Discourse Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hdsp20

Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions

Naomi Kamoen [a], Bregje Holleman [a], Pim Mak [a], Ted Sanders [a] & Huub van den Bergh [a b]

[a] Utrecht Institute of Linguistics OTS Utrecht University, The Netherlands

[b] Graduate School of Teaching and Learning University of Amsterdam, The Netherlands

Accepted author version posted online: 09 Jun 2011.Version of record first published: 08 Jun 2011.

To cite this article: Naomi Kamoen, Bregje Holleman, Pim Mak, Ted Sanders & Huub van den Bergh (2011): Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions, Discourse Processes, 48:5, 355-385

To link to this article: http://dx.doi.org/10.1080/0163853X.2011.578910

Discourse Processes, 48:355–385, 2011

ISSN: 0163-853X print/1532-6950 online
DOI: 10.1080/0163853X.2011.578910

Routledge
Taylor & Francis Group

Agree or Disagree? Cognitive Processes in Answering Contrastive Survey Questions

Naomi Kamoen, Bregje Holleman, Pim Mak, and Ted Sanders
Utrecht Institute of Linguistics OTS
Utrecht University, The Netherlands

Huub van den Bergh
Utrecht Institute of Linguistics OTS
Utrecht University, The Netherlands
Graduate School of Teaching and Learning
University of Amsterdam, The Netherlands

Survey designers have long assumed that respondents who disagree with a negative question ("This policy is bad.": *Yes* or *No*; 2-point scale) will agree with an equivalent positive question ("This policy is good.": *Yes* or *No*; 2-point scale). However, experimental evidence has proven otherwise: Respondents are more likely to disagree with negative questions than to agree with positive ones. To explain these response effects for contrastive questions, the cognitive processes underlying question answering were examined. Using eye tracking, the authors show that the first reading of the question and the answers takes the same amount of time for contrastive questions. This suggests that the wording effect does not arise in the cognitive stages of question comprehension and attitude retrieval. Rereading a question and its answering options also takes the same amount of time, but happens more often for negative questions. This effect is likely to indicate a mapping difference: Fitting an opinion to the response options is more difficult for negative questions.

Correspondence concerning this article should be addressed to Naomi Kamoen, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands. E-mail: n.kamoen@uu.nl

In surveys, language is used to address people's opinions. Therefore, surveys are an excellent means to study the relation between language and attitudes. Among other things, survey studies show that seemingly irrelevant linguistic characteristics influence how respondents express their attitudes (e.g., Ayidiya & McClendon, 1990; Hippler & Schwarz, 1987; Schuman, 2008; Tourangeau, Rips, & Rasinski, 2000).

One such influential characteristic is the polarity of the question. In survey handbooks, it is often advised to mix positive and negative wordings throughout the questionnaire (e.g., Dillman, 2000; Sudman & Bradburn, 1982). This means that, for example, in a survey on hotel quality, a positive question about the service ("The service in this hotel is good.": *Yes* or *No*; 2-point scale) should be followed by a negative question about the rooms ("The rooms in this hotel are ugly.": *Yes* or *No*; 2-point scale). The assumption that lies at the heart of this advice is that at the level of the individual question, the choice for positive or a negative wording is irrelevant: Respondents who answer *yes* to a *positive* question are expected to answer *no* to an equivalent *negative* question.

As early as in 1941, Rugg tested this assumption. He posed one group of respondents the positive question, "Do you think the government should allow public speeches against democracy?" (*Yes* or *No*; 2-point scale), and another group of respondents the equivalent negative question, "Do you think the government should forbid public speeches against democracy?" (*Yes* or *No*; 2-point scale). Results showed that, contrary to expectations, respondents react differently to such so-called contrastive questions: Respondents are more likely to answer *no* to the negative *forbid* question than to answer *yes* to the equivalent positive *allow* question.

Since Rugg's (1941) study, wording effects for contrastive questions have been studied extensively, with a peak in the 1980s and 1990s. Initially, this research focused on response effects for forbid and allow questions (e.g., Glendall & Hoek, 1990; Hippler & Schwarz, 1986; Krosnick & Schuman, 1988; Loosveldt, 1997; Narayan & Krosnick, 1996; Schuman & Presser, 1996; Waterplas, Billiet, & Loosveldt 1988). Although a significant effect of question wording was not observed in each and every study (e.g., Bischop, Hippler, Schwarz, & Strack, 1988), a meta-analysis has shown that the so-called forbid/allow asymmetry can be generalized beyond the question level: All in all, respondents are more likely to answer *no* to *forbid* questions than to answer *yes* to equivalent *allow* questions (Holleman, 1999b). This means that respondents express their opinions more positively when the question is phrased negatively.

