Preventing mismatch answers in standardized survey interviews

Yfke P. Ongena · Wil Dijkstra

Abstract Interaction analysis of question–answer sequences from a telephone survey shows that so-called mismatch answers, i.e. answers that do not correspond to the required answering format, are the most frequently occurring problematic verbal behavior. They also are likely to trigger suggestive interviewer probing. Explanations for the occurrence of mismatch answers concern cognitive and conversational factors. In both a non-experimental and an experimental study it was found that questions with formal response alternatives yield more mismatch answers than questions with colloquial response alternatives (i.e., words that are frequently used in ordinary conversations). Effects of the wording of questions were only found in the non-experimental study, indicating that formal question wording yields fewer mismatch answers than colloquial question wording. The findings suggest that, especially in case of questions that are formulated as agree–disagree opinion assertions, the chance of mismatch answers is highly reduced when colloquial response alternatives are used.

Keywords Behavior coding · Question wording · Interaction analysis · Question–answer sequence · Response error · Survey-interview

1 Introduction

The study of verbal interactions in the interview is one of several methods to obtain information about the quality of responses. In interaction analysis the 'paradigmatic sequence'

Y. P. Ongena (✉)
Department of Political Science and Research Methods, School of Management and Governance, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: Y.P.Ongena@utwente.nl

W. Dijkstra
Department of Social Research Methodology, Vrije Universiteit, Amsterdam, The Netherlands

(Schaeffer and Maynard 1996) is a crucial concept. In a paradigmatic question–answer sequence (henceforth: Q–A sequence) the question is read as scripted, the answer of the respondent is directly scorable and the interviewer may acknowledge the answer neutrally. The following Q–A sequence clearly deviates from this pattern:

Excerpt 1 Q–A sequence with an example of a mismatch answer[a]

---

1.   I: Do you strongly agree, agree, disagree or strongly disagree with the statement "For
     me, television commercials are too much alike"?
2.   R: Well yes,
3.   R: They are all the same

---

[a] This Q–A sequence was translated into English, taken from a Dutch Television Survey (Smit and Neyens 2000)

The required answering format is one of four categories as indicated by the interviewer. In line 2 the respondent provides an answer ("Well, yes") that is not formatted according to these response alternatives, i.e. the respondent gives a so-called mismatch answer (Dijkstra 1999). Because such a deviation from the paradigmatic sequence requires the interviewer to take action in order to obtain a scorable response, such deviations are called problematic. It was shown that response validity was lower for sequences with problematic deviations than for non-problematic and paradigmatic sequences (Dijkstra and Ongena 2006). Mismatch answers appear to be the most frequently occurring problematic deviations from the paradigmatic Q–A sequence (Ongena 2001). Mismatch answers also are the most important cause of *interviewer* problematic deviations. As Smit (1995) concluded, interviewers often try to repair respondent's inadequate answers, and suggestive probing seems to be the most 'effective' way. The respondent's mismatch answer in Excerpt 1 would for example be followed by a suggestive probe of the interviewer like 'Would that be strongly agree?'. Such suggestions are likely to be immediately accepted by the respondents (Smit et al. 1997).

To prevent respondents from giving mismatch answers, or stimulate them to give more precisely formatted answers we should know more about the causes of such mismatch answers. In this paper we will formulate some hypotheses about such causes and test the hypotheses in a non-experimental and an experimental study.

2 Causes of mismatch answers

According to Houtkoop-Steenstra (2000), the occurrence of mismatch answers (or, as she labels them, 'unformatted' answers) is remarkable for two reasons. First, "From a cognitive perspective, one would expect that it is much easier to merely repeat a line that was just presented by the interviewer than it is to formulate a different answer" (p. 183). And from a social perspective, "one might expect respondents to be willing to please the interviewer and to make her task as easy as possible" (p. 183). Nevertheless, mismatch answers do occur frequently, and Houtkoop-Steenstra argues that respondents may have a good reason not to provide formatted answers, but she does not discuss what such a reason might possibly be. To account for the occurrence of mismatch answers, Dijkstra and Ongena (2006) also distinguished between cognitive and social factors.

According to the first factor mismatch answers may be an indication of problems in cognitively processing the question. Especially retrieval of information may be related to the occurrence of so-called *cognitive* mismatch answers. When respondents retrieve information, they need to process a lot of information, for example enumeration of instances of frequent behavior during a relatively long period. Respondents may express this retrieval of

information through verbal considerations. While providing verbal considerations, they are more likely to forget the required alternatives, and formulate an answer in their own words, thus increasing the probability of a mismatch answer.

The studies described in this paper concern the second factor, which relates to a more motivational problem. Respondents may view the survey interview as an ordinary conversation, and as a consequence they may be convinced that colloquial answers are adequate. Although Schuman and Presser (1981) argue that most respondents "accept the framework of questions and try earnestly to work within that framework" (p. 299), the occurrence of mismatch answers shows they do not always follow "the rules of the game". This factor is also related to the principle of satisficing (Krosnick 1999); respondents may not be very motivated to give an optimal answer, not only with regard to accuracy, but also with regard to exact formulation We assume that mismatch answers especially occur when respondents are distracted from their task of providing exactly formatted answers, either by cognitive problems or by social factors. Both respondent and question characteristics may be involved in causing this distraction. For respondent characteristics, age may be relevant (older respondents might have more difficulties in focusing on the task than younger ones). For question characteristics, the question wording may play a role. Dijkstra and Ongena (2006) showed that especially assertions (with a Likert-type scale ranging from strongly disagree to strongly agree) received a lot of mismatch answers, consisting of just 'yes' or 'no'. Such responses are perfectly normal in ordinary conversations. Assertions, such as "For me, television commercials are too much alike", are rather conventional expressions. Therefore, a colloquial style of responding (i.e., providing yes–no answers instead of agree–disagree answers) may be triggered when questions resemble expressions that are common in conversations.

