Bidirectional causality in methods research of interviews with standardized questionnaires: Anticipation and repair as sources of interviewer effects

Johannes van der Zouwen[1] , Johannes H. Smit[2] and Stasja Draisma[2]
Vrije Universiteit, Faculty of Social Sciences, Amsterdam, The Netherlands[1]
Vrije Universiteit Medical Centre, Department of Psychiatry, Amsterdam, The Netherlands[2]

Paper submitted for presentation at the XVIth World Congress of Sociology, in the joint session on 'Causality Revised' of RC-51 and RC-33

Abstract

The paper explores the conditions under which methods research concerning standardized questionnaires can be based on unidirectional causality only, and when it should use designs that can also handle bidirectional causality.

If the questions in the questionnaire become quite difficult, self-administered questionnaires and computer-assisted telephone interviews have to be replaced by personal or diagnostic interviews. Here the interviewers are instructed to make attempts at repair of inadequate initial answers and to anticipate problems respondents will encounter, by making the questions less difficult to answer.

The exercise of these additional control tasks by the interviewers is investigated by a detailed analysis of transcribed question-answer sequences of part of a standardized interview with 233 elderly people. The questionnaire contained detailed retrospective questions about their physical activities. The four interviewers were professional research nurses.

The hypothesis is tested, and confirmed, that the more difficult the topic of the question is for this particular category of respondents, the more the interviewers will initiate repairs and use a deviating questioning strategy and, thus, the lower the proportion of paradigmatic sequences will be. There is a large interaction effect with interviewer number and only a very small effect with respondent characteristics like age and physical condition.

Keywords: standardized interviews; repair by interviewers; anticipation by interviewers; decomposition of questions

Contact: j.van.der.zouwen@fsw.vu.nl

1. Introduction

In social science a discussion is going on about the most appropriate conception of causality: is the unidirectional conception of causality, symbolized by X→Y, sufficient, or even the only possibility? Or does the particular character of social processes makes it necessary to use also a bidirectional form of causality, symbolized by X←→Y ? (Van der Zouwen and Van Dijkum, 2001; Van Dijkum, 2006).

In our view these questions have an empirical aspect, for part of an answer can be found by conducting empirical research. To illustrate this view, we explore the conditions under which methods research concerning standardized questionnaires, can be based on unidirectional causality only, and under what conditions this research should use designs that can handle bidirectional causality.

This paper also shows how it is possible to deal with forms of bidirectional causality in research after anticipation and repair as sources of interviewer effects in standardized questionnaires with rather difficult questions.

2. Different Questions, Different Modes of Administration

In social sciences, like sociology and public opinion research, most data are obtained by using standardized questionnaires. The questions asked in standardized questionnaires strongly differ with respect to the difficulty respondents will have to answer them properly. These differences in 'task difficulty' [1] partly determine the mode of administration of the questionnaire.

Sometimes the questions can be answered so easily, that it suffices to distribute the questionnaire in class, by mail or via internet. However, researchers using the *self-administered questionnaire* do not have the possibility to identify and correct misunderstandings by respondents.

If the questions are less easy to answer, or if part of the respondents will have difficulties with reading and responding to a written questionnaire, the researcher has to call in interviewers. The fastest and cheapest interviewing method is the *computer assisted telephone interview,* where dozens of interviewers phone the respondents from a call center (e.g., in public opinion polls). The task of these interviewers is normally restricted to reading out the questions and entering the response codes into the central computer.

If the response task becomes still more difficult, for example if respondents have to reconstruct their social network or their professional career, or if many of them have problems with understanding the questions, a far more expensive method has to be chosen, i.e., the *personal interview,* conducted 'face to face' at the respondent's home. Here the interviewers have to give explanations if requested, and probe ('repair') inadequate initial responses; without biasing the information obtained. These tasks have to be fulfilled by well-trained (*free lance) interviewers*.

If it is to be expected that the majority of the respondents, for example due to the topic of the questionnaire or due to sample characteristics, like old age, has problems with

understanding and answering the questions, the researcher has to call in *professional interviewers*, who are able to guide the respondents through the questionnaire, and therefore receive the most extensive training. The interviewer has sometimes to adjust the wording of the question to the particular situation of the respondent, and to interpret the answers given by the respondents. In case these interviews are conducted for diagnostic purposes, they are called *diagnostic interviews*.

3. Different Method Effects, Different Research Designs

A central topic in methodology is the effect of the method of research on the result obtained by this method; method effects for short. In this case: the effect of the way in which the questionnaire is administered, on the information obtained; or shortly: 'mode effects' (De Leeuw and Van der Zouwen, 1988; De Leeuw, 1992). One has to bear in mind that the following classification is rather schematic: in survey practice the borders between different modes of administration are less rigid than suggested below.

In *self-administered questionnaires* the method of data collection about equals the questionnaire constructed by the researcher. Various characteristics of the entire questionnaire and of the separate questions (X), have been proven to affect the responses (Y); there exist question or wording effects [X→Y] (e.g., Groves, 1989; Schuman and Presser, 1996). The reverse, effects of the response distributions on the questions, is impossible here. Therefore, studies of method effects can be conducted with a research design based on unidirectional causality, i.e., the *split ballot design* in which different versions of the questionnaire (X1, X2) are randomly distributed over the respondents. Next it is investigated what the differences are between the response distributions of these different sub-samples (Y1, Y2), and what question characteristics lead to the largest differences (see Box 1).

Box 1. Task of the interviewers, related modes of administration, assumed causal relationships, and research design of the corresponding methods research

Task of the interviewers	Related mode of administration	Causal relations	Research design of methods research
No interviewers present	Self-administered questionnaire	X→Y	Split ballot with different versions of the questionnaire
Reading out questions Recording responses	Computer-assisted telephone interview	Xt1→Yt2	Split ballot with different (types of) interviewers
Like above plus: Giving explanations Probing inadequate initial responses	Personal interview	Xt1→Yt2 Yt2→Xt3	Sequence analysis of strings of coded behaviors
Like above plus: Adjusting wording of the questions Interpreting the answers	Diagnostic interview	Xt1→Yt2 Yt2→Xt3 Y*t1→Xt2	Sequence analysis of strings of coded behaviors coupled with information about individual respondents

In *computer-assisted telephone interviews,* not only the questionnaire, but also the interviewer may have an effect on the responses; for example because respondents, while answering, reckon with the gender or estimated age of the interviewer. Research after these methods effects also uses the *split ballot design*. Every respondent is interviewed with the same questionnaire, but interviewers with different characteristics, e.g., male or female, (X1, X2), are randomly assigned to the respondents. The response distributions of these sub-samples are compared in order to assess this type of interviewer effect: [Xt1→Yt2].

