Journal of Official Statistics, Vol. 21, No. 1, 2005, pp. 103–120

The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-administered Questionnaire on Alcohol Consumption

Harrie Jansen[1] *and Tony Hak*[2]

The three-step test interview (TSTI) is a recently developed observation-based procedure for the identification of response problems in self-administered survey questionnaires. The TSTI was applied in field test interviews to a quantity-frequency-variability questionnaire on alcohol consumption. For an assessment of its productivity the results are compared to a previously performed expert review. Most response problems that were identified in the expert review were confirmed in the field test interviews. Additionally, the TSTI identified many unexpected problems, mostly stemming from unanticipated “deviant” drinking patterns and from local normative connotations attached to drinking alcohol. From these findings we conclude that the TSTI is a powerful test tool with a high ecological validity.

Key words: Data quality; pretesting; cognitive methods; ecological validity; focused interview; think aloud; observation.

1. Introduction

Survey questionnaires often make high demands on the motivation and skills of respondents, either by the contents of the information asked for or by the way of questioning (such as question wording and the formatting of response categories). This may cause problems of data quality in terms of validity, accuracy, reliability and completeness (item nonresponse).

In the field of alcohol consumption surveys, these problems have been discussed mostly in terms of the accuracy of self-reports. Debates and research focus on the issue of underreporting in surveys as compared to sales data (Midanik 1982; Garretsen 1983; Lemmens 1991; Lemmens et al. 1992; Midanik et al. 1999). Of course underreporting is a general phenomenon in self-reports on routine behaviour (Mingay et al. 1994) but it seems to be higher with alcohol consumption than, for example, with soft drinks (Reinhard and Horwitz 1995). Consequently it is quite common to consider methods that generate higher consumption rates to be better than methods generating lower rates, at least for general population surveys (cf. Lemmens et al. 1992; Rehm and Spuhler 1993; Romelsjö et al. 1995).

For aims like the identification of risk factors or the analysis of relationships between drinking and biographical characteristics, the quality of self-reports has to be assessed at

[1] IVO, Heemraadssingel 194, 3021 DM Rotterdam, The Netherlands. Email: jansen@ivo.nl
[2] RSM Erasmus Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. Email: thak@rsm.nl

the levels of subgroups and individuals. Furthermore, for specific subgroups like "heroic" youngsters, as well as for specific institutional settings like selection interviews for clinical treatment, overreporting might be more of a problem than underreporting because of the attributed benefits (help, prestige) of high reporting in these cases (Del Boca and Noll 2000). It is now widely recognised that we cannot rely on comparison of aggregate frequency counts as a test for the quality of questionnaires on alcohol consumption, and this applies to other questionnaires as well. For most purposes data quality in surveys has to be assessed and enhanced at the individual level. To that end cognitive theories have been applied increasingly to the studying of response processes in survey questioning (Sirken et al. 1999). In this article we will first review literature on respondent problems in answering questions and on methods to reveal these problems. Subsequently we will present the Three-Step Test-Interview (TSTI) and evaluate its productivity empirically in comparison to an expert review of the same questionnaire on alcohol consumption.

1.1. Classification and identification of problems in answering survey questions

Data quality problems can be defined as respondent problems, i.e., problems of respondents with regard to acting adequately in relation to the questionnaire. In order to enable respondents to report their facts validly and accurately, two types of conditions must be met. First, motivational conditions: respondents must be enabled to identify themselves positively with the research aims and the task should be perceived as worthwhile. Second, cognitive conditions: the respondents must understand the question, and the response categories given in the questionnaire should provide them with an adequate model to formulate or denote their answers. There may be a trade-off between motivational and cognitive conditions (Lahaut et al. 2003).

Almost all test studies focus on the cognitive part of response behaviour in laboratory settings by applying the four-stage model of the response process (Tourangeau 1984; Sudman et al. 1996). This model includes: comprehension/interpretation of the question, retrieving the requested information, judgment formation (evaluating the adequacy of the retrieved information) and reporting the answer (in writing or by marking the right box). In an elaboration of this model, Midanik and Hines (1991) specified different strategies for retrieving information. They hypothesised that response problems may stem from a mismatch of question formats to habitual retrieving strategies of respondents. An important detail in their study is the inclusion of normative elements. They discovered that many respondents after initial retrieval of information adjust their reports to some normative idea of what seems plausible. In later research they included "context" as well, which means a further elaboration towards coverage of noncognitive aspects of the response process (Midanik et al. 1999).