In Excerpt 1 we gave an example of a mismatch answer. The problem does not seem a cognitive one, because the respondent does not indicate problems in retrieving an answer. The respondent just gives an answer that is perfectly normal in ordinary conversations, and in addition elaborates on his answer. Elaborations are not useful in survey interviews (unless they indicate problems in misunderstanding of questions that can subsequently be solved), but very normal in ordinary conversations. When respondents view the interview as an ordinary conversation they may react in a conversational (or colloquial) way, instead of focusing on the task of giving precisely formatted answers. Therefore, we classify the mismatch answer as a colloquial mismatch.

3 Hypotheses

3.1 Colloquial and formal questions

Cannell et al. (1977) showed that although respondents may be sufficiently motivated to participate in interviews, they generally do not know what is expected of them. Thus, according to Cannell et al., the questionnaire and the techniques interviewers use, should clarify the respondents' general task. These should provide cues as to how respondents can answer questions most efficiently, and motivate them to meet the requirements of accurate responses. The specific instructions they used in their experiment yielded more precise and elaborate reports. These instructions clarified the goal of the survey and provided cues that clarified intended task performance (e.g., "we'd like you to be as exact as you can").

Although Cannel et al.'s suggested instructions were merely concerned with general accuracy of reports, and not specifically with properly formatted answers, we think they could be applied with the aim to decrease the number of mismatch answers as well. However, including

such instructions will increase interview time. Moreover, specific instructions are not likely to be effective when they are not repeated throughout the interview, but when repeated too often they may offend respondents. We argue that the wording of questions can be used to serve the same goal of cueing respondents to give precisely formatted answers. A question itself may signal respondents about the character of the survey. We assume that colloquially worded questions, in which the wording of ordinary conversations is used, may give false signals to the respondent about the required degree of accuracy in reporting, whereas formally formulated questions alert the respondent to the formal character of the survey, reminding them to answer with precisely formatted answers. Therefore, we expect that a question that is formulated as a colloquial question will generate more mismatch answers, than a formally worded question (hypothesis 1).

3.2 Colloquial and formal response alternatives

In addition to question wording, the wording of response alternatives may also affect the likelihood of mismatch answers. For respondents it is much easier to produce an answer that is a normal expression in conversations, than to use uncommon words as formulated by the researcher. For example, it is much more common for respondents to answer a question with 'yes' or 'no', than to use the words 'agree' or 'disagree'. In addition, when alternatives are used for several questions in a row (as is often the case for batteries of assertions), alternatives are sometimes presented only at the introduction of the battery. Even though interviewers may be required to repeat the alternatives for each assertion, it is not known to what extent they actually do this. Thus, the alternatives may have to be stored in short-term memory. It is likely that when colloquial words are used, such words are more easily remembered than formal, more unusual words and allow for a natural or even automatic way of responding. Therefore, we expect that questions with colloquial response alternatives will generate fewer mismatch answers, than the same questions with comparable formal response alternatives (hypothesis 2).

4 Study 1: a non-experimental study

4.1 Method

A non-experimental study was conducted to investigate whether the different types of questions and alternatives (formal versus colloquial) could be identified in a normal survey, and were related to the occurrence of mismatch answers. For these analyses Q–A sequences from interviews of the Dutch pilot of the European Social Survey (ESS) were used. These concern face-to-face interviews that were conducted by means of a CAPI program, in the spring of 2002 (ESS 2007). The questionnaire consisted of 268 different questions. This large number enabled us to find examples of all types of questions and alternatives. Tapes of 23 interviews with a good recording quality, available from seven different interviewers, were digitized. Due to time constraints, not all 268 questions of all interviews were transcribed. Eight interviews were transcribed completely, for the other 15 interviews at least the first 100 questions of the interview were transcribed.

Coding of the data was done by two different coders. A third coder coded random selection of 30% of the 3623 Q–A sequences. Comparison of the 1100 twice-coded Q–A sequences yielded a Kappa of 0.70.

Table 1 Frequency of mismatch answers for four question types

Q–A sequences	Colloquial Questions		Formal Questions			
	Formal alternatives (A)		Formal alternatives (B)		Colloquial alternatives (C)	
With mismatch	32	26%	63	17%	15	2%
Without mismatch	90	74%	299	83%	556	97%
Total	122	100%	362	100%	571	100%

4.2 Distinction of colloquial and formal questions

In order to non-experimentally test the hypotheses, we distinguished versions of each question type (colloquial and formal) among the ESS questions. Not all questions were appropriate to test the hypotheses. We did not analyze questions without an explicit list of alternatives, such as open-ended questions (e.g., type of job of the respondent), and open-ended questions with numbers as response options, such as "How many days a week do you watch television?" (7 and 19 questions, respectively). In addition, we did not consider the three questions that were hardly ever asked since they were observed or inferred by the interviewer (i.e., the gender, country of birth and nationality of the respondent). The use of show cards was unevenly distributed among the ESS questions. Furthermore, these questions were expected to complicate interpretation of results since use of show cards can decrease the occurrence of mismatch answers (e.g., Prüfer and Rexroth 1985). Therefore questions with show cards (156 questions) were also excluded from the analysis. Eventually 83 different questions (concerning 1055 Q–A sequences) were included in the analysis.

The distinctions between colloquial and formal questions were made on the basis of the colloquial character of the question. When question wording was considered to include common words and a sentence structure that is generally normal to use in ordinary conversations, the question was categorized as colloquial. All other questions were considered formal. An assertion such as 'Politicians do not care what people like me think' was considered to consist of common words and to be a normal expression in ordinary conversations. A question like 'Do you consider yourself as a member of a minority group that is discriminated in this country?' is not likely to be formulated as such in ordinary conversations, and was categorized as formal. Of course, there are questions that can be considered far more colloquial than other questions within the same category, but we chose to dichotomize the questions into the two categories that the hypotheses refer to, and not to complicate the categorization with gradations of the colloquial character of the questions.