In *personal interviews,* the interviewers give explanations if asked for by the respondents, and probe inadequate initial answers. The behavior of the interviewer (X) is partly dependent on preceding behavior of the respondent (Y). This means that besides the relationship Xt1→Yt2, there exists a bidirectional causal relationship, the feedback loop Yt2→Xt3. To disentangle these two causal relationships the split ballot design does not fit anymore. What is needed is a precise description of the question-answer process, based on a (audio) *recording* of the process, a coding of relevant actions of interviewer and respondent *(behavior coding*), and an analysis of strings of subsequent actions *(sequence analysis*) (see e.g., Dijkstra, 1999, 2002, 2005).

In *diagnostic interviews,* interviewers not only have to react to behavior of the respondent; they also have to anticipate problems the respondent will encounter when answering the question, and to adjust the question wording to avoid these potential problems. The expected behavior of the respondent (Y*), like giving a 'don't know answer', affects the present behavior of the interviewer, via the feedforward loop Y*→X. Thus, in diagnostic interviews three types of causal relations are present Xt1→Yt2; Yt2→Xt3, and Y*t1→Xt2. The disentanglement of these three causal relationships requires a method of analysis that can detect anticipatory behavior, and background information – 'feeding' these anticipations - of the individual respondents. But before discussing this method of analysis, we will first present some theoretical notions borrowed from cybernetics about control processes in standardized interviews.

4. Different Forms of Control in Standardized Interviews

A standardized interview can be viewed as a controlled communication process involving three actors (Van der Zouwen and Smit, 2005, 2006; Van der Zouwen, 2006).

[Figure 1 about here]

The first one is the *researcher*, who constructs the questionnaire, instructs the interviewers, and analyzes the collected responses. The second one is the *interviewer*, who reads out the questions from the questionnaire, and evaluates the initial answers of the respondent. If an initial answer is inadequate, e.g., not complete, the interviewer has to probe further until she [2] obtains an adequate answer. In the middle of Figure 1 the third actor is located, the *respondent*, who by means of an internal cognitive process, interprets the question, searches for relevant information, and reports his answer.

The activities of each of these three actors are aimed at a specific goal. The goal of the researcher is to get comparable, relevant and unbiased information about the respondents. To reach this goal he has to control the interview process.

In cybernetics different forms of control are distinguished. The first form of control is *open loop control*, exercised by the researcher, in Figure 1 depicted by the arrows leading from box 1, via boxes 2, 3 and 4, to box 5. Ideally, in a standardized interview only open loop control takes place. Then the interview process directly proceeds from phase 1 till phase 5: The question as worded by the researcher (phase 1) is posed about exactly so by the interviewer (phase 2), and leads to the cognitive process within the head of the respondent (phase 3), resulting in an adequate answer (phase 4), correctly noted by the interviewer (phase 5), and entered into the data file of the researcher.

But sometimes things go wrong, for example in the cognitive process within the respondent, leading to an inadequate answer. In that case, the interviewer is instructed to make an attempt at repair, for example by asking a probing question. In that case there emerges a *feedback loop* from phase 4 back to phase 2: the respondent is asked to reconsider and redo his initial answer (Moore and Maynard, 2002).

Previous research (Van der Zouwen and Smit, 2005) has shown that sometimes the interviewers 'override' the open loop control of the researcher. If they expect that the question as worded in the questionnaire will lead for this respondent to inadequate answers, they may rephrase the question such that the question becomes easier to answer, hopefully leading to an adequate initial answer (Houtkoop-Steenstra, 1995). This anticipatory behavior of the interviewer, based on expectations (Hyman et al. 1954, 1975) about the respondent (box 6), is depicted by another arrow from box 2 to box 3: the 'official' question is replaced by one 'invented' by the interviewer. This alternative control can be described as a *feedforward* loop: on the basis of the expected inadequate answer of the respondent to the 'official' question by the researcher, the interviewer constructs a question that hopefully avoids the necessity of repair.

These three different forms of control of the interview process lead to different interactions between the interviewer and the respondent, resulting in different strings or sequences of verbal utterances of both actors. If open loop control is successfully exercised, the resulting sequence can be labeled as '*paradigmatic*', that is, the "ideal sequence from the standpoint of survey design" (Schaeffer and Maynard, 1996, p. 66). It ".. consists of the interviewer posing the question as worded in the questionnaire, followed by an adequate answer of the respondent. The interviewer may close the sequence with repeating the answer or saying something like okay" (Dijkstra, 2005, p. 612). In a paradigmatic sequence the number of turns is very small: two or three.

If the initial answer of the respondent is inadequate and the interviewer does an attempt at repair, the question-answer sequence becomes at least two turns longer. So the exercise of feedback control, or repair, shows itself by a longer question sequence.

Feedforward control by the interviewer, or anticipation, has different forms. A very common form is adding to the question itself a kind of candidate answer, making it a suggestive question. If the candidate answer is about correct, the resulting sequence

will be as long as the paradigmatic one. Another form of anticipation is that the interviewer decomposes the topic or subject of the 'official' question into several components, and asks the 'official' question about each of these components. Examples of this *partial questioning strategy,* 'invented' by the interviewers are reported in preceding papers (Van der Zouwen and Smit, 2005, 2006). This form of anticipatory interviewer behavior will lead to long question-answer sequences, because for every component of the question topic a separate question must be asked and answered.

From this short exposition it appears that the type of control exercised within the interview is reflected by the length of the resulting question answer sequence: successful open loop control leads to a short paradigmatic sequence, repair leads to some additional turns, and feedforward control, when using a partial questioning strategy, leads to many additional turns, and thus to a much longer sequence.

The preceding statements about the effects of the difficulty of the question and the task related capacity of the respondent on the control over the interview process as exercised by the interviewer, in its turn affecting the length of the sequence, are reflected in Figure 2.

[Figure 2 about here]

Since the main subject of this paper is repair and anticipation – by partial questioning - of interviewers, we focus the analysis on sequences that are significantly longer than the paradigmatic ones. The method of analysis employed will be illustrated by data from a survey after physical activities of elderly people.

5. The LAPAQ Questionnaire

We chose this material because it was shown by Schwarz et al. (1998) that elderly people often have difficulties with answering survey questions. Moreover, that part of the interview we selected for the analysis, mainly consisted of rather detailed retrospective questions about the frequency and duration of quite 'mundane' activities like walking or housekeeping, that is questions that are rather difficult to answer (Knaüper et al., 1997; 2004). Nevertheless, the proportion of 'don't know answers' and refusals was remarkably low. It seems as if the four research nurses who interviewed these elderly respondents did all they could to get eventually adequate answers to these difficult questions. Thus we expected that much repair and partial questioning has taken place here.