Similar response effects have been shown for other contrastive word pairs and for other scale types. For example, the forbid/allow asymmetry was also shown to arise on a 7-point agree–disagree scale (Holleman, 2000). Furthermore, Waterplas et al. (1988) found response effects for other contrastive verb pairs,

such as *abolish/maintain*. In addition, response effects for contrastive adjectives have been shown (e.g., Javeline, 1999; see also a summary of older work in Molenaar, 1982). As with forbid and allow questions, these studies show that the size and the direction of response effects varies at the level of the individual question: An effect cannot actually be observed for every question in every experiment. However, in a meta-analysis for various contrastive adjectives, Kamoen, Holleman, and Van den Bergh (2007) observed that, when generalizing over questions, people express their opinions more positively when the question is phrased negatively. This is similar to the effect found for forbid and allow questions.

The unexpected effect of question polarity has practical implications for science, politics, and all other areas in which surveys are widely used. What are we measuring with contrastive questions? To answer this question, the reason *why* the choice for a positive or a negative wording influences the answers has to be determined. This can be done by learning more about the cognitive processes involved in answering contrastive survey questions (e.g., Belli, 2005; Holleman & Murre, 2008; Tanur, 1999).

Tourangeau et al. (2000; see also Tourangeau & Rasinski, 1988) introduced a model describing the cognitive processes underlying question answering. Their cognitive model distinguishes four steps. First, a respondent answering a survey question should interpret the question. This means that the respondent has to make a logical representation of the question, and has to determine what kind of opinion or attitude is asked for. In the second stage, the respondent has to retrieve relevant attitudinal information from the long-term memory, which means that beliefs related to the attitude object in the question are gathered. Tourangeau et al. (2000) described this stage as a kind of sampling process in which the most accessible beliefs are activated. Respondents with well-formed attitudes do not retrieve separate beliefs; instead, they activate a summary evaluation of their beliefs directly. In the third stage, the respondent has to render a judgment by integrating all information retrieved. Sometimes this step may be unnecessary, for example, for respondents who retrieve an extreme summary evaluation like *smoking is despicable*. In other cases, however, respondents must weigh and scale their beliefs to come to a judgment. In the fourth stage, the respondent has to fit the judgment made to the answering options in the question. During this process, the answer may be adapted for reasons of social desirability (for more information on the four stages, see Tourangeau et al., 2000).

The "Tourangeau model" (Tourangeau et al., 2000) gives a sensible description of the question-answering process, and has been very influential in the area of survey research: Many research findings have been integrated into this model. However, before the model can be used to test hypotheses about the stage in which the choice for a positive or a negative question wording becomes

relevant, we have to know how the different stages of the model are related. Tourangeau et al. (2000) presented two views on the sequencing of the cognitive processes. On the one hand, they argued that the four cognitive stages follow each other linearly: A question must be understood before information can be retrieved, all relevant information must be retrieved before being integrated, and an opinion must be formed before being mapped onto the response options (pp. 14–16). On the other hand, Tourangeau et al. (2000) gave various reasons for why there may be an overlap among components of the model (pp. 14–16). For example, during the incremental process of question interpretation (Stage 1), respondents may already start retrieving attitudinal information and world knowledge from memory (Stage 2). This suggests that the stages of question comprehension and attitude retrieval cannot easily be separated. Tourangeau et al. (2000) also noted that the judgment stage (Stage 3) can occur alongside the retrieval stage (Stage 2). For example, respondents may update their judgment while retrieving information from memory. Furthermore, it may be possible to backtrack from the judgment stage to the retrieval stage if not enough information is initially retrieved. The judgment stage can even be skipped entirely, as respondents may activate a summary evaluation and match that to the answering scale. Therefore, it is also difficult to separate the retrieval stage from the judgment stage.