In the same way we categorized the types of alternatives; alternatives that consisted of common words and a simple structure (e.g., 'yes' and 'no') were considered colloquial alternatives, all other alternatives were considered formal.

Panel A in the appendix shows some examples of questions categorized for their colloquial character and types of alternatives. For colloquial questions all accompanying alternatives happened to be categorized as formal. Hence, we could not distinguish between formal and colloquial alternatives for colloquially worded questions.

4.3 Results of study 1

Table 1 shows the frequency and percentages of mismatch answers and non-mismatch answers for the available combinations of question types, and alternatives.

Hypothesis 1 concerned the effect of question wording on the number of mismatch answers. Comparing column A and B shows that colloquial questions yield more mismatch answers (26%) than formal questions (17%, $\chi^2 = 3.04$, df = 1, $p < 0.05$). Thus hypothesis 1 could be confirmed.

Also hypothesis 2 could be confirmed. Comparing column B and C shows that questions with formal alternatives yield more mismatch answers (17%) than questions with colloquial alternatives (2%, $\chi^2 = 63.28$, $p < 0.01$).

We have to keep in mind that the questions were not designed in a proper split ballot experiment. The questions do not only differ with respect to colloquial and formal wording of questions and alternatives, but also with respect to the topics being asked about, and the order of occurrence in the questionnaire. For example, all background questions were categorized as formal questions (e.g., respondents' age, number of persons in the respondents' household, etc.). The topic of the question may very well affect the percentage of mismatch answers.

Also, perception, factual and opinion questions were very unequally distributed over colloquial and formal questions and alternatives. It was not possible to test the hypotheses controlling for perceptual, factual and opinion questions, because combinations of all these kinds of different questions with colloquial and formal questions yielded empty cells.

4.4 Conclusions non-experimental study

The results indicated confirmation of our expectations. Colloquial questions yielded more mismatch answers than formal questions, and formal alternatives yielded more mismatch answers than colloquial alternatives. However, some of these variables are confounded with other characteristics of questions. This makes interpretation of results rather difficult. For example, assertions occurred only as questions with formal alternatives in the ESS questionnaire.

Although a large number of different questions, and as a result a relatively large number of Q–A sequences was available, data from only 23 different respondents and 7 interviewers were analyzed. Because of these low numbers, the results cannot easily be generalized to other surveys. Moreover, distinctions of the colloquial character of questions and types of alternatives were based upon the researcher's assumptions about common words used in ordinary conversations. A better strategy would be to use empirical data concerning words and formulations used in ordinary conversations of the intended population, in order to formulate new colloquial and formal questions and alternatives. In an experimental design, the same questions can be manipulated, and confounding of variables can be avoided.

Another problem of the ESS-data was that they concerned face-to-face interviews. In face-to-face interviews, non-verbal communication can play an important role in the interaction. For example, specific task-related behaviors, such as acknowledgements or confirming responses like a nod or specific gesture may occur, without auditory communication. This non-verbal visual communication between interviewer and respondent is not available to the researcher, when only auditory information is used. In case of telephone interviews, there is no visual communication and as a result audio-recordings are sufficient to perform interaction analysis (Ongena and Dijkstra 2006).

Finally, telephone interviewing enables conducting a large number of interviews in a relatively short period of time. For these reasons, it was decided to test the hypotheses in an experimental design, varying the formulation of questions and alternatives and to conduct the interviews by telephone.

5 Study 2: an experimental study

5.1 Manipulation of colloquial and formal questions

Four different questionnaires were designed for CATI interviews in which colloquial and formal questions were asked, accompanied by either colloquial or formal response alternatives.

Our definition of 'colloquial questions' primarily concerns the word choice in ordinary conversations. Words that are used in surveys may differ from those typically used in ordinary conversations. For example, in survey questions, 'research-theoretical concepts' may be used that would be awkward to use in ordinary conversations, such as 'main activity' in a common survey question like 'What is currently your main activity: employed, unemployed, retired, or in education?' Adhering to 'lexical availability' and 'frequency of use' factors (Brennan and Clark 1996), we used words that are used most frequently in ordinary conversation, but will have more or less the same meaning as their formal equivalents to formulate colloquial questions. Formal questions then contain words that are less frequently used in ordinary conversations.

The difference in word use especially concerns the *nouns*, *verbs, adjectives* and *adverbs* used. Although other word types (i.e., articles, conjunctions, pronouns, prepositions) and the grammatical structure used may differ between conversations and survey interviews, we did not address this in our manipulations.

To determine what words are used most frequently we used information from the Spoken (Dutch) Corpus project. This project aims to yield a resource of 1,000 h of speech (approximately ten million words) originating from adult speakers of standard Dutch (Oostdijk et al. 2002). Within the corpus, two different sub-corpuses are relevant as indicators for ordinary conversations, i.e., the 'spontaneous conversations' and 'telephone conversations'. These sub-corpuses contain 1,733,244 and 593,980 words, respectively.

To maximize the external validity of the study, we selected, as far as possible, existing questions from actual surveys on health and health related topics. In this paper, only the attitude and perception statements that we included in the questionnaire will be discussed. We used eight questions from the Dutch translations of the short version (SF-20) of the general health perception questionnaire (henceforth: GHPQ, Kempen et al. 1995; Kriegsman et al. 1995), and six opinion questions on costs of public health (Elchardus et al. 2002; Van de Berg et al. 1986). Furthermore, three opinion assertions were included concerning government and health issues (RVD 2003). Based upon the Corpus frequency of the nouns, verbs, adjectives and adverbs in the questions, we determined whether they could be considered colloquial or formal. The criterion for a question being colloquial rather than formal was that the Corpus frequency of the words used was above the average of the telephone conversations. Subsequently, with replacement of words (i.e., colloquial words with formal words or vice versa), we constructed the opposite version. We took care that the Corpus frequency of the relevant words in the colloquial version was at least five times as large as in its formal counterpart. The questions used in the three batteries of assertions in the experiment are explained in more detail in panel B of the appendix.