The topic of the survey, from which part of the data is used, is "the prevention of fall accidents in older persons" (Stel, 2003). It has been shown in the literature that the performance of physical activities is an important factor in the reduction of falls and fractures (Graafmans, Bouter and Lips, 1998). For that reason, a couple of instruments was used in this survey to measure these activities (Stel et al., 2004).

One of these instruments is called the LAPAQ, a questionnaire for face-to-face interviews that covers the frequency and duration of walking outside, bicycling,

gardening, light and heavy household activities, and sport activities during the previous two weeks.

The survey was performed within the framework of LASA, the Longitudinal Aging Study Amsterdam (Deeg and De Serière, 1994). For this survey, respondents were recruited from 1509 men and women, aged 65 years and older as of January 1 in 1996, and who participated in a follow-up study on falls within LASA. Subjects who completed the third data-collection cycle within LASA in 1998/1999 and fell at least once between January 1998 and January 1999 (n=328) were eligible for this additional study, as well as a random sample of 196 subjects who did not fall during that year [3].

One year after the third data collection of LASA in 1998/1999, four trained research nurses visited the participants two times at home. During the first visit, they administered among other things the LAPAQ questionnaire.

Concerning each topic of the LAPAQ questionnaire (abbreviated with X), a few questions are asked. The first question is the *base question* "Do you do X?" If the answer is positive, then for four of the six topics the base question is followed by "In the past two weeks, did you do X?." If the answer to this question is positive (too) then the *frequency question* is asked: "How many times did you engage in X during the last two weeks?" This series of questions about this topic is closed by the *duration question* "How long, on average, were you engaged in activity X?" (see the Appendix for the exact wording of the questions) [4].

The series of subsequent questions, responses and remarks related to one of the six topics is called a *macrosequence*. A macrosequence contains the question-answer sequences of the separate questions for each topic; see Box 2 for an example [5].

Box 2. An example of a coded macrosequence (# 11826-1)

01. Ib: I now want to ask you some questions about your physical movements pattern. Do you walk now and then? And by 'walking' we mean walking outside, to visit someone or to go shopping.
02. Rb: No, I only walk for my pleasure.
03. Ib: Yes, but you never go out shopping?
04. Rb: Yes.
05. Ib: So that happens now and then. And if you visit somebody; then you walk to that person? Your sister-in-law lives nearby, and then you go there on foot?
06. Rb: Yes
07. Ip: Have you walked during the past two weeks?
08. Rp; No, not much, because it was quite windy.
09. Ip: Yes, but you have done it?
10. Rp: Yes
11. If: And how many times did you walk during the past two weeks?
12. Rp: No
13. Ip: And not to your brother?
14. Rp: Yes, but that is here.
15. Ip: But then you walk through the corridors?
16. Ri: Yes and he comes along too.
17. Ip: Yes, but how often have you visited him the last two weeks?

18. Rf: Two or three times.
19. If: The last fourteen days; how often have you been there; two times, three times?
20. Rf: I come there once every day.
21. If: Everyday, also Saturday and Sunday?
22. Rf: Yes.
23. If: That is 14 times then.
24. Id: And how long do you walk then? If you walk through the corridors how long is that; a few minutes or is it ten minutes?
25. Rd: I walk a quarter of an hour back and forth.
26. Id: I mean in the house, because the house is so big. Because if you go to your brother you walk through the corridors and these corridors are quite long.
27. Ri: Yes, but I do not go up the stairway.
28. Ii: No, then you go with the elevator.
29. Id: But you walk daily about fifteen minutes?
30. Rd: Yes, that's right.

If a macrosequence contains all questions and answers about the topic, it is called a *complete* macrosequence; otherwise it is called *incomplete*. For example if the respondent answers with 'no' to the base question ("Do you do X?") and the interviewer goes to the next topic, the result is an incomplete macrosequence, since the follow up questions and answers do not occur in the sequence.

6. Content Analysis of Long Macrosequences

As explained above, repair and use of a partial question strategy by the interviewer, will especially occur in longer question-answer sequences, and thus also in longer macrosequences. Therefore, to study repair and partial questioning in more detail, we selected for each of the six topics of the LAPAQ questionnaire the ten longest macrosequences. These sixty long macrosequences contained many more turns of interviewer and respondent than the corresponding set of paradigmatic sequences. A content analysis of these macrosequences showed the following six main 'causes' of these additional turns:

1. *Definition problems*: A discussion takes place between respondent and interviewer about the definition and the delineation of the physical activity mentioned in the questionnaire (for example turns 02 to 05 in Box 2). These problems especially occur with the topics 'walking' and 'sport'. Neither from the question, nor from its explanation in the questionnaire, it is quite clear what activities fall under the label 'walking'; e.g., is taking out the dog, or shopping to be viewed as 'walking' or not? Similar problems arise with the topic 'sport': is walking a sport, or a means of going from one place to another?
2. *Partial questioning and probing*: The interviewer – deviating from the questionnaire – rephrases the questions and decomposes the activity into individual tasks within household work, or into different destinations for walking or cycling (turns 05 and 13). Partial questioning especially occurs with questions about activities that are hard to answer in its entirety. For 'gardening' and 'sport' there seems to be no need for using the partial questioning strategy.

3. *Completion probing*. If the interviewer uses a partial questioning strategy, it is important that all relevant household tasks, or all relevant walking or cycling destinations are dealt with. This explains the frequent use of probing questions like: "And are there other tasks you do?" Much of this probing occurs in a leading manner "And to the doctor, you also walk?" (e.g., turn 13).
4. *Confirmation and check.* Some answers of respondents are unclear, or – at least to the interviewer – not very probable. Therefore, some interviewers are inclined to check the initial answer for clarity and validity (e.g., turn 23). This interviewer behavior is quite independent of the topic of the question; it is rather related to characteristics of the interviewer.
5. *Conferring* by interviewer and respondent about what is, given the particular situation of this respondent, the best fitting answer to the question. For example, in how far has time spent with 'doing the laundry' to be counted, if most of the work is done by the automatic washer? Or, on what type of sport does the respondent spend more time, and thus has to be questioned first? (see also turns 24-29).
6. *A complex situation* of the respondent. The frequency questions and the duration questions assume a kind of regularity in the performance of the activities. However, sometimes the behavior of the respondent shows so much irregularity, at least over the last two weeks, that it is difficult to gain a clear view of his physical activities, which leads to additional turns to clarify the situation. This especially occurs with the topics 'walking', 'cycling' and 'household tasks'.

From Table 1 it appears that some 'causes' of additional turns in long macrosequences are very unevenly distributed over the six topics of the LAPAQ questionnaire, adding support to the hypothesis that the control behavior (especially partial questioning and the related completion probing) of interviewers will be quite topic specific.