To enable the measurement of the cognitive stages, the Tourangeau model (Tourangeau et al., 2000) can be simplified into a two-stage model (Chessa & Holleman, 2007; Holleman, 1999a). In the first stage of this model, the question is understood, information is retrieved, and a judgment is formed (Stages 1, 2, & 3 of the original model). We refer to this stage as the *comprehension-retrieval stage*. In the second, so-called *mapping stage*, the judgment made is fitted to the presented response options (Stage 4 of the original model).

The distinction between these two cognitive stages is relevant with respect to the validity of survey questions. If question wording influences the comprehension-retrieval stage, this means that question wording causes respondents to activate a different judgment about the attitude object in the question. This also implies that either positive or negative questions (or both) do not measure the attitude the researcher intends to measure. In contrast, if the wording of the question affects the mapping stage, this means that respondents come to the same judgment when answering positive and negative questions, but this judgment is translated differently to the response options depending on the question wording. This suggests that positive and negative questions are equally valid, but that the meaning of the response options differs due to the choice for a positive or a negative wording of the question: Although response options like *yes* and *no* are straight opposites, the meaning of *yes* as an answer to a *positive* question is not identical to the answer *no* to a *negative* question (cf. Holleman, 2000).

For forbid and allow questions, two previous studies have investigated whether response effects arise in the comprehension-retrieval stage or in the mapping stage. First, a correlational study by Holleman (1999a) showed that forbid and allow questions measure the same underlying attitude.[1] As the answers to forbid and allow questions do differ, Holleman (1999a) concluded that wording effects arise in the mapping stage. Second, in a reaction time study, Chessa and Holleman (2007) showed that the comprehension-retrieval process for forbid and allow questions (the reaction time measured from onset of the question until a button press to indicate the question was read) takes the same amount of time, whereas mapping the answer to the response options (the reaction time from onset of the answers to giving an answer) takes longer for forbid questions. As a longer processing time reflects processing complexity (e.g., Bassili, 1996; Bassili & Scott, 1996; Fazio, 1990), Chessa and Holleman concluded that mapping an answer to the response options is more difficult for negative questions.

In sum, there seems to be converging evidence for the conclusion that wording effects for contrastive questions arise in the mapping stage, at least for forbid and allow questions. However, this conclusion can be criticized on experimental grounds. First, the correlational study relies heavily on an a priori separation of the cognitive processes at hand and, in fact, only gives indirect evidence for cognitive differences. Second, the reaction time study suffers from problems with the ecological validity: By offering question and answering options separately, there was no possibility for respondents to switch between question and answering options. Recent studies into answering behavior (Galesic, Tourangeau, Couper & Conrad, 2008; Graesser, Cai, Louwerse, & Daniel, 2006) showed that respondents frequently switch between question and response options. Disabling switching may, therefore, have caused a distortion in the cognitive processes measured. Hence, it is uncertain whether the differences in answering time actually reflect mapping differences.

In addition, the main conclusion drawn by Holleman (1999a) and Chessa and Holleman (2007) can be debated on theoretical grounds. In a classical study, Clark (1976) showed that sentences with a negative term take more processing

[1] More specifically, Holleman's (1999a) analyses focused on the congenericity of forbid and allow questions. The basic idea of such an analysis is to compare scores that are undone of measurement errors (Jöreskog, 1971). Holleman (1999a) showed that when correcting for measurement error, the correlation between forbid and allow questions is identical to the correlation between two identical forbid questions, as well as to the correlation between two identical allow questions. This suggests that forbid and allow questions measure the same underlying attitude, which implies that response effects do not arise in the comprehension-retrieval stage. From this it follows that response effects arise in the mapping stage, as the answers to equivalent forbid and allow questions do differ when not corrected for measurement error.

time than sentences containing a positive term. The conclusion drawn from these results is that negative terms cause an increased cognitive load and are, therefore, inherently more difficult to comprehend than their positive counterparts. Although this conclusion is based on experimental evidence obtained outside of a survey context, the inherent difficulty of negative terms is also expected to show in a survey context. Based on Clark (1976), a difference in the comprehension-retrieval process is, thus, to be expected.

Considering the problems with previous survey studies and the contrasting evidence from a general discourse perspective, this study investigates, once more, whether response effects for contrastive questions arise in the comprehension-retrieval stage or in the mapping stage. An answer to this question is important to obtain insight into the validity of contrastive questions. From a general discourse perspective, this research question is also relevant because surveys provide a natural language use context to study the processing of positive and negative terms. Eye-tracking methods are used to investigate the cognitive processes underlying question answering for a broad range of contrastive word pairs. With this method, respondents can pursue their answering behavior in a natural way. If differences in fixation time occur during the initial reading of the question, the comprehension-retrieval process is likely to be affected by the choice of wording. If differences in the question-answering process occur after the respondent has looked at the response options, these probably mainly reflect mapping differences.