5.2 Data collection

A sample of telephone numbers was drawn from a website with telephone listings of all local communities in The Netherlands. In February and March 2004, 1525 different telephone numbers were dialed. Eventually 40% of these calls resulted in a completed interview (including both eligible and non-eligible cases). The largest number of non-response calls

are refusals, as could be expected. In 34.4% of the calls respondents could not be persuaded to engage into an interview. The AAPOR 'Response Rate 1' (or minimum response rate) is 44.5%.[1] There was no significant difference between the four ballots for the respondent background variables such as age, education and gender.

The 12 interviewers were all female social-science students aged between 19 and 28. They received 15 h of training within 1 week before their first interview evening. Each interviewer had to conduct interviews with all four versions of the questionnaire. The interviewers were told that different question wordings were used to control for question wording effects on response distributions, and for this reason it was important that they read the questions exactly as worded.

During the fieldwork, the interviewers were monitored using digital recordings. If necessary, interviewers were instructed to improve their behavior. This was especially necessary for some interviewers with regard to persuading reluctant respondents, neutral probing, and adequate clarification of questions. Reading the questions exactly as worded appeared not to be a problem.

The number of completed interviews (610) comprised a dataset of 25,670 Q–A sequences. In this paper only 10,130 Q–A sequences concerning three batteries of opinion and perception assertions will be discussed. The digital recordings of all 610 completed interviews were of good quality. The sound files were transcribed by three graduate students. Coding of these transcripts was done by three other graduate students and for a small part by the researcher. In order to assess the reliability of the coding, the researcher coded a random sample of 10% of all Q–A sequences, excluding the Q–A sequences originally coded by the researcher. These codes were compared with the original coding. The percentage of agreement in the two coded files appeared to be 82%, and the Cohen's Kappa value was 0.81. According to the scale that Landis and Koch (1977) proposed to describe the degree of agreement, a Kappa value of 0.81 can be considered as "almost perfect".

5.3 Determination of eligible Q–A sequences

To test the hypotheses, we compared for all questions the percentage of Q–A sequences during which a mismatch answer occurs. Of course only cases in which questions are read exactly as worded can be compared. Fortunately in our case almost 99% of the questions was read exactly as worded or with minor changes. In addition we excluded any interactions that took place before, during, or after the interviewer's question reading, as those interactions may have consequences for the interpretation of the relation between question wording and the occurrence of a mismatch answer. We also excluded Q–A sequences with mismatch answers that occurred *after* initially adequate answers. Self-corrections, mismatch answers immediately followed by an adequate answer without intervention of the interviewer, were categorized as adequate answers. Finally, we ended up with 95% of our original Q–A sequences. About 18% of these Q–A sequences contained a mismatch answer according to our operationalizations.

---

[1] In this study the number of non-eligible cases is high, due to a large number of non-working numbers. This might be caused by the fact that the telephone numbers were sampled about 2 months before data collection started. The AAPOR rate RR3 that accounts for estimation of non-eligible cases within cases of unknown eligibility is 45.5.

5.4 Results of study 2

Panel A of Table 2 shows the percentage of Q–A sequences with and without mismatch answers, for the wording of the assertions of the three different batteries of assertions. Overall, the results do not show a confirmation of the question wording hypothesis for all assertions. For the eight GHPQ-assertions, the percentage of Q–A sequences with a mismatch answer was 11 in both conditions. Although for the six 'Public Health' assertions, the percentages of Q–A sequences with a mismatch answer were in the expected direction (19 for the colloquial versions of the questions; 17 for the formal versions) the difference was not significant. For the 'Government and Health' assertions we did find a significant overall difference between the colloquial and formal versions. As the results in the table show, the colloquial versions yield mismatch answers in 22% of the Q–A sequences, whereas for the formal versions this percentage is 16. Hypothesis 1 can only partly be confirmed.

With respect to hypothesis 2, panel B of Table 2 shows the percentage of Q–A sequences with and without mismatch answers for the different response alternatives used. The results show that, according to our expectation, assertions with formal answer alternatives yield more mismatch answers than assertions with colloquial alternatives.

For the eight GHPQ-assertions, the difference in the percentage of Q–A sequences with a mismatch answer is 19 (formal alternatives) versus 4 (colloquial alternatives). This difference is 27 (formal alternatives) versus 10 (colloquial alternatives) both for the 'public health' assertions and the 'Government and Health' assertions.

In the latter case (the 'Government and Health' assertions), not only the wording of the alternatives was manipulated, but also the number of alternatives (two alternatives in case of colloquial wording, and five alternatives in case of formal wording). This was done in order to compare common survey practice (the five alternatives) with alternatives that—as responses to assertions—in our view are least likely to yield mismatch answers (just 'yes' or 'no'). We expected to find even more pronounced differences between these two types of alternatives than for the more properly manipulated alternatives of the GHPQ and 'public health' assertions. Although this does not appear to be the case, the differences for all types of assertions are quite striking.

5.5 Modeling the effects of question, respondent variables

In order to consider the effects of both manipulation of questions and response alternatives, as well as effects of respondent characteristics such as age and level of education, a logistic regression was run.

The unstandardized regression coefficients and the odds ratio (exponent of B, Exp (B) in the table) of significant variables are reported in Table 3. For the categorical variables question wording, type of alternatives and level of education of respondents, we used the following values as a reference category: the 'colloquial questions', 'colloquial alternatives' and 'higher education'.

Model 1, which includes only the main effects of question wording and the response alternatives, shows that the odds of a mismatch answer in a Q–A sequence with formal alternatives are four times higher (odds ratio = 4.03) than those of a Q–A sequence with colloquial alternatives. The effect of question wording, however, is not significant.