Table 1. The different 'causes' of extra turns in long macrosequences, for each of the six topics

	Walking	Cycling	Gardening	Sport	Light HW	Heavy HW	Total
1. Definition problems	2	0	1	3	0	0	6
2. Partial questioning and probing	8	8	0	0	10	10	36
3. Completion probing	6	6	0	1	0	1	14
4. Confirmation and check	2	2	1	2	2	1	10
5. Conferring between I and R	3	4	4	8	2	1	22
6. Complex situation of R	2	3	0	1	4	3	13
Total number of 'causes'	23	23	6	15	18	16	101
Number of analyzed macrosequences	10	10	10	10	10	10	60

7. The Coding Scheme

To test the hypothesis about the topic specificity of task related interviewer behavior we need interviews with transcribed macrosequences for all six topics. This requirement is fulfilled in 233 interviews, so the database for the quantitative analysis consists of 6*233 = 1398 macrosequences.

Each macrosequence is divided into a series of successive turns, that is speech acts of the interviewer (I), the respondent (R); and possibly a third person (T), but only if he or she interferes with the answering process. The turns are coded with a coding scheme, using the SEQUENCE VIEWER software, developed by Dijkstra (1999, 2002). The coding scheme consists of only two variables. The first variable regards the ACTOR to whom the turn can be attributed (I, R or T). The second variable concerns the QUESTION from the questionnaire that is under discussion in this turn.

In Box 3 the coding scheme for the variable QUESTION is presented. The question numbers q1, q5 etc. refer to the LAPAQ questionnaire presented in the Appendix.

Box 3. Codes of the coding variable QUESTION and the related questions from the LAPAQ questionnaire

b: Do you do X? (q1, q5, q9, q15; q24, q27)
p: Did you do X during the past two weeks? (q2, q6, q11, q17)
f: How many times did you do X during the past two weeks? (q3, q7, q12, q17, q25, q28)
d: How long did you usually do X each time? (q4, q8, q13, q18, q26, q36)
h : Questions and answers related to a particular task in the household (partial questioning strategy)
m: Two questions or topics are mixed into one turn
i: Questions or answers that are irrelevant to the topic

8. The Number of Additional Turns of a Macrosequence

8.1 Correcting for the different number of questions to be asked

The length of a macrosequence, as measured by the number of turns it contains, not only depends on task related behavior of interviewer and respondent, but also on the number of questions from the questionnaire that have to be asked and answered. So to make valid comparisons between different types of respondents and between different topics, the length of the macrosequence has to be corrected for differences regarding the number of questions that have to be asked. This last number depends on the answer of the respondent to the 'base question', i.e., "Do you do X?". If the answer is 'no', then the interviewer must skip the follow-up questions and proceed to the next topic of the questionnaire. If the answer is 'yes', then for the first four topics the follow-up question "Did you do X during the past two weeks?" has to be asked. If the answer to this question is negative the interviewer must proceed to the next topic, otherwise the frequency and duration question have to be posed.

As each question requires at least two turns to be asked and answered, we can subtract from the length of each macrosequence (LM) the number of turns it would contain if the macrosequence only consisted of paradigmatic question answer sequences.

In Box 4 an overview is presented of the number of turns that have to be subtracted from the length of the macrosequence (LM) to obtain the number of 'extra' or 'additional' turns (AT). It is possible that the score on the variable AT is negative, for example if the questions are asked very suggestively, e.g., "You do X, I understand", and the respondent does not respond to this suggestion. Or if the respondent answers the question ""Do you cycle?" by saying "Yes, I cycle everyday, about half an hour", so answering four questions in one turn.

Box 4. Number of turns to be subtracted from the length of the macrosequence (LM) to obtain the number of additional turns (AT)

Topics	Answer to 'base question': "Do you do X?"	Answer to 'past question': "Did you do X in the past two weeks?"	Number of turns to be subtracted from LM
Walking, cycling, gardening, or sport	No	--	2
	Yes	No	4
	Yes	Yes	8
Light or heavy household work	No	--	2
	Yes	--	6

8.2 Distribution of AT by topic

In Table 2 the distribution of the AT score for each of the six topics of the LAPAQ questionnaire is presented for all 233 interviews. About five percent of the macrosequences have a negative AT-score, meaning that the observed sequence is even shorter than the paradigmatic one. This especially occurs with the topic 'gardening'; for example, as the interviewer rephrases question q9 "Do you have a garden?" for respondents living in a flat as "I see that you don't have a garden". Twenty-one percent of the macrosequences has an AT-score of zero; they contain no additional turns. This percentage is much higher (41%) for the topic 'cycling'. And another twenty-one percent has only a small AT-score, with three or even less additional turns.

So about half (48%) of the macrosequences have so few additional turns that it is unlikely that repair, let alone partial questioning, has taken place. Probably the main form of control here is the open loop control by the researcher, as implemented by the interviewer [6].

Table 2. Distribution of the number of additional turns of the macrosequence (AT) for each of the six topics of the LAPAQ questionnaire

AT score	Walking	Cycling	Gardening	Sport	Light HW	Heavy HW	Total
Less than 0	4	15	36	4	9	7	75
0	8	95	79	44	20	49	295
1-3	25	34	79	55	50	56	299
4-7	53	40	33	82	57	44	309
8-11	46	20	5	28	40	34	173
12-15	40	10	1	6	19	14	90
16-19	20	5	0	8	3	14	50
20-23	15	7	0	3	7	5	37
24-27	7	2	0	2	10	4	25
28-31	7	2	0	1	11	4	25
More than 31	8	3	0	0	7	2	20
Total	233	233	233	233	233	233	1398

Chi-square = 447.7; df = 50; $p<0.001$**

The other half of the macrosequences have AT-scores of 4 turns or more and therefore are certainly not paradigmatic: it is quite likely that some repair or partial questioning has taken place. The proportion of these longer sequences is very unevenly distributed over the six topics of the questionnaire: it varies from 17% (gardening) to 84% (walking).

An AT-score of 16 means that the macrosequence is at least twice as long as the paradigmatic one. Such really long sequences did not show up with the topic gardening. On the other hand, in 16%, respectively 25%, of the sequences concerning 'light household work' and 'walking', such high AT-scores were obtained, indicating that especially these topics require a lot of additional control activities of the interviewers.

9. The Three Sequence Types

Whether the additional turns in a relatively long macrosequence are generated by a partial questioning strategy of the interviewer or not, can be seen from the coding of the second coding variable of the individual turns. For the topics light and heavy household work, the presence of code 'h' indicates that partial questioning has occurred. For the other four topics, partial questioning is indicated by a repetition of the code-string f-d, that is, a frequency question followed by a duration question (for one component of the question), and again succeeded by a frequency and duration question for another component. If neither code 'h' or the code string f-d…f-d does occur in a relatively long macrosequence, we may conclude that only repair activities have led to the additional turns.