Conrad and Blair (1998) studied response problems in interviews on a more pragmatic basis, i.e., without any assumptions about strategies applied or strategies required. Their aim was to design an easily applicable coding scheme for the analysis of "think-aloud" protocols and retrospective interviews in order to improve reliability of the coding. They preferred a three-stage scheme of the cognitive process, because of the practical impossibility for the analyst to distinguish between the retrieval and evaluation of information. Furthermore they provided a list of problem types that may emerge in each of these three stages: lexical, temporal, logical, computational and omission/inclusion problems. Their matrix of response problems (problem types by stages) has the advantage of giving a detailed overview of the most frequently occurring cognitive problems of respondents. That makes their inventory useful as a checklist in reviewing questionnaires or in the construction of cognitive interviews.

Notwithstanding theoretical differences in the classification of problems, the methods used by Midanik (1991, 1999), Conrad and Blair (1998) and most other researchers to identify response problems in survey research are quite similar (Sudman et al. 1996; Campanelli 1997). The most common methods are:

1. Expert analysis, or expert review. It is (or should be) common practice that researchers who develop survey questionnaires consult fellow researchers to evaluate and discuss draft questionnaires. In order to make this more accountable and sophisticated, coding schemes have been developed on the basis of cognitive psychological principles. (Sudman et al. 1996: pp. 28–29).
2. "Think-aloud" procedures, also referred to as "protocol analysis": the subject is asked to think aloud while preparing to answer (Ericsson and Simon 1993).
3. "Cognitive interviews," in which respondents are probed retrospectively on their interpretations of questions and specific words used in them (Loftus et al. 1985; Köhnken et al. 1995).

In practice "cognitive interviewing" often covers both "think-aloud" observation and retrospective interviewing on interpretations. It seems appropriate to distinguish between these methods analytically.

1.2. The three-step test-interview (TSTI)

Usually, by conducting pretesting research in laboratory situations with "professional" respondents, motivational and contextual factors are neglected. Furthermore questions tend to be studied separately, thereby missing routing errors and interaction effects between questions. In order to overcome some of these limitations of reliability and ecological validity, we designed a three-step procedure for testing the quality of a self-administered questionnaire on alcohol consumption.

To start with, we choose a "natural" setting for testing: the home, where subjects normally complete survey questionnaires. Our procedure starts with tape-recorded concurrent thinking aloud as a first step. The researcher asks the subject to complete the questions while reading and thinking aloud; while the subject is doing this, the researcher writes notes on the respondent's verbal (deliberations about the questions) and nonverbal (marking answers, skipping questions, hesitating etc.) behaviours.

In the second step the researcher returns to the start of the questionnaire and conducts a focused interview about various observations during the first step, in order to fill up missing observational data or to check the validity of the observations. In this focused interview respondents are not asked for their interpretation of terms, e.g., "alcoholic beverage," in general, but we observe and ask how they have actually interpreted such concepts in this case while responding. The focus remains on clarification of observed behaviour. Typical question formats are: "how" (did you arrive at this answer?) and "what" (were you thinking: you first marked response category (1) and then you marked (2) – what happened?)

Then, in the third step, we proceed with a debriefing interview with questions about the interpretation by the respondent of the observed response problems, exploring reasons why they found it difficult to understand a question or to find an accurate response category, for instance in terms of a mismatch between the underlying model of the response categories and their actual drinking habits and biographical context.

In our study, we added after the TSTI proper a short substantive interview about drinking habits throughout the year: daily habits, weekly habits and nonweekly events (birthday parties, holidays, etc.). This interview was intended as a validation device to appraise the effects of response problems on data quality.

2. An Empirical Evaluation of the Productivity of TSTI in Comparison to an Expert Review

In order to assess the power of TSTI as a method to detect respondent problems, we will compare the results of a field study by TSTI with the results of an expert review that we did before on the same questionnaire. The aim of this study is, thus, to answer two questions:

i) Does TSTI "discover" the same problems that were identified in the expert review?
ii) Does TSTI identify specific response problems that were not identified in the expert review?

The subject of our study is a "Quantity-Frequency-Variability" (QFV) questionnaire on alcohol consumption which was used in several surveys in the Netherlands between 1983 and 1999 (Garretsen and Knibbe 1983; Bongers and Van Oers 1998; Verdurmen et al. 2000).

It consists of six questions:

1. A preparatory question asking which type of alcoholic beverage the subject drinks most often, followed by a list of categories of alcoholic beverages from which one must be chosen. The function of this question is, first to operationalise the concept of alcoholic beverage (e.g., it does not include light beer), and then to single out non-drinkers.
2. A question on the frequency of drinking six or more glasses of alcoholic beverages in one day. This question serves to identify "binge drinking."
3. A question asking for the number of days per month the subject usually drinks any alcoholic beverage.
4. A follow-up to Question 3, asking for the number of glasses drunk on average on such "drinking days."
5. Two questions on the variation in alcohol consumption over the last ten years, i.e., periods of drinking much more or much less than nowadays.