In model 2 the interaction of question wording and types of alternatives is included. The results are more or less equal to model 1, whereas the interaction between question wording and wording of the alternatives is not significant.

Table 2 Percentage of Q–A sequences with mismatch answers for three batteries assertions

Q–A sequences	GHPQ assertions (n = 8 questions, 4805 Q–A sequences)			Public Health assertions (n = 6, questions, 3530 Q–A sequences)			Government & Health assertions (n = 3, questions, 1795 Q–A sequences)		
	Colloquial	Formal	Total	Colloquial	Formal	Total	Colloquial	Formal	Total
A. Effects of question wording									
With mismatch	379 11%	153 11%	532	466 19%	191 17%	657	183 22%	151 16%	334
Without mismatch	3108 89%	1255 89%	4273	1967 81%	906 83%	2873	647 78%	814 84%	1461
Total	3397 100%	1408 100%	4805	2433 100%	1097 100%	3530	830 100%	965 100%	1795
	χ^2 = n.s. (1.5% Q–A sequences excluded)			χ^2 = n.s. (3.6% Q–A sequences excluded)			χ^2 = 12.07, $p < 0.01$ (1.9 % Q–A sequences excluded)		

Q–A sequences	Colloquial 'yes', 'maybe' or 'no'	Formal 'true', 'possibly true' or 'false'	Total	Colloquial 'yes' or 'no'	Formal 'agree' or 'disagree'	Total	Colloquial 'yes' or 'no'	Formal 'strongly agree', 'agree', 'neutral', 'disagree' or 'strongly disagree'	Total
B. Effects of response alternatives									
With mismatch	86 4%	446 19%	532	173 10%	484 27%	657	91 10%	243 27%	334
Without mismatch	2303 96%	1970 82%	4273	1589 90%	1284 73%	2873	813 90%	648 73%	1461
Total	2389 100%	2416 100%	4805	1762 100%	1768 100%	3530	904 100%	891 100%	1795
	χ^2 = 269.42, $p < 0.01$ (1.5% Q–A sequences excluded)			χ^2 = 179.59**, $p < 0.01$ (3.6% Q–A sequences excluded)			χ^2 = 87.72** $p < 0.01$ (1.9 % Q–A sequences excluded)		

Table 3 Logistic regressions for the odds of mismatch answers occurring in a Q–A sequence

	Model 1		Model 2		Model 3		Model 4		Model 5	
	B	Exp (B)	B	Exp (B)	B	Exp (B)	B	Exp (B)	B	Exp (B)
Question variables										
Question wording (ref: colloquial)										
Formal	−0.33	0.97	−0.21	0.81	−0.19	0.82	−0.23*	0.79	−0.22	0.80
Alternatives (ref: colloquial)										
Formal	1.39**	4.03	0.32**	3.72	0.39**	4.00	1.54**	4.66	1.55**	4.71
Interaction Q * A (ref: colloquial)										
(Formal * formal)			0.24	1.27	0.19	1.20	0.21	1.23	0.23	1.26
Constant	−2.58**	0.08	−2.53**	0.08	−4.30**	0.01	−3.02	0.05	−4.98**	0.01
Respondent variables										
Age of respondent (continuous variable)					0.03**	1.03			0.03**	1.03
Education of respondent (ref: higher education)										
No education/primary					0.58**	1.80			0.66**	1.93
Lower vocational					0.51**	1.66			0.51**	1.66
Vocational education					0.42**	1.52			0.37*	1.45
Pre-university education					0.08	1.09			−0.07	0.93
Random effects										
Random intercept							1.71	5.53	1.56	4.78
Akaike Information Criterion	8037.54		8032.40		7562.38		54279.71[a]		54055.42	

n = 10, 130 Q–A sequences, 610 respondents

* *p* < 0.05

** *p* < 0.01

[a] In the R-program, scaling of the log likelihood is different for generalized linear models as compared to ordinary logistic regressions. The Akaike Information Criterion of models 3 and 4 can only be compared with each other and cannot be compared to the Akaike Information Criterion from the ordinary logistic regression (i.e., model 1 and 2)

These parameters appear to remain almost the same when both respondent variables age and education are included (model 3). The effects of the respondents' age show that for each year a respondent is older, the odds of a mismatch answer occurring, increases with 3%. The lower levels of education of respondents also differ significantly as compared to the highest levels of education, indicating that the odds of a mismatch answer are higher for respondents in the three lowest levels of education than for respondents in the two highest levels of education (i.e., pre-university and higher educated respondents). However, the effects of respondent variables did not substantially change the effects we found for wording of the alternatives. No significant interaction effects between respondent and question variables could be found, and therefore are not presented here. We found that model 3 predicted the odds of mismatch answers more accurately (as indicated by a lower value for the Akaike Information Criterion, Burnhan and Anderson 2002). Model 3 does not change our conclusion about the effects of wording of alternatives on the odds of a mismatch answer.

In models 4 and 5 the nesting structure of the data (i.e., respondents answered multiple questions of the same version in a battery) is taken into account. Model 4 is a generalized linear mixed model and includes only question variables (similar to model 2). Parameter estimation in generalized linear mixed models is complicated because some kind of approximation is involved (Snijders and Bosker 1999). We used penalized quasi-likelihood (PQL) for which the approximation is around an estimate for the fixed and random part. In model 4, next to the effect of response alternatives, the effect of question wording turns out to be significant. The odds ratio (0.79) shows that the odds of a mismatch answer in a Q–A sequences with formal question wording are lower than in those of a Q–A sequence with colloquial question wording.

We tried to improve the fit of the model by including the respondent's age and education. Model 5 shows results similar to those of model 3. Also for model 5 the effects of respondent variables did not substantially change the effects we found for wording of the alternatives.

From the analyses we conclude that the effects of types of alternatives are much stronger than the effects of question wording, and there is no interaction between question wording and the types of alternatives.