By this reasoning we end up with three different types of macrosequences:
Type A: relatively short macrosequences (AT maximally 3) with only open loop control;

Type B: relatively long macrosequences (AT more than 3) without partial questioning, so only open loop control and repair activities; and
Type C: relatively long macrosequences (AT more than 3) with partial questioning and open loop control, and possibly also repair activities.

Of all 1398 macrosequences, 669 (48%), belong to Type A (only open loop control), 451 (32 %) belong to Type B (open loop control plus repair activities) and 278 (20%) are of Type C (open loop control plus partial questioning and possibly repair activities).

10. The Relationship between Type of the Macrosequence and Question Topic

Table 3 presents the distribution of the macrosequences over the three types, for each topic separately. As hypothesized, there are indeed striking differences between the topics of the questionnaire with respect to the kind of control exercised by the interviewers.

Table 3. The distribution of the macrosequences over the three Types of Control per Question Topic

Topic	Type A open loop control only	Type B open loop control and repair	Type C open loop control, partial questioning, and possibly repair	Total	N
Walking	16%	68%	16%	100%	233
Cycling	62%	31%	7%	100%	233
Gardening	83%	15%	1%	100%	233
Sport	44%	52%	3%	100%	233
Light HW	34%	19%	47%	100%	233
Heavy HW	48%	7%	45%	100%	233
Total	48%	32%	20%	100%	1398

Chi-square = 585.3; G-square = 603.7; df = 10; p<.001**; Cramer's V= 0.46

Open loop control seemed to work quite well with the topic of 'gardening' (83%), and – to a lesser degree (62%)- with 'cycling', resulting in generally short macrosequences.

Feedback control, in the form of repair activities, is very much exercised with the topic of 'walking' (68%) mainly because the definition of 'walking' is not quite clear to the respondent (and sometimes not even to the interviewer).

Feedforward control in the form of partial questioning is especially used with the topics 'light and heavy household work' (47% respectively 45%), and to a far lesser degree (16%), with the topic 'walking'. The questions about these topics are difficult to answer and it is quite easy to decompose the questions about these topics into questions about separate components, i.e., separate household tasks and separate destinations for walking.

We must bear in mind that the large differences between the six topics with respect to the control activities of the interviewers cannot be the result of differences between characteristics of interviewers, or of differences between respondents, because every interviewer and every respondent asked and answered questions about all six topics. Neither can the differences between the topics be the result of a different question format, as the questions for all six topics had about the same format; only the activity itself (X) differed (see the Appendix for the exact question wording).

11. The Interviewer and the Type of the Sequence

The exercise of partial questioning is certainly not scripted; it is based on the initiative of individual interviewers who think that decomposing the questions will lead to better results. The expectation therefore is that the degree to which partial questioning is used will strongly differ between interviewers. The data of Table 4 give strong support to that expectation.

Interviewer #2 used partial questioning in 35 percent of 'her' macrosequences. The three other interviewers did not differ very much with respect to the type of control used in their interviews.

Table 4. Distribution of all macrosequences over Type of Control, by Interviewer Number

Interviewer Number	Type A	Type B	Type C	Total	N
#1	50%	35%	15%	100%	420
#2	41%	25%	35%	100%	360
#4	59%	27%	14%	100%	156
#8	48%	37%	15%	100%	462
All interviewers	48%	32%	20%	100%	1398

Chi-square = 74.6; df = 6; p<.001**

The difference in control behavior between interviewer #2 and her three other colleagues, especially appeared in the macrosequences about the topics 'walking' and 'cycling'. Here interviewer #2 used the partial questioning strategy (Type C) in 42% respectively 22% of the interviews, that is six to ten times as much as the other interviewers (p<.001**). For the four other topics the differences between the interviewers with respect to control behavior were not significant [7].

12. The Respondent and the Type of the Sequence

In Figure 2 we depicted our hypothesis that the control behavior of interviewers would be affected not only by the difficulty of the question to be answered, but also by characteristics of the respondent. We assume that important background characteristics like age, gender and cognitive and physical functioning are related with the degree to which interviewers add their own closed loop control to the open loop control as designed by the researcher. Stated otherwise: we hypothesize a relationship

between the distribution of the macrosequences over the three types of control behavior (A, B, and C), and relevant respondent characteristics.

We choose the following respondent characteristics:

1. *Gender* of the respondent: male *versus* female;
2. *Age* of the respondent: 65-75 *versus* 76 and over;
3. Score on the Mini Mental State Examination *(MMSE),* a test for the evaluation of the cognitive function of the subject, ranging from 0 to 30 points (Folstein, Folstein and Hugh, 1975): below 24 *versus* 24 and above;
4. Score on a Disability scale indicating the difficulty subjects have with doing activities of daily life (Van Gool et al. 2005): score 0-2 *versus* score 3-6 [8].

For this analysis we have selected from the 1398 macrosequences the 931 complete macrosequences, i.e., macrosequences in which not only the base question was asked, but also the frequency question and the duration question. The reason for this selection is that incomplete sequences will almost by definition belong to type A (short sequences). And it is very likely that there is a relationship between the probability of performing an activity by a respondent and some of the respondent characteristics like Age and Disability.

From the data of Table 5 it appears that none of four respondent characteristics chosen have a significant effect on the type of control exercised by the interviewer. So this hypothesis has to be rejected.

Table 5. Distribution of the complete macrosequences over the three Types of Control for respondent characteristics Gender, Age, Cognitive Functioning (MMSE) and Disability.

Respondent characteristic		Type A Open loop control only	Type B Open loop control and repair	Type C Open loop control, partial questioning and possibly repair	Total	N
Gender	Male	33%	39%	28%	100%	435
	Female	27%	42%	31%	100%	496
	Total	30%	40%	30%	100%	931
	Chi-2	4.1	df = 2	p =.13 ns		
Age	65 -75	30%	42%	28%	100%	491
	>76	29%	39%	32%	100%	440
	Total	30%	40%	30%	100%	931
	Chi-2	1.5	df = 2	p =.48 ns		
MMSE	<24	24%	45%	31%	100%	62
	>23	30%	40%	30%	100%	869
	Total	30%	40%	30%	100%	931
	Chi-2	1.1	df = 2	p =.57 ns		
Disability	0,1,2	31%	41%	28%	100%	720
	3-6	28%	37%	35%	100%	211
	Total	30%	40%	30%	100%	931
	Chi-2	3.6	df = 2	p =.16 ns		

13. Summary

In the research project described in this paper, elderly respondents (65+) were asked detailed retrospective questions about 'mundane' physical activities like walking, cycling and household tasks. On the basis of methodological literature concerning the difficulty of retrospective questions (especially those with low saliency), and the more frequently occurring inadequate response behavior of elderly respondents, one would expect many 'no answers' and 'don't know answers.'