2.1. *The expert review*

We executed an elaborate expert review of this questionnaire in two stages. First, we read the questions carefully, compared them with similar and (almost) identical questions in other alcohol survey questionnaires and searched in the literature for any information available about problems identified with these question types and, if possible, with these very questions. The result of this review was a list of known and potential problems that (might) occur when respondents answer these questions. In the second stage of this review we conducted intensive interviews with one of the original authors and two current users of the questionnaire, querying them about the exact concepts that they intended to measure with these questions, about reasons for specific wordings of these questions and response categories, and about any problems noticed by them. The result of these interviews was not only an extended list of known and potential problems that (might) occur when respondents answer these questions, but also a yardstick for validity, namely precise descriptions of the concepts that these questions were intended to measure. The results of this expert review (Hak and Jansen 1997) will be used here as a benchmark for the evaluation of the productivity of the TSTI.

2.2. The field test with TSTI: sampling and saturation

This study is not aimed at producing frequency distributions of response problems in the population of respondents, but at producing an inventory of problems that subjects encounter when completing the questionnaire. Therefore the most important criterion for the quality of our sample is neither its size nor its statistical representativeness, but rather its theoretical representativeness, i.e., the degree to which it is able to represent all types of response problems that are of relevance in the study population. In order to ascertain the theoretical representativeness of our sample, we adapted the procedure of "theoretical saturation" as developed in the grounded theory approach (Glaser and Strauss 1967). This implies that after each interview (or observation) it is ascertained whether this interview produced relevant new information. If it did not, a decision needs to be taken about whether and where to look for deviant cases that might yield something not yet seen. When the analysis of a number of new cases does not produce new information that is relevant to the study, one may claim that "saturation" has been achieved and data gathering may be stopped. We followed a variant of this procedure by starting with seven Rotterdam Dutch-speaking subjects representing a wide variety of lifestyles (1–7 in the list below). These seven interviews produced enough information for a first report on the quality of the questions, but because of the absence in this sample of subjects with lower levels of education, which might be considered a deviant category from the ones represented before, it was hypothesized that some categories of relevant response problems might have been missed. Therefore we interviewed three subjects with lower levels of education (8–10 in the list). These interviews generated two new problems. Next, we interviewed six subjects from another database (11–16). Each of these interviews was analysed before conducting a next one, in order to monitor "saturation." Interview 14 produced one entirely new ("normative") problem. Interviews 15 and 16 produced two new variants of the (known) "unit of capacity" problem, the problem deriving from not drinking from standard glasses (i.e., containing known amounts of alcohol), for example drinking from a bottle. If all "unit of capacity" problems are defined to be one and the same problem, then Interviews 15 and 16 did not produce new problems. After Interview 10 no new problem locations had been detected, and we had detected such a small number of new "real" problems that we decided to stop after 16 interviews.

3. Results of the Field Test: Problem Locations and Problem Types

In the analysis of protocols we could rather easily locate the places in the questionnaire where respondent problems originated. It also appeared rather easy to understand the reported problems on the basis of the interviews. For our aim of gaining an insight into respondents' problems, we felt little need for a more sophisticated way of coding the problems in terms of theoretically defined cognitive tasks such as in Conrad and Blair (1998). An overview of all the problems that we observed during the 16 TSTI's and in the preceding expert analysis is reported in the Appendix.

Here we will detail the case of Question 2 only.

3.1. Question 2 (Q2): frequency of drinking six or more glasses in one day

Question 2 was the most complex of the six questions tested. It asks subjects to say how often they have consumed six glasses or more of alcohol beverages in one day during the last six months. It is preceded by an introduction which is meant to counteract possible shame by "normalising" that amount of consumption. This was not identified by expert analysis as a problem. In the test-interviews we discovered response problems in seven places in this question:

Research has shown that a considerable part of the population drink six glasses or more of alcoholic beverages in one day, more or less regularly. ①

In the past six months ②, | did you ever drink six or more ③ | glasses ④

of alcoholic beverages ⑤ | in one day ⑥ ?

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

To illuminate the logical structure of the TSTI we present here an example of the results of the consecutive three steps of which the TSTI consists. The example has to do with interview data concerning one question, which has been taken from the transcript of the test interview with Respondent R9.

In Step 1 (observation) it was observed that he first ticked Category 3 (yes, 3 or 4 times a week) and then, after some hesitation, also Category 9 (don't know).