METHOD

Experimental Design

Respondents completed a Web survey with 90 questions on smoking policies. Two versions of the survey were constructed: Questions phrased positively in Version 1 were worded negatively in Version 2, and vice versa. Hence, both survey versions contained both positive and negative survey questions. Sixty questions in the survey were manipulated this way. Thirty questions served as fillers and had an identical wording in the two survey versions.

Participants

Participants ($N = 56$) registered for the experiment on a Web site of the humanities faculty. They were randomly assigned to either of the two survey versions. The eye movements of one half of the participants were measured ($N = 28$). Table 1 shows some demographic characteristics of the respondents.

TABLE 1
Characteristics of the Respondents in the Sample

Gender	*Age*	*Educational Level*	*Smoking Behavior*
Female: 84%	16–20: 36%	University student: 93%	Non-smoking: 79%
Male: 16%	21–25: 55%	College degree student: 7%	Party smoker: 13%
	26–30: 5%		Some cigarettes per day: 7%
	>30: 4%		One pack per day: 2%

Note. $N = 56$.

Procedure

Respondents each came to the laboratory at Utrecht University for individual sessions. They were told that they were going to answer attitude questions on Dutch smoking policies. Eye movements were measured using a Tobii 1750 remote eye tracker (Tobii Technology, Danderyd, Sweden). The hardware of this eye tracker resembles an ordinary computer monitor. Near-infrared beams capture the respondents' eye movements. The frame rate of the Tobii is 50 Hz, which means that the position of the eyes is tracked every 20 ms.

After a short explanation of the eye tracker, the experimenter started the calibration procedure to locate the position of the respondents' eyes. Once the calibration had succeeded, the survey was started. The respondents completed the survey in about 15 min, and they were paid €5 for their participation.

The Survey

The survey started with an introduction to the topic of the survey: Dutch smoking policies. At the time the experiment was carried out (May and June 2008), smoking policies were a hot topic in the Dutch public debate, as they were about to be changed. The respondents were instructed to answer the questions about this topic truthfully, and to click with the left mouse button on the answering option that best matched their opinion. Furthermore, it was explained that before each question, an asterisk would appear on the screen. Respondents were instructed to look at the asterisk and to click on it. By clicking on the asterisk, the question would appear exactly in that spot. This was done to ascertain that respondents started reading the question from the beginning (i.e., the first word).

Then, the actual survey started. The questions in the survey were presented one by one to the participants. Important in this respect is the visual presentation of the questions, as all sorts of design characteristics may affect the mapping process (e.g., Couper, Conrad, & Tourangeau, 2007; Tourangeau, Couper, &

Conrad, 2004, 2007). To prevent respondents from getting distracted, the visual design of each Web page was kept as simple as possible (see Figure 1).

The questions in the survey were organized per subtopic, and each subtopic was introduced on a separate Web page in the survey. All manipulated questions were about one of the four topics in Table 2. Respondents answered several positive and negative questions about one and the same topic. Within each of the four clusters of questions, various word pairs were used for the manipulations, such as *bad/good* (*slecht/goed*), *forbid/allow* (*verbieden/toelaten*), and *difficult/easy* (*moeilijk/makkelijk*). As shown in Table 2, the answering scales (2-point yes or no scales vs. 5-point agree–disagree scales) were varied between clusters, but they were kept identical within each cluster. When analyzing clusters of questions, we can, therefore, generalize over linguistic contrasts while explicitly distinguishing between 2-point scale yes or no questions and 5-point scale agree–disagree questions.

