to acknowledge that this kind of comments constitute ‘opinions’, 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 pre-test).

- (b) Respondents might be asked to paraphrase questions and to comment on their definitions of terms or, in other words, some form of ‘cognitive interviewing’ might be done in this stage of the TSTI.
- (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 to the interviewer in their own words.

Conclusion

As in any kind of interview, respondents should be asked whether they have anything to add to what has been said and whether they have liked the experience. Respondents should be encouraged to bring forward any recommendation (about the test interview or about the questionnaire that has been tested) that they find relevant. They should be thanked for their cooperation. If respondents request this, a promise should be made (and kept) that they receive a summary of the study’s findings.

Box 1

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

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

Step 2: Focused interview

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: And you also marked "don't know"

R9: Well, the one time three and the other time nothing

Step 3: Semi -structured interview

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.

Box 2

Item:

Illegal aliens provide the Netherlands with a valuable human resource

Step 1: Observation

R: *illegal aliens provide the Netherlands with a valuable human resource*erm.....I immediately think, erm, 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, erm, a straightforward opinion ... so, it is uncertain....because one can be valuable also if one is legal.

Step 2. Focused interview

R: Well ... here I have ... I think, in my opinion, I lack knowledge a bit erm ... I have *uncertain* ...erm well, *a valuable human resource*, well, what can I say about it?
I: well erm there might also be a kind of logical reasoning behind it ... that the item suggests a difference between illegal and legal people
R: oh right yes that's what I said
I: you said if this applies to everyone, erm, why should I confirm it here for illegal people only? that kind of reasoning
R: yes yes could be yes
I: but now ... you say, you say also the facts ... you don't know the facts
R: no, you're right, the way I thought was ... indeed I thought that the distinction between illegal or normal, just a normal dutch person ... that is not clear to me ...and if there is a distinction .. well, I don't know either whether they are valuable or damaging ... those are two things, a bit two things ... of which I don't know

Item:

Illegal aliens cost the Netherlands millions of guilders each year

Step 3. Semi-structured interview

I: take the item *illegal aliens cost the Netherlands millions of guilders each year* .. such an item ... you take it literally
R: erm
I: you erm you know for sure 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 interprets 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 erm illegal aliens ... in that case you have been almost deliberately, deliberately 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: and 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

Box 3

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

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

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 was nothing ... I do it without any problem.

T3

Step 1: Observation

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 sometimes this answer refers to a walk to the shopping center (450 meter) that on another occasion might be considered a 'short' walk (because, at that moment, it can be walked without much trouble).