In Step 2 (focused interview) this conversation occurred:

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.

In Step 3 (debriefing interview) it appeared that this respondent is a shift worker at a beer brewery. He only drinks alcohol in periods when he does not work (about one in four weeks). In these free weeks he drinks quite often, six or more glasses of beer a day. But that may vary a lot too: from three times in one week to five times in another week. In weeks when he is at work, he does "not drink," which means not more than one or two glasses after work in his case.

3.2. Overview of problems discovered in Q2

① Intro has adverse effect on R7:

Think-aloud: R7: *Those people are alcoholics.*
Focused interview:
R7: This leads you astray. that social research has shown that a considerable part of the population drinks six glasses or more – to me it is a kind of a freak [. . .. . .], it disturbs me.

② Past six months
In the expert analysis it was doubted whether respondents are able to measure "past six months" accurately. This proves to be a problem indeed: R4 and R7 measure the six-month period exactly, others do not.

Debriefing interview:
R7: The past six months was important; should you ask this in June, then I would think of wintertime, but now I think of last summer, and drinking in summer and winter are different in my case.

In the expert review it was noticed that computing weekly or monthly means seems very complicated over six months, especially with high frequencies. R4, R8, R9, and R13 confirm this problem by reporting difficulty in calculating the mean frequency because of an irregular or divergent drinking rhythm.

Debriefing interview:
I: The question is about the period of six months, can you survey that period?
R8: Yes, I think so, but it is more about the beginning of this six months, because now I drink only a little.
I: You are now in a period of little drinking, and before the summer you were drinking more?
R8: Yes, sure.

③ Six or more
In the expert analysis it was hypothesised that some people may interpret this question as asking for occasions of drinking more than usual. This is confirmed by R7, who interprets this question as a question on occasions of drinking (too) much alcohol, which in her case is drinking three (or more) glasses.

I: Did you think "how often did I drink more than six glasses"?
R7: Yes . . .I get very sick of it. . .So I know perfectly when I am over three glasses.
I: Okay, but the question is about six or more.
R7: Yes then I am very drunk.

④ Glasses
It was foreseen in the expert review, that various respondents probably count bottles of beer.
This appeared to be true in the case of R8. Furthermore R15 counts big glasses (35 cc) with beer and R16 counts soft-drink glasses with Riesling.

⑤ Alcoholic beverage
The expert review expected a selection effect.
This appeared to be true in the case of R8: while thinking aloud, he asks whether to report beer only or not, because in the first question he marked "beer" as the beverage he usually drinks.

⑥ In one day
The expert review expected that respondents do not count days but drinking occasions with six or more glasses.
In many cases in the interviews it was not ascertained whether all the drinks in a day had been counted; therefore evidence is not clear on this point.

⑦ Response categories
In the expert analysis no remarks were made about this part, but it appeared to cause many problems in the field:
- R7 wanted to mark "2 or 3" but this option is not available.
- As we showed above, R9 did not understand that he was supposed to calculate a mean (which would have been very complicated in his situation) and ticked two categories, indicating his range of periods of drinking.
- R13 marked "3 to 5 times these six months" but he meant 3 to 5 times a week.
- R16 wrongly marked the first "Yes" (= every day).

4. The Productivity of TSTI Compared to the Expert Review

In this section we will compare the productivity of the TSTI field test with the expert review both quantitatively, i.e., in terms of the number of problems identified and located in the questionnaire, and qualitatively, i.e., in terms of the kinds of problems identified by the two methods.

4.1. Quantitative comparison

In the expert review six phrases in Q2 were marked as possibly problematic: ②③④⑤⑥ and ⑦. Except for ⑥ all these phrases appeared to cause problems in the TSTI study. This was also the case with phrase ① that had not been nominated. So we may conclude that there is a high degree of correspondence in the identification of problem locations. This conclusion holds also for the other questions, as may be seen from Appendix 2, which shows all the results in brief: from 17 problem locations that were identified in the expert review, 14 were affirmed in the field TSTI. In addition the TSTI study identified 10 new problem locations. This makes this TSTI study significantly more productive than the expert review in terms of the identification of problem locations. Furthermore TSTI discovered a larger variety of problems at many problem locations.

4.2. Qualitative comparison

There were differences in content of problems also between the expert review and TSTI. The TSTI revealed a number of specific problems that could not easily be derived from the analysis of question wording (grammar, internal logic). Many problems appeared to originate not so much from the complexity of formulations or of tasks defined in itself (which could be foreseen quite well by an expert review), but from mismatches of standard questions with "non-standard" habits (e.g., R9 in Q2). Furthermore, it appeared that some parts of the population display specific normative connotations (e.g., regarding the introduction of Q2) that probably would not show up in convenient student laboratory samples. Both types of problem (mismatch and local normative connotations) are bound to socio-biographical peculiarities.