Please note that, although the same kind of response scale was used within each cluster, the filler questions were used to make sure that the answering scales

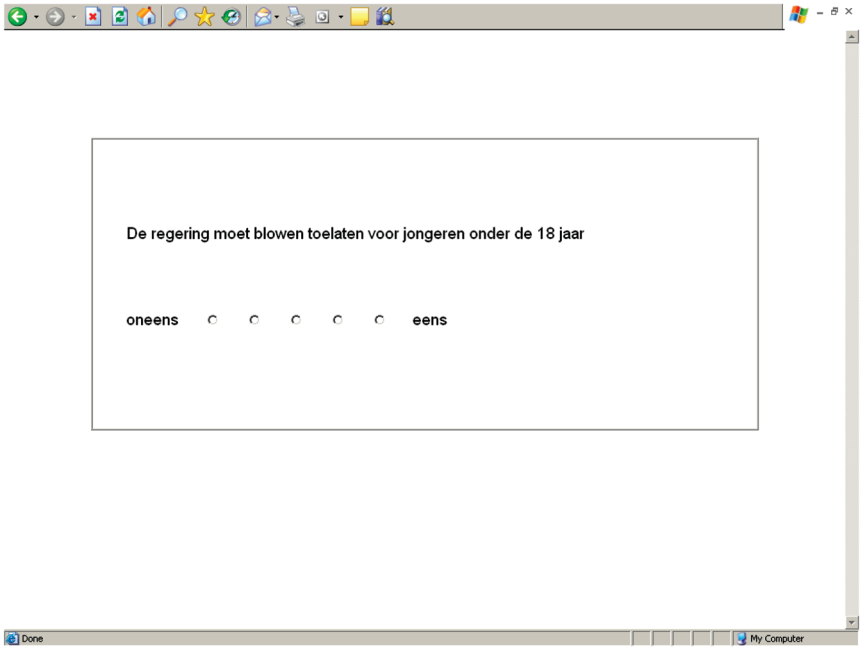

FIGURE 1 Example of the survey layout (color figure available online). *Note.* The question in the example can be translated as follows: "The government should allow smoking pot for youngsters under 18," and the answers range from "disagree" to "agree."

TABLE 2
Characteristics of the Four Clusters of Questions

Cluster	*N Questions*	*Scale Type*	*Example Question*
New smoking policies	10	2-point	I think it is a good/bad idea to increase taxes on cigarettes from July 1st onwards.
Other measures to prevent smoking	7	2-point	It is a good/bad idea to exclude smoking couples from having fertility treatments.
Role of politics in smoking prevention	14	5-point	I think it is just/unjust for the government to fine violators of the smoking laws.
How to protect the weak against smoke	14	5-point	In my view it is acceptable/unacceptable that people smoke in ice bars.

(2-point vs. 5-point) were varied throughout the entire survey. The filler items were sometimes 2-point yes or no questions and other times 5-point agree–disagree questions. Moreover, about one half of the filler items were frequency indicators using 5-point often–never scales. Because the type of response scale was varied between questions, respondents were unaware of the response options until the answers were fixated. In addition, because very distinct scale types were used (e.g., 2-point scales do not include a middle option and do not allow for much nuance, whereas 5-point scales do), the mapping process can only start when the respondent fixates on the answers. Hence, these design characteristics facilitate the separation of the two cognitive processes of interest. We also made sure that the question wording did not reveal the type of response scale used. For example, the question, "I think it is fascinating/boring to think about Dutch political issues," can be followed by a 2-point yes or no scale, a 5-point agree–disagree scale, or a 5-point often–never scale. In addition, the answering options were outside of the visual span of the respondent as long as the respondent fixated on the question. Therefore, all fixations on the question are reflections of comprehension-retrieval processes until the respondent fixates on the answers.

Measures of the Comprehension-Retrieval Process and the Mapping Process

The assumption underlying the use of eye-tracking measures is that respondents fixate on the word they process (Rayner, 1998). To analyze recordings of eye movements, the experimental material is divided into regions. Time measures on these regions are compared between experimental conditions. Many scientific

fields using eye tracking have developed more or less standard regions and standard eye-tracking measures for assessing processing difficulties. Unfortunately, no such standard measures have yet been developed for assessing the comprehension-retrieval process and the mapping process in answering survey questions. In the remaining part of this section, we discuss the measures we used, and we argue why these measures reflect the two cognitive processes we investigate.

For analyzing the eye-tracking data, each Web page was divided into two regions: the question and the answering options. For each of the two regions, the first pass reading time and total fixation time for all (possible) rereading turns was measured. This resulted in a broad categorization of the question-answering process in four processing measures: the first pass reading time for the question, the first pass reading time for the answering options, the remaining total fixation time for the question, and the remaining total fixation time for the answers. These measures are illustrated in Figure 2. As rereading a question or an answer will probably not happen for each trial, we also dichotomously registered whether rereading the question or the answers occurs. These two measures for the occurrence of rereading are analyzed in addition to the four processing time measures.