6 Discussion

The goal of the studies presented in this paper was to present and test hypotheses about the occurrence of so-called mismatch answers. One reason for the occurrence of mismatch answers is that people are used to participate in ordinary conversations, and apply their style of responding to survey interviews. When survey questions resemble expressions commonly used in ordinary conversations, respondents will not be focused on the task of giving precisely formatted answers, yielding a high number of mismatch answers. A formal question on the other hand, triggers respondents to focus adequately on the task of formulating precise answers.

In a non-experimental study we were able to confirm this hypothesis. In an experimental study, we could only find partial support for this hypothesis, but in general colloquial and formal questions did not yield a clear difference in the percentage of mismatch answers.

With a second hypothesis we aimed to test the effects of the colloquial character of response alternatives. We assume that respondents are not accustomed to use the formal words that are common in surveys, but instead answer questions in the same way as they do in common conversations. When response alternatives are used that are formulated according to language in ordinary conversations, respondents will have less difficulties with using such

colloquial alternatives, and as a consequence give fewer mismatch answers than when formal alternatives are used. We could confirm this hypothesis in both studies. In this case, using an operationalization of response alternatives based upon word frequencies in ordinary conversations was an effective strategy.

6.1 Success of the manipulations

We were primarily concerned with the use of feasible and realistic survey questions. Therefore we used assertions that were derived from actual surveys. It might have been more efficient to focus on internal validity, by means of creating more extreme manipulations. However, very extremely formal manipulations may yield questions that, although they alert to the formal character of the task, are difficult to understand. Furthermore, extreme manipulations might have alerted interviewers of the experimental character of the study. In this study it was very important that interviewers were unaware of the actual hypotheses being tested. Their knowledge of the expected outcomes could have influenced their behavior in the interaction with the respondent. For example, they could have stressed the importance that respondents formulate their answers as precisely as possible, and they could have done this in different ways for different versions of the same question. In that way it would have been impossible to distinguish effects of interviewer behavior from effects of question wording. Fortunately the interviewers' behavior did not indicate any suspicion with regard to the goals of the study. We could not find any differences in interviewer behavior, related to the different versions of the questions. Moreover, interviewers were surprised when they were, at the end of the study, informed about the experimental character of the study.

For the manipulation of questions it may have been better to base the manipulations of question wording on actual frequencies of grammatical structures or complete sentences, instead of word frequencies. However, to our knowledge a frequency database with such information does not exist.

6.2 Consequences for survey practice

The results of this study showed that especially in case of assertions colloquial alternatives decrease the chance of mismatch answers. Respondents typically treat assertions as yes–no questions, i.e., they typically reply with 'yes' or 'no' and such answers do not fit with the five-point Likert-type scale. Therefore, colloquial alternatives should be used for assertions. Alternatively, if a researcher wishes to use a Likert-type scale, we suggest to use adapted wording of assertions with additions such as "To what extent do you find that…".

Mismatch answers trigger inadequate interviewer behavior such as suggestive probing, which had been shown to influence the response distribution (Smit et al. 1997). According current practice of standardized interviewing, interviewers are not allowed to use any kind of suggestive probe (Fowler and Mangione 1990). Therefore, since the increased likelihood of suggestive probing after mismatch answers, we assume that the occurrence of such answers influences data quality negatively.

A clear-cut consequence of the occurrence of mismatch answers is the fact that it extends the interaction, which will add to interview costs. More attention should be paid to mismatch answers in interviewer training. Interviewers need to learn how to recognize mismatch answers, and how to react appropriately. To better prepare interviewers, examples should be given of typical mismatch answers that could occur for specific questions that are used in the interviews.

6.3 Suggestions for further research

It is important that the relation between mismatch answers and response validity is further studied. We do not know how our manipulations to decrease the occurrence of mismatch answers, by themselves influence response validity. Colloquial responses, that are adequate when colloquial response alternatives are used, might not necessarily yield more valid and reliable data. One of the possible consequences of listing colloquial responses is that, while allowing respondents to use convenient words, respondents may view their task as less important, and as a result process questions less thoroughly.

In addition, the idea that respondents give mismatch answers because they view the interview as a conversation needs to be explored further. Next to question wording, the topic, order and number of colloquial and formal questions within a questionnaire will probably influence this process. Respondents may be more likely to view the interview as an ordinary conversation when topics are discussed that are common for ordinary conversations (e.g., the weather, the prices of goods, political affairs, sports, etc.), than when they are not so common (e.g., specific number of hours and minutes respondents spend on routine activities). When a questionnaire begins with (or consists mostly of) colloquial topics, formal questions will not make much difference in changing the respondent's view of the interview, and vice versa. Finally, next to questionnaire characteristics, interviewer and respondent characteristics may play a role in how respondents view the interview. Interviewers may differ in the extent to which they have a colloquial style of interviewing, thereby evoking a conversational interview. The interviewer style may depend on interviewer characteristics such as age and level of education, and the way interviewers are trained. For respondents characteristics the extent to which they have experience with survey interviews, their age and level of education may be important factors that influence their view of interviews.

Appendix

A. Examples of formal and colloquial questions and alternatives in study 1 (ESS)

See Table 4

B. Question wording study 2 (Experiment)

General health perception assertions

In the questionnaire eight questions were derived from the General Health Perception Questionnaire (Brook et al. 1979). In the GHPQ the subjective perception of one's own health status, in the past, at present and as expected in the future is measured. The questions are formulated as assertions. Two (Dutch) translations of the short version (SF-20) of this questionnaire were available (Kempen et al. 1995; Kriegsman et al. 1995). The assertions in original wording (as indicated with an asterisk), and reworded questions, are listed in Table 5. As might be observed by face value in this table, the original GHPQ-wording was typically colloquial. However, for the English translations of the questions, which are presented for illustration purposes, we did not use word frequency data. As a consequence, the difference may be less apparent as in their Dutch equivalents.