In summary: the expert review detected mainly cognitive problems stemming from the complexity of logical operations respondents have to perform, and those stemming from inconsistencies within and between questions. It did not identify specific groups of respondents that could be expected to experience trouble in responding, but these are identified in TSTI:

- people doing shift-work whose drinking patterns follow the shift-rhythm,
- people who get tipsy after three glasses of wine,
- people who have changed their drinking habits in the past six months,
- people who have just returned from a bacchanal holiday.

4.3. The relevance for survey quality

From the fact that so many unexpected drinking habits have been detected in this small sample of 16 respondents, it may be concluded that these are not rare anomalies. This might also be seen as a result of our decision to conduct this study as a field test: even a small "theoretical" sample provides some indication about the extent of problems, whereas desk study results inevitably indicate *potential* problems. Furthermore complementing thinking aloud and focused interview with a third element, the debriefing interview, yields relevant insight into the contextual origins of problems in answering these questions. The first two steps of TSTI (observation and think aloud, followed by the focused interview on what was observed) are productive in terms of the identification of problems, whilst the third step (debriefing interview) is productive in terms of diagnosing the causes of these problems.

In the substantive validation interview that we added to the TSTI proper, answers about past drinking behaviour sometimes differed from answers given by the respondent when completing the questionnaire in the first (think-aloud) step of the TSTI. During the validation interview, both the researcher and the respondent invariably considered data produced in this post-interview to be more accurate than data produced in the first step. The reason for this was that in this interview the subject often remembered drinks that were forgotten initially, or that quantities and frequencies were recalled much more precisely. When confronted with differences between the self-administered report in the think-aloud and the oral report in the validation interview, subjects always rated the latter as being the better. Probably the learning effect of the preceding questioning strongly contributed to the experienced validity of the interview data (cf. Means et al. 1994 on training cognitive strategies). We can never expect perfect validity of any self-report, but on the basis of our experience we feel safe in claiming that after think aloud, focused interview, debriefing interview and the detailed validation interview, we got a highly valid self-report that can be used as a criterion, a "gold standard", for the assessment of the quality of the primary report in the self-administered survey, as Cutler et al. (1988) did in their study on the Health Survey Questionnaire.

5. Discussion

The important practical question now remains: what to do with this knowledge? How can the questionnaire be improved in order to prevent these response problems and make the alcohol survey more valid and reliable? First of all we must say that our primary aim was to develop an instrument (TSTI) for successfully identifying and diagnosing respondent problems, not also for solving them right away. From our study we conclude that it is commendable to accompany every population survey with a TSTI field test, because cognitive laboratory tests do not discover context-bound response problems. Even very well pretested questionnaires will, when fielded, show problems in some groups of respondents. Identification of these problems and groups will provide a basis for estimating biases and for making decisions about – for example – conducting additional interviews with special groups.

Although this was not the topic of our study, we think that many problems in interview surveys could relatively easily be solved by asking cognitive and interpretive questions when respondents have problems choosing fixed answers. Therefore interviewers should be trained in tailoring the interview (cf. Houtkoop-Steenstra 2000).

For mail and web surveys finding solutions might be more complicated. However, the validation interview that we conducted after the TSTI has given some indication of directions in which solutions could be found. In accordance with cognitive theory, we found that people remember social events (meals, birthday parties, feasts, and holidays) better than "occasions of drinking," because most often drinking is a secondary activity. Most people can easily estimate the frequency of social events and then memorize the typical number of drinks at specific events. Another suggestion, which we are currently exploring, is to start with a day-by-day "typical week" report, followed up with quantity-frequency questions on daily, weekly and yearly events. Frequencies of drinking 6 + or whatever number of units per occasion can then be computed afterwards electronically.

Finally we would like to address the external validity of this comparison between an expert review and a field TSTI. It is important to note that this study concerned questions that require respondents to recall events (such as "drinking") and make calculations about elements of these occasions (such as calculating the "average number of glasses"). TSTI appeared to be an effective test strategy for these questions. It might be the case that TSTI is less productive in testing other types of questions and questionnaires such as those concerning attitudes or knowledge (see Hak et al. forthcoming). Also, expert reviews might become more productive in the future if experts learn from cognitive laboratory and field research. Nevertheless test interviews with the target population will remain indispensable for the assessment of survey data quality. The Three-Step Test-Interview seems to be a valuable addition to the methodological repertoire in data quality assessment.