However, contrary to that expectation, the proportion of such inadequate answers appeared to be quite low, indicating that the interviewers must have done their very best to obtain adequate answers. For that purpose they, in principle, had three different methods at their disposal. Firstly, an optimal execution of the prescribed 'open loop control', that is, precisely following the text of the questionnaires and the instructions given by the researcher. Secondly, via the 'feedback loop control' of repair: if nevertheless inadequate answers are given, to perform further probing until eventually an adequate answer is obtained. And thirdly via 'feedforward loop control'; if the interviewer expects problems with posing this particular question to that particular respondent, she may adjust the question, for example by decomposing a general question into its components, making the question probably easier to answer (partial questioning).

To investigate to what degree feedback loop and feed forward loop controls are actually used by interviewers, *the classical design of methods research of interviewer effects, i.e., the split ballot design, does not fit* because this experimental design can only deal with the unidirectional causal relation between interviewer behavior and resulting answers of respondents, but not with interviewer behavior that is the *result* of the expected or observed initial answers of the respondents. For our research purpose we therefore had to use another research design, i.e., a detailed analysis of the interactions between interviewer and respondent preceding the final response of the respondent.

From the analysis of question-answer sequences it appeared that *the type of control* exercised in these sequences, and as a consequence, the length of these sequences, *differs strongly by topic of the question*. With 'gardening' the prescribed open loop control was sufficient, but the topic 'walking' required a lot of repair, whereas the topics of 'light and heavy household tasks' initiated much feedforward loop control in the form of partial questioning.

The *degree to which interviewers vary* with respect to the type of control they use *differs strongly by topic*. For 'walking' and 'cycling' there was a large difference between one interviewer and the other three; but this difference was hardly present for the other four topics.

Differences regarding the behavior of interviewers and respondents between the six topics *cannot be explained by differences in question wording*, because for all the six topics the same question format is used. The only difference between the questions is the topic (X) mentioned in the question. *Neither can differences regarding interviewer*

characteristics explain the differences between the six topics as each interviewer posed questions about all topics. *Nor can the differences* between the six topics *be explained by respondent characteristics* as the respondents were about randomly assigned to the interviewers. Moreover *respondent characteristics* like age, gender, physical functioning (Disability-scale) and cognitive functioning (MMSE) *are hardly related to the control behavior of the interviewers.*

So we can conclude that more difficult topics not only lead to more need for repair, but also to larger differences between the interviewers concerning their readiness to use feedforward loop control, in the form of partial questioning, in addition to, or as an alternative for, the open loop control as designed by the researcher.

14. Epilogue

One may wonder whether this conclusion about the effect of the question topic on the control behavior of the interviewers, could also have been reached if we had *not* performed a (rather time consuming) sequence analysis, but only a 'classical' study of interviewer effects. That is, by relating the interviewer number to the scores on the main dependent variable of this study, i.e., the time spent during the last two weeks to each of the six activities mentioned in the LAPAQ-questionnaire.

Table 6. Interviewer effects per topic; for complete macrosequences only

Topic	F	Degrees of freedom	p	Number of macrosequences
Walking	7.4	3; 198	.001**	202
Cycling	9.5	3; 103	.001**	106
Gardening	0.6	3; 69	.592 ns	73
Sport	2.9	3; 119	.036*	123
Light HW	4.1	3; 221	.007**	225
Heavy HW	5.4	3: 166	.001**	170

To answer this question we performed a one way ANOVA analysis with time spent to each activity in the last two weeks as dependent variable, and interviewer number as fixed factor. The resulting Table 6 shows significant interviewer effects for five of the six topics. *But we cannot infer from this table that the differences in control behavior among the interviewers are the most likely causes of these interviewer effects*. So the sequence analysis not only pointed at the existence of interviewer effects, but it also identified repair behavior of the interviewers and partial questioning as the main sources of these interviewer effects.

We end this paper with some remarks regarding the method used in this research project, i.e., sequence analysis.

1. With modern software like SEQUENCE VIEWER, sequence analysis is far less time consuming than it was in the past. It requires that the interviews have been (audio) recorded; the transcription, the coding and especially the analysis can be partly automated.
2. The content analysis of the sequences may be helpful to detect interviewer behavior that deviates from that what is expected by the researcher (e.g.,

suggestive questioning, and partial questioning), making the responses collected by different interviewers less comparable. The content analysis thus may lead to an improved interviewer training and monitoring.

3. The content analysis may also point at ill-designed questions. A clear example is all the confusion about what exactly is meant by the researcher by the term 'walking': is walking through the corridors of a large home for the elderly 'walking' or not? The detection of these semantic problems may lead to an improved question wording.

Acknowledgements

This paper is based on data collected in the context of the Longitudinal Aging Study Amsterdam (LASA). This program is conducted at the Vrije Universiteit in Amsterdam and is supported by the Ministry of Health, Welfare, and Sports.

Notes

[1] Questions from standardized questionnaires differ strongly with respect to the difficulty respondents will have to answer them correctly. One instrument to measure the question difficulty is the Task Difficulty Score, TDS (Van der Zouwen and Smit, 2004); another the Question Understanding Aid, QUAI (Graesser et al. 2006). The higher the score, the more effort respondents have to do to answer the question and the higher the need they will have to get help from an interviewer.
[2] For the sake of readability, interviewers are indicated in this paper by she and her, whereas researchers and respondents are indicated with he and him.
[3] Of the 524 eligible subjects, 85 did not participate (14 deceased, 34 refused, 31 were not capable, and 6 could not be contacted). Of the 439 participants in the study, there were four non-participants of the LAPAQ: one was bedridden and three used an electrical wheelchair for transportation.
[4] Dependent on the topic some additional questions are asked. For example, for the topic 'gardening' a question reads "Did you turn over the soil in your garden?"
[5] Each sequence is identified by the number of the interview followed by the number of the topic. For example, macrosequence #11622-1 is the sequence about 'walking' in interview 11622.
[6] This does not automatically mean that these macrosequences are all paradigmatic: much unwanted suggestive questioning can have taken place in these interviews, but that remains unnoticed here because suggestive questioning generally does not lead to additional turns.
[7] To avoid artificial results, the analysis of the interaction by the Topic on the relationship between Interviewer Number and Type of Control exercised, is done for the 931 complete macrosequences only.
[8] The Disability scale used here is a Dutch variant of the ADL-scale of Katz (1983)

REFERENCES

Deeg, D.J.H., and Westendorp – de Serière, M. (Eds.) (1994), *Autonomy and well-being in the aging population. Report from the Longitudinal Aging Study* Amsterdam 1992-1993. Amsterdam: VU University Press.