Table 4

	Formal question	Colloquial question
Formal	(*n* = 28 questions)	(*n* = 8 questions)
alternatives	To what extent do you consider yourself associated with this party?	Taking all things together, are you
	1 Very associated	1 Very happy
	2 Fairly associated	2 Fairly happy
	3 Hardly associated	3 Not so happy
	4 Not at all associated	4 Not at all happy
Colloquial	(*n* = 47 questions)	
alternatives	Do you consider yourself as a member of a minority group that is discriminated in this country?	–
	1 Yes	
	2 No	

Table 5 Original colloquial wording and formal rewording of GHPQ SF-20 assertions (Dutch wording in italics)

In original (colloquial) wording	Reworded as a formal question
I worry about my health[a]	My health causes me worries
Ik maak me zorgen over mijn gezondheid*	*Mijn gezondheid baart mij zorgen*
Getting sick sometimes is part of life	Getting sick once in a while is part of life
Af en toe ziek worden hoort bij het leven*	*Sporadisch ziek worden behoort bij het leven*
I have been feeling bad lately	Recently I have had a poor physical health
Ik voel me de laatste tijd slecht*	*Ik heb de afgelopen periode een slechte lichamelijke gesteldheid gehad*
I believe that sometimes I am just going to be sick	I acknowledge that from time to time I will become sick
Ik accepteer het dat ik soms gewoon ziek wordt	*Ik aanvaard dat ik van tijd tot tijd nu eenmaal ziek wordt*
I am as healthy as an ox	My health is excellent
Ik ben zo gezond als een vis	*Mijn gezondheidstoestand is uitstekend**
I seem to get sick a little easier than other people	It appears to me that I get sick easier than other people
Ik lijk wat makkelijker ziek te worden dan andere mensen	*Het komt me voor dat ik gemakkelijker ziek word dan andere mensen*
I am ill[b]	–
Ik ben ziek[b]	–
I was so sick once I thought I might die	I have on one occasion been so sick I thought I would pass away
Ik ben wel eens zo ziek geweest dat ik dacht dat ik doodging	*Ik ben een enkele maal zo ziek geweest dat ik dacht te overlijden*

[a] From the original wording "I never worry about my health" the word 'never' was omitted because it is obviously problematic to use negations in question wording
[b] For this question only a colloquial version was available

Table 6 Original colloquial wording and formal rewording of public health assertions (Dutch wording in italics)

Original (formal) wording	Reworded as a colloquial question
A person is responsible for his own extra public health costs caused by smoking	You should pay for your owns costs of public health caused by smoking
*Iemand is zelf aansprakelijk voor de extra ziektekosten door roken**	*De extra ziektekosten door roken moet je zelf betalen*
A person is responsible for his own extra public health costs caused by drinking alcoholic beverages	You should pay for your own extra public health costs caused by drinking alcoholic beverages
*Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit het drinken van alcohol**	*De extra ziektekosten door het drinken van alcohol moet je zelf betalen*
A person is responsible for his own extra public health costs caused by careless driving	You should pay for your own extra public health cost caused by careless driving
*Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit nalatigheid in het verkeer**	*De extra ziektekosten door nalatigheid in het verkeer moet je zelf betalen*
A person is responsible for his own extra public health costs caused by sports	You should pay for your own extra public health cost caused by sports
*Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit sporten**	*De extra ziektekosten door sporten moet je zelf betalen*
Elderly should contribute more to health insurance than youngsters	The elderly should pay more health insurance premium than youngsters
*Ouderen (…) zouden meer moeten bijdragen aan de ziektekostenverzekering dan jongeren**	*Ouderen zouden meer ziektekostenpremie moeten betalen dan jongeren*
People with a healthy lifestyle, for instance with attention for their food, should contribute less to the health insurance	People who live very healthy by for instance taking notice of their food should pay less for the health insurance
Mensen met een gezonde levensstijl, bijvoorbeeld met aandacht voor hun voeding, zouden minder moeten bijdragen aan de ziektekostenverzekering	*Mensen die heel gezond leven door bijvoorbeeld te letten op hun eten, zouden minder moeten betalen aan de ziektekostenverzekering**

Public health assertions

The six assertions concerning the respondent's opinion on costs of public health were included were derived from (Dutch) studies (Bernts 1991; Elchardus et al. 2002; Van de Berg et al. 1986). The questions in the original version (as indicated with an asterisk) and reworded questions are presented in Table 6.

Government and health assertions

In the questionnaire, three assertions taken from the Dutch Government 'Belevingsmonitor' ('Perception monitor' RVD 2003) were included. These concerned the respondent's opinion about a smoke free hotel and catering industry, education about smoking, and inspection of food safety. For these assertions, as shown in Table 7 , we used the original formal wording and a colloquial rewording.

Table 7 Original colloquial wording and formal rewording of government and health assertions (Dutch wording in italics)

Original (formal) wording	Reworded as a colloquial question
The government should provide civilians with a smoke free hotel and catering industry	The government should take care of smoke free cafes and restaurants for the people
De overheid dient er voor zorgen dat burgers gebruik kunnen maken van rookvrije horeca*	*De regering moet er voor zorgen dat er voor de mensen rookvrije cafés en restaurants zijn*
It is necessary that the government educates civilians about the consequences of smoking and passive smoking	I think the government should educate the people about the consequences of smoking and passive smoking
Het is nodig dat de overheid de bevolking voorlicht over de gevolgen van roken en meeroken*	*Ik vind dat de overheid voorlichting aan de mensen moet geven over de gevolgen van roken en meeroken*
I trust that inspections are good so stores are supplied with safe food	I think our food is well checked so what is in stores is safe
Ik vertrouw erop dat er goed wordt gecontroleerd, zodat winkels zijn voorzien van veilig voedsel*	*Volgens mij wordt ons eten goed gecontroleerd zodat wat in de winkel ligt veilig is*