APPENDIX 1

Social Characteristics of Respondents

Table 1. Social characteristics of respondents

Gender	Age	Household/ family type	Level of education	Profession/job
1. man	49	single	higher vocational	manager
2. woman	74	widow	elementary school	housewife
3. woman	44	married + children	university	teacher secondary school
4. man	31	single	university	unemployed
5. woman	27	couple	higher vocational	social worker
6. man	27	couple	university	researcher
7. woman	53	single	higher vocational	unemployed
8. man	29	single	lower general	administrative worker
9. man	47	unknown	elementary school	manual worker
10. woman	49	married + children	middle vocational	carer
11. woman	48	married couple	university	teacher
12. man	61	married + children	lower vocational	hostler
13. man	18	with parents	lower vocational	plumber
14. man	67	married couple	lower vocational	(ex-) furniture maker
15. man	34	couple	higher vocational	IT consultant
16. woman	56	married + children	middle vocational	teacher creative skills adults

APPENDIX 2

Results from the Field TSTI Study Compared to the Expert Review[*]

Question 1.	expert review	field TSTI
Which alcoholic beverage do you take ***mostly*** ① *if you drink?*		- ① R5 reads "*mostly*" as "most frequently," R2 and R6 read it as "largest quantity"
Please mark <u>one</u> category②		+② R8 asks at the next question whether this concerns beer only, because this is the drink he marked here
1 ( ) beer		- ② R8: one answer is impossible
2 ( ) wine, sherry, port or vermouth ③	+② restriction to one category here may induce restriction to this type of drink in answering later questions	- ② R3 overlooks this instruction; R9 does not understand it; both of them mark two categories
3 ( ) liquor, eggnog, black currant gin, lemon gin		- ③ R4: strange category, I would take beer and wine together
4 ( ) gin, brandy, rum, cognac, whisky, vodka or other spirits		
5 ( ) soft drinks mixed with alcoholic drink	+ ⑤ a respondent who most of the time drinks light or nonalcoholic beer but occasionally drinks alcoholic drinks, is unjustly routed towards Q. 5	- ④ R1 avoids this category although it would be the right one; R7 and R13 take this for want of better
6 ( ) it depends and varies strongly ④	- ⑥ non-alcoholic beer is listed here as an alcoholic drink; this may cause overreporting	+⑤ R9 sometimes drinks light beer and goes to Q.5 although he drinks normal beer also
7 ( ) nonalcoholic ⑤ or light ⑥ beer → go to Q. 5		
8 ( ) I never drink alcohol → go to question 5		

[*] An identified problem is marked with "+" if it also was identified in the other study and with "-" if it was not revealed in the other study

Question 2.	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,	+② R's probably will not assess the period of "six months" accurately	+② R4 and R7 assess the past half year accurately; others don't.
did you ever drink six or more ③ glasses ④ of alcoholic beverages ⑤ on one day ⑥ ?	+② especially R's with high drinking frequencies probably will extrapolate from last week to six months	- 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	+③ not every R counts glasses when drinking	+③ R7 takes this as a question for occasions of drinking significantly more than average – in her case still far below six glasses
4 ( ) Yes 1 or 2 times a week ⑦	+③ "six glasses" may be taken as drinking a lot	
5 ( ) Yes, 1 to 3 times a month 6 ( ) Yes, 3 to 5 times this half year	+④ it is known that people at home often drink beer from bottles and other drinks from non-standard glasses	+④ R8, R15 and R16 count bottles or big glasses
7 ( ) Yes, 1 or 2 times this half year 8 ( ) No, never this half year	+⑤ as a selection effect from question 1 some R's will not count all types of alcoholic drink	+⑤ R8 asks whether this concerns beer only or other drinks as well
9 ( ) Don't know	- ⑥ some R's probably count drinks by occasion rather than by day	- 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 other study and with "-" if it was not revealed in the other study

Question 3. ①	expert review	field TSTI
How many days a month do you drink② on average ③ ? ④ ⑤	±① order: it seems more adequate to place this question before Q 2 (6+ days)	±① the previous Q probably stimulates to think only of days with large drinking quantities – which some R's do
1 ( ) 28 or more days		- ② R8, R9 and R12 take "days you drink" as days of heavy drinking. A beer with dinner is no drinking to them. R14 feels strange in marking "28 or more" of drinking although he takes one gin every day
2 ( ) 24 - 27	+② see at ⑥	
3 ( ) 21 - 23 ⑥	- ④ there is no explicit restriction to alcoholic drinks; but probably this goes without saying	- ③ R3 has difficulty in calculating a mean because of large variation in his drinking pattern
4 ( ) 15 - 20	+⑤ no period is defined here; this may cause use of unjust reference periods	+⑤ all R's take a period of several months, but: R8 changed his habit some weeks ago; he takes this last week's drinking. R7 stresses the difference between summer and wintertime drinking. R5 reports a change of his pattern during this period
5 ( ) 12 - 14	+⑥ these categories suggest a high precision; some R's will probably start with a week's estimate and then extrapolate by multiplying with four. In that case these categories seem unwieldy	
6 ( ) 9 - 11		
7 ( ) 6 - 8		
8 ( ) 3 - 5		+⑥ most respondents extrapolate from a weekly pattern. R4 calculates from "once in 4 to 5 days" towards # of days a month
9 ( ) 2 or less		