Dijkstra, W. (1999), A New Method for Studying Verbal Interactions in Survey-Interviews. *Journal of Official Statistics 15*, 67-85

Dijkstra, W. (2002), Transcribing, coding and analyzing verbal interactions in survey interviews. In: D.W. Maynard, H. Houtkoop, N.C. Schaeffer and J. van der Zouwen (Eds.), *Standardization and Tacit Knowledge, Interaction and Practice in the Survey Interview.* New York: Wiley, 401-425.

Dijkstra, W. (2005), Question-Answer Sequence. In: S.J. Best and B. Radcliff (Eds.), *Polling America; An Encyclopedia of Public Opinion*. Greenwood Press, Estport CT and London, pp. 612-616

Folstein, M.F., Folstein, S.E. and McHugh, P.R. (1975) Mini-mental state. A practical method for grading the cognitive state of patients for a clinician. Journal of Psychiatric Research, 12, 189-198.

Graafmans, W.C., Bouter, L.M. and Lips, P. (1998) The influence of physical activity and fractures on ultrasound parameters in elderly people. *Osteoporosis International, 8*, 449-454.

Graesser, A.C., Cai, Z., Louwerse, M.M., and Daniel, F. (2006), Question Understanding Aid (QUAID); A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly*, 70(1) 3-22.

Groves, R.M. (1989), *Survey Errors and Survey Costs*. Wiley & Sons, New York.

Houtkoop-Steenstra, H. (1995), Conversational Problems in Telephone Survey Interviews. In: *Proceedings of the International Conference on Survey Measurement and Process Quality.* Bristol. American Statistical Association, Alexandra VA, pp. 286-289.

Hyman, H.H., Cobb, W.J., Feldman, J.J., Hart, C.W. and Stember, C.H. (1954/1975), Interviewing in Social Research. The University of Chicago Press, Chicago.

Katz, S. (1983), Assessing Self-Maintenance: Activities of Daily Living, Mobility and Instrumental Activities of Daily Living. *Journal of the American Geriatrics Society*, 31(12), 721-726.

Knaüper, B., Belli, R.F., Hill, D.H., and Herzog, A.R. (1997), Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. *Journal of Official Statistics, 13,* 181-199.

Knaüper, B., Schwartz, N. and Park, D. (2004), Frequency Reports Across Age Groups. *Journal of Official Statistics*, 20(1), 91-96.

De Leeuw, E.D.(1992), *Data Quality in Mail, Telephone and Face to Face Surveys.* Dissertation Vrije Universiteit Amsterdam TT-Publikaties Amsterdam.

De Leeuw, E.D. and J. van der Zouwen (1988), Data quality in telephone and face to face surveys: a comparative meta-analysis. In: R.M. Groves et al. (Eds.), *Telephone Survey Methodology*. Wiley, New York, pp. 283-299.

Moore, R.J. and Maynard, D.W. (2002), Achieving Understanding in the Standardized Survey Interview: Repair Sequences. In: D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. van der Zouwen (eds.), *Standardization and Tacit Knowledge; Interaction and Practice in the Survey Interview*. Wiley, New York, 281-312.

Schaeffer, N.C. and Maynard, D.W. (1996), From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews. In: N. Schwarz and S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* Jossey-Bass, Francisco.

Schaeffer, N.C. and Maynard, D.W. (2002), Standardization and Interaction in The Survey Interview. In: J.F. Gubrium and J.A. Holstein (Eds.), Handbook of Interview Research; Context & Method. Sage, Thousand Oaks, pp. 577-601.

Schuman, H. and S. Presser (1996), *Questions and Answers in Attitude Surveys; Experiments on Question Form, Wording, and Context*. (2nd edition) Sage, Thousand Oaks.

Schwarz, N., Park, D. Knaüper, B., and Sudman, S. (Eds.)(1998), Cognition, Aging, and Self-Reports. Psychology Press, Philadelphia, PA.

Stel, V.S. (2003), *Prevention of fall accidents in older persons. From risk profile to intervention strategy.* Unpublished dissertation Vrije Universiteit Amsterdam.

Stel, V., Smit, J.H., Pluijm, S.M.F., Visser, M., Deeg, D.J.H., and Lips, P. (2004), Comparison of the LASA Physical Acitivity Questionnaire with a 7-day diary and pedometer. *Journal of Clinical Epidemiology 57*, 252-258

Van der Zouwen, J. (2006) Interviewer and Survey Researcher: Mutual Dependencies. *Bulletin de Méthodologie Sociologique*, No. 89, 49-64.

Van der Zouwen, J. and J.H. Smit (2004), Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question-Answer Sequences: A Diagnostic Approach. In: S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer (Eds.), *Methods for testing and Evaluating Survey Questionnaires*. John Wiley & Sons, New York, pp. 109-130.

Van der Zouwen, J. and J.H. Smit (2005), Control processes in survey interviews: a cybernetic approach. *Kybernetes, The international journal of systems & cybernetics*, 34 (5/6), 602-616

Van der Zouwen, J. and J.H. Smit (2006), The effectiveness of repair strategies of interviewers in survey interviews. In: C. van Dijkum, J. Blasius & C. Durand (Eds.), *Recent Developments and Applications in Social Research Methodology*. Barbara Budrich Publishers, Opladen CD-ROM.

Van der Zouwen, Johannes, Johannes H. Smit & Marleen H.L. van der Horst (2005), Reporting the Frequency and Duration of Household Tasks by Elderly Respondents: The Effect of Different Interview Strategies on Data Quality. Paper presented at the 60th AAPOR Annual Conference, May 12-15, 2005, Miami Beach, FL.

Van der Zouwen, J., Smit, J.H., and van der Horst, M.H.L. (2005), Expectations of Interviewers as Determinant of Deviant Interviewer Behavior. Paper presented at the First Conference of the European Association for Survey Research, Barcelona, 18-22 July, 2005.

Van der Zouwen, J. and C. van Dijkum (2001), Towards a Methodology for the Empirical Testing of Complex Social Cybernetic Models. In: F. Geyer and J. van der Zouwen (Eds.), *Sociocybernetics: Complexity, Autopoiesis, and Observation of Social Systems. Greenwood Press, Westport CT and London, pp. 223-240.*

Van Dijkum, C. (2006), Causality in the Social Sciences Revised: a Short History and Outlook to the Future. Paper presented at the 16th ISA World Congress of Sociology, Durban South Africa, 23-29 July, 2006.