Table 8 Wording of colloquial and formal answer alternatives for the assertions

Colloquial answer alternatives	Formal answer alternatives
GHPQ-Assertions	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with 'yes', 'maybe' or 'no'	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with 'true', 'possibly true' or 'false'
Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'ja', 'misschien' of 'nee'	Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'waar', 'mogelijk waar' of niet waar'
Public health assertions	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with 'yes' or 'no'	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with 'I agree' or 'I disagree'.
Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'ja', of 'nee'	Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'mee eens', of 'oneens'
Government and health assertions	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with 'yes' or 'no'	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with the following five answer possibilities 'strongly agree, agree, neutral, disagree or strongly disagree'.
Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'ja', of 'nee'	Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'zeer mee eens', 'mee eens', 'neutraal', 'mee oneens' of 'zeer mee oneens'

Colloquial and formal response alternatives

In Table 8 the colloquial and formal alternatives for the three batteries of assertions are presented. To illustrate the context of the alternatives, the introductory statements that belong to the answer alternatives are included.

References

Bernts, T.: Leven zonder zorg. Oordelen over risico's, rechtvaardigheid en solidariteit in de gezondheidszorg. Swets en Zeitlinger, Amsterdam (1991)

Brennan, S.E., Clark, H.H.: Conceptual pacts and lexical choice in conversation. J. Exp. Psychol.: Learn. Mem. Cogn. **22**(6), 1482–1492 (1996)

Brook, R.H., Ware, J.E., Jr., Davies-Avery, A., Stewart, A.L., Donald, C.A., Rogers, W.H., Williams, K.N., Johnston, S.A.: Overview of adult health status measures fielded in Rand's Health Insurance Study. Med. Care, **17** (7 Suppl), iii–x, 1–131 (1979)

Burnhan, K.P., Anderson, D.R.: Model Selection and Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York (2002)

Cannell, C.F., Oksenberg, L., Converse, J.M.: Striving for response accuracy: experiments in new interviewing techniques. J. Mark. Res. **XIV**, 306–315 (1977)

Dijkstra, W.: A new method for studying verbal interactions in survey interviews. J. Off. Stat. **15**(1), 67–85 (1999)

Dijkstra, W., Ongena, Y.P.: Question–answer sequences in survey-interviews. Qual. Quant. **40**(6), 983–1011 (2006)

Elchardus, M., Tresignie, C., Derks, A.: Project: het draagvlak van de solidariteit. Deelrapport 3: Levensstijl- en levensloopaansprakelijkheid.: Vakgroep Sociologie, Onderzoeksgroep TOR, Vrije Universiteit Brussel (2002)

ESS: European Social Survey. Available from http://www.europeansocialsurvey.org/ (2007). Cited March 5 2007

Fowler, F.J., Mangione, T.W.: Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Sage, Newbury Park, CA (1990)

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

Kempen, K.I.J.M., Brilman, E.I., Heyink, J.W., & Ormel, J.: Het meten van de algemene gezondheidstoestand met de MOS Short-Form General Health Survey (SF-20): een handleiding (p. 32). Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken, Rijksuniversiteit Groningen (1995)

Kriegsman, D.M.W., Van Eijk, J.T.M., Deeg, D.J.H.: Psychometrische eigenschappen van de Nederlandse versie van de RAND General Health Perception Questionnaire: de Vragenlijst Algemene Gezondheidsbeleving (VAGB) [Psychometric properties of the RAND General Health Perception Questionnaire in The Netherlands]. Tijdschrift Sociale Gezondheidszorg **73**(6), 390–398 (1995)

Krosnick, J.A.: Survey research. Annu. Rev. Psychol. **50**, 537–567 (1999)

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **45**, 255–268 (1977)

Ongena, Y.P.: Problematic deviations in question–answer sequences, p. 40. Unpublished manuscript, Vrije Universiteit, Amsterdam (2001)

Ongena, Y.P., Dijkstra, W.: Methods of behavior coding of survey interviews. J. Off. Stat. **22**(3), 419–451 (2006)

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., Baayen, H.: Experiences from the Spoken Dutch Corpus Project. In: González Rodriguez, M., Paz Suárez Araujo, C. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation, pp. 340–347 (2002)

Prüfer, P., Rexroth, M.: Zur Anwendung der Interaction-Coding-Technik. ZUMA-Nachrichten **17**, 2–49 (1985)

RVD (2003) Belevingsmonitor Rijksoverheid, juni 2003. Den Haag: Rijksvoorlichtingsdienst/Publiek en Communicatie

Schaeffer, N.C., Maynard, D.W.: From paradigm to prototype and back again: interactive aspects of cognitive processing in standardized survey interviews. In: Schwarz, N., Sudman, S. (eds.) Answering Questions Methodology for Determining Cognitive and Communicative Processes in Survey Research, pp. 65–88. Jossey-Bass, San Francisco (1996)

Schuman, H., Presser, S.: Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. Academic Press, New York (1981)

Smit, J.H.: Suggestieve vragen in survey-interviews. Vóórkomen, oorzaken en gevolgen. Academisch proefschrift, Vrije Universiteit, Amsterdam (1995)

Smit, E.G., Neyens, P.C.: Segmentation based on Affinity for Advertising. J. Advert. Res. **40**, 35–43 (2000)

Smit, J.H., Dijkstra, W., Van der Zouwen, J.: Suggestive Interviewer Behaviour in Surveys: An Experimental Study. J. Off. Stat. **13**(1), 19–28 (1997)

Snijders, T.A.B., Bosker, R.J.: Multilevel Analysis. An introduction to basic and advanced multilevel modeling. Sage publications, London (1999)

Van de Berg, J., Jansen, R., Haveman, H.: Solidariteitsvoorkeuren ten aanzien van ziektekostenverzekering 1985. Maandbericht Gezondheidsstatistiek **5**(8), 5–16 (1986)