* An identified problem is marked with "+" if it also was identified in the other study and with "-" if it was not revealed in the other study

Question 4.	expert review	field TSTI
If you drink alcohol on a day ① *how many glasses do you take on average*② *(round up half glasses*③ *)?* 1 ( ) 11 or more glasses 2 ( ) 7-10 glasses 3 ( ) 6 glasses ④ 4 ( ) 4-5 glasses 5 ( ) 3 glasses 6 ( ) 2 glasses 7 ( ) 1 glass	+② it seems questionable whether R takes "average" as the mean or as a median or modal value; especially with diversity in quantity this will cause problems +③ the distinction between whole and half glasses will probably appear unmanageable +④ different ranges in categories may be disturbing	-① R1 does not count per drinking day, but per day. R5 chooses a modal drinking day, = a day when she drinks wine; on a "beer day" she drinks a lot more. R8 counts only days of *heavy* drinking +② R4 calculates a real arithmetic mean; all others report a median or modal value - ② R1 and R7 give a "normative" answer, i.e. what they strive for - ② R10 reports her maximum (3 glasses) which is quite higher than the mean +③ R1 and R5 notice that this distinction does not work; R8, R15 and R16 often drink from big glasses. +④ R7 comments on the "construction of answering categories"

* An identified problem is marked with "+" if it also was identified in the other study and with "-" if it was not revealed in the other study

Questions 5 & 6	expert review	field TSTI
5. *Did you have periods during the past ten years* ① *of drinking significantly* ② *more or less than nowadays* ③ *?* 1 ( ) no, my drinking pattern has approximately ④ remained the same 2 ( ) yes, I have been drinking significantly more than nowadays 3 ( ) yes, I have been drinking significantly less than nowadays	+① the reference period "past ten years" overlaps with "past half year" in previous questions as the reference period for actual drinking pattern +② "significantly" and +④ "approximately" is vague, giving little guidance to R	- ① R1, R7 and R9 compare past ten years as a whole to the period before ten years ago. +① R8 draws the line between "past 10 years" and "nowadays" at 2 months ago when he diminished drinking +② the term "significantly" is interpreted differently. R5 selects an extreme period (year living in U.S.A.). R3 in "think aloud": what a word; there are different periods depending of taking courses for example. - ③ R10 comes up against a wrong routing. After the answer "no" (code 1) she unjustly goes to Q 6 which does not apply in her case.
6. Can you tell globally ① *how long a period you were drinking significantly* ② *more or less than nowadays?* .... Years .... Months	- ① "globally" is vague - ② "significantly" also	

* An identified problem is marked with "+" if it also was identified in the other study and with "-" if it was not revealed in the other study

6. References

Bongers, I.M. and Van Oers, J.A. (1998). Mode Effects on Self-reported Alcohol Use and Problem Drinking: Mail Questionnaires and Personal Interviewing Compared. Journal of Studies on Alcohol, 59, 280–285.

Campanelli, P. (1997). Testing Survey Questions: New Directions in Cognitive Interviewing. Bulletin de Methodologie Sociologique, 55, 5–17.

Conrad, F. and Blair, J. (1998). From Impressions to Data. Increasing the Objectivity of Cognitive Interviews. Paper presented at the 1996 Annual Meeting of the American Statistical Association.

Cutler, S.F., Wallace, P.G., and Haines, A.P. (1988). Assessing Alcohol Consumption in General Practice Patients–A Comparison Between Questionnaire and Interview (findings of the Medical Research Council's general practice research framework study on lifestyle and health). Alcohol, 23, 441–450.

Del Boca, F.K. and Noll, J.A. (2000). Truth or Consequences. The Validity of Self-reported Data in Health Services Research on Addictions. Addiction, 95, S347–S360.

Ericsson, K.A. and Simon, H.A. (1993). Protocol Analysis. Verbal Reports as Data. Cambridge, MA: MIT Press.