Van Gool, C.H., Kempen, G.I.J.M., Penninx, B.W.J.H., Deeg, D.J.H., Beekman, A.T.F., and van Eijk, J.Th. M. (2005), Impact of depression on disablement in late middle aged and older persons: results from the Longitudinal Aging Study Amsterdam. *Social Science & Medicine* 60, 25-36.

APPENDIX

Selected questions from the LASA Physical Activity Questionnaire (LAPAQ) (Stel, 2003)

q1. Do you **walk** outside? With walking outside we mean walking to go shopping or doing other daily activities, like visiting someone. We do not mean: a walking tour. 1. no (go to question q5); 2. yes.

q2. Did you walk during the past two weeks? 1. no (go to question q5); 2. yes.

q3. How many times did you walk during the past two weeks? …. times.

q4. How long did you usually walk each time? …hours; … minutes.

q5. Do you **cycle**? With cycling we mean cycling to go shopping or doing other daily activities, like visiting someone. With cycling we do not mean: a cycling tour. 1. no (go to question q9); 2. yes.

q6. Did you cycle during the last two weeks? 1. no (go to question q9); 2. yes.

q7. How many times did you cycle the past two weeks? …. times.

q8. How long did you usually cycle each time? … hours; … minutes.

q9 Do you have a **garden** (including allotment)? 1. no (go to question q15); 2. yes.

q11. Did you work in the garden during the past two weeks? 1. no (go to question q15; 2. yes.

q12. How many times did you work in the garden during the past two weeks? … times

q13 How long did you usually work in your garden each time? … hours; ….minutes

q15 Do you do **sports**?

q17 How many times did you this sport during the past two weeks? … times

q18. How long did you usually do this sport each time? …hours; …minutes

q24 Do you do **light household tasks**? With light household tasks we mean washing the dishes, dusting, making the bed, doing the laundry, hanging out the laundry, ironing, tidying up, and cooking meals. 1. no (go to question q27); 2. yes.

q25. How many days did you do light household tasks during the past two weeks? … days

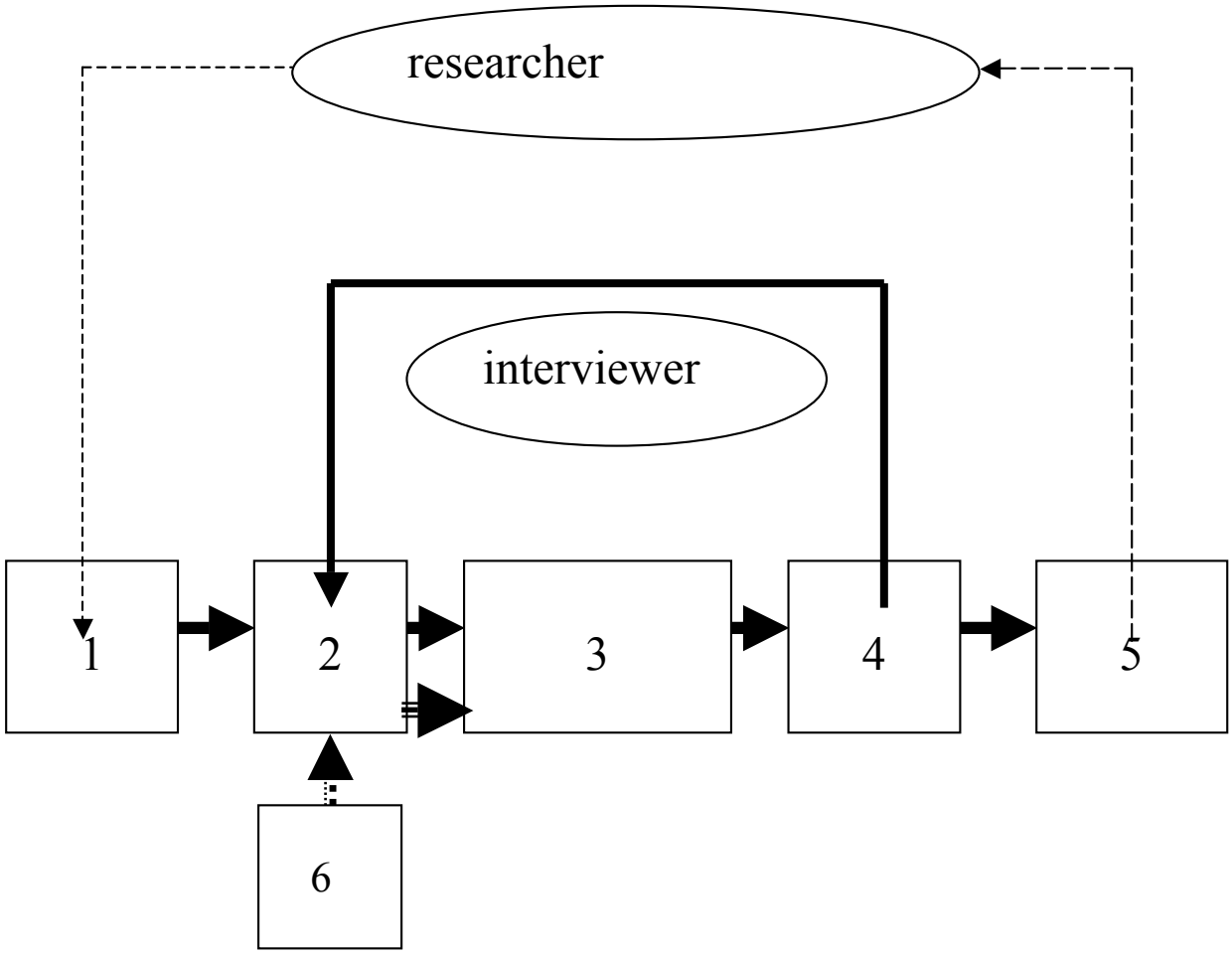
q26. How long per day did you usually do light household tasks? … hours; … minutes

q27. Do you do **heavy household tasks**? With heavy household tasks we mean window cleaning, changing the bed, beating the mat, vacuuming, washing or scrubbing the floor, and chores with sawing, carpeting, repairing or painting. 1. no (go to question q30); 2. yes

q28 How many days did you do heavy household tasks during the past two weeks? … days

q29 How long per day did you usually do heavy household tasks? … hours; …minutes

Figure 1. The interview as controlled communication process



1. The question from the questionnaire as constructed by the researcher and the instructions for the interviewers
2. The (probing) question as posed by the interviewer
3. The cognitive process within the respondent
4. The answer given by the respondent
5. The response code assigned by the interviewer
6. Expectations of the interviewer about the respondent

Figure 2. The relationship between question difficulty, characteristics of the respondent, control behavior of the interviewer, and length of the question answer sequence