Garretsen, H.F.L. (1983). Probleemdrinken. Prevalentiebepaling, beïnvloedende factoren en preventiemogelijkheden. Nijmegen: Katholieke Universiteit Nijmegen. [In Dutch].

Garretsen, H.F.L. and Knibbe, R.A. (1983). Alkohol prevalentie onderzoek Rotterdam/Limburg: Ministerie WVC. [In Dutch].

Glaser, B. and Strauss, A. (1967). The Discovery of Grounded Theory. Chicago: Aldine.

Hak, T. and Jansen, H. (1997). Het meten van alcoholconsumptie per postenquête. Een onderzoek naar cognitieve problemen bij de beantwoording van vragen naar frequenties en hoeveelheden. Rotterdam: IVO. [In Dutch].

Hak, T., Van der Veer, K., and Ommundsen, R. (forthcoming). An Application of the Three-Step Test-Interview (TSTI). A Validation Study of the Dutch and Norwegian Versions of the Illegal Aliens Scale. International Journal of Social Research Methodology: Theory and Practice.

Houtkoop-Steenstra, H. (2000). Interaction and the Standardized Survey Interview. The Living Questionnaire. Cambridge: Cambridge University Press.

Köhnken, G., Schimossek, E., Aschermann, E., and Höfer, E. (1995). The Cognitive Interview and the Assessment of the Credibility of Adults' Statements. Journal of Applied Psychology, 80, 671–684.

Lahaut, V.M.H.C.J., Jansen, H.A.M., Mheen, D. van de, and Garretsen, H.F.L. (2003). Comparison of Two Formats of the Weekly Recall and Quantity-frequency Alcohol Measures. Journal of Substance Use, 8, 164–169.

Lemmens, P.H.H.M. (1991). Measurement and Distribution of Alcohol Consumption. Maastricht: Rijksuniversiteit Limburg.

Lemmens, P.H.H.M., Tan, E.S., and Knibbe, R.A. (1992). Measuring Quantity and Frequency of Drinking in a General Population Survey. A Comparison of Five Indices. Journal of Studies on Alcohol, 53, 476–486.

Loftus, E.F., Fienberg, S.E., and Tanur, J.M. (1985). Cognitive Psychology Meets the National Survey. The American Psychologist, 40, 175–180.

Means, B., Swan, G.E., Jobe, J.B., and Esposito, J.L. (1994). The Effects of Estimation Strategies on the Accuracy of Respondents' Reports on Cigarette Smoking. Autobiographical Memory and the Validity of Retrospective Reports. N. Schwarz and S. Sudman (eds). New York: Springer Verlag.

Midanik, L. (1982). The Validity of Self-reported Alcohol Consumption and Alcohol Problems. A Literature Review. British Journal of Addiction, 77, 357–382.

Midanik, L.T. and Hines, A.M. (1991). Unstandard Ways of Answering Standard Questions. Protocol Analysis in Alcohol Survey Research. Drug and Alcohol Dependence, 27, 245–252.

Midanik, L.T., Hines, A.M., Greenfield, T.K., and Rogers, J.D. (1999). Face-to-face Versus Telephone Interviews. Using Cognitive Methods to Assess Alcohol Survey Questions. Contemporary Drug Problems, 26, 673–693.

Mingay, D.J., Shevell, S.K., Bradburn, N.M., and Ramirez, C. (1994). Self and Proxy Reports of Everyday Events. Autobiographical Memory and the Validity of Retrospective Reports. N. Schwarz and S. Sudman (eds). New York: Springer Verlag.

Rehm, J. and Spuhler, T. (1993). Measurement Error in Alcohol Consumption. The Swiss Health Survey. European Journal of Clinical Nutrition, 47, S25–S30.

Reinhard, S.C. and Horwitz, A.V. (1995). Caregiver Burden. Differentiating the Content and Consequences of Family Caregiving. Journal of Marriage and the Family, 57, 741–750.

Romelsjö, A., Leifman, H., and Nyström, S. (1995). A Comparative Study of Two Methods for the Measurement of Alcohol Consumption in the General Population. International Journal of Epidemiology, 24, 929–936.

Sirken, M., Hermann, D., Schechter, S., Schwarz, N., Tanur, J., and Tourangeau, R. (1999). Cognition and Survey Research. New York: Wiley and Sons.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). Thinking About Anwers. The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive Sciences and Survey Methods. In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds). Washington DC: National Academy Press.

Verdurmen, J., Toet, J., and Spruit, I. (2000) Alcohol- en druggebruik in de gemeente Utrecht. Utrecht: Trimbos instituut. [In Dutch].

Received November 2002
Revised September 2004