All six eye-tracking measures play an important role in characterizing the question-answering process. The first pass reading time for the question necessarily only reflects comprehension-retrieval processes (see The Survey section). If negative questions are more difficult to comprehend than their positive

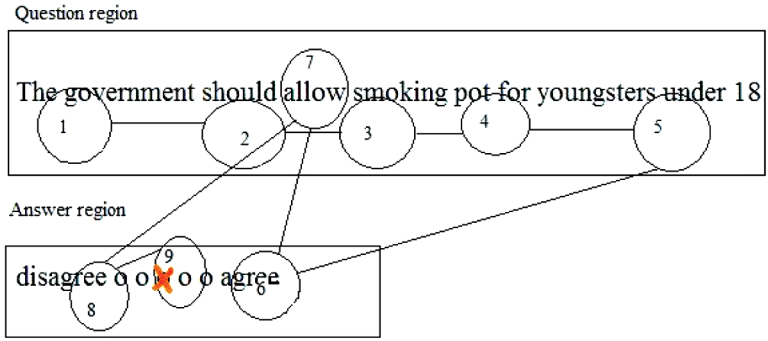

FIGURE 2 Fictitious example of an eye-movement pattern to explain the four time measures. *Note.* The circles indicate fixations, and the number in each circle indicates the fixation number and, thus, the fixation sequence. In this example, Fixations 1 through 5 are part of the first pass reading time of the question, Fixation 6 is the first pass reading time for the answers, Fixation 7 is the remaining total fixation time for the question, and Fixations 8 and 9 comprise the remaining total fixation time for the answers.

counterparts (Clark, 1976), a time difference on this measure is, hence, to be expected. More important, as the Clark account predicted negative questions to be *inherently* more difficult to comprehend, a time difference on the first pass reading time for the question is always to be expected, irrespective of whether a wording effect actually occurs for a certain question.

From the moment the respondent looks at the response options, the mapping process may start. This by no means implies that when a respondent first looks at the response options all comprehension-retrieval processes are abruptly ended. However, considering that this survey contains relatively easy questions (see Table 2), and considering that these questions are answered by skilled readers (see Table 1), it is likely that, starting from the moment the respondent views the response options, the main cognitive activity he or she is involved in is mapping an answer to the response options. Nevertheless, it is useful to distinguish different measures for the mapping process, as a mapping effect may occur in at least three distinct ways.

First, there may be differences between positive and negative questions in the time needed for reading the answering options the first time. If a respondent answers a question immediately upon the first reading of the response options, the first pass reading time for the answers gives a more-or-less unbiased estimate of the mapping process. In such cases, all potential mapping activities must have taken place in the time needed to read the response options.

Second, for those trials in which rereading actually occurs, there may be differences between positive and negative questions in the average time needed for rereading the question and the response options. Although it may be argued that rereading a question or an answer reflects extended comprehension-retrieval processes, we, because of the easy questions and the skilled respondents, a priori assume that, in rereading, the respondent reviews the question in light of the answers and switches back to the answers again to give an answer. Hence, we assume that rereading time reflects mainly mapping processes.

Third, there may be differences between positive and negative questions in the dichotomous registration of rereading: Negative questions and answers to negative questions may be reread more often than positive questions and answers to positive questions. Again, for reasons previously mentioned, for the time being, we assume that a difference in the occurrence of rereading reflects a mapping difference; we return to alternative explanations in the Conclusion and Discussion section. For the sake of clarity, we would like to stress the difference between the dichotomous registration of rereading and the remaining total fixation times: The remaining total fixation times are only calculated for those trials in which rereading actually occurred, whereas the dichotomous registration of rereading applies to all trials. Thus, a distinction can be made between negative questions, which cost more time to reread, and those that are reread more often.

Analyses

Answers to the survey questions. The answers to the survey questions are analyzed per cluster of questions (see Table 2).[2] For each cluster, a separate multilevel model was constructed. In each model, one mean answer for all positive questions in a cluster is estimated, and another mean is estimated for all negative questions. The model also estimates question variance (because one question can be more controversial than another), between-person variance (because one person can express his or her opinion more positively than another), and residual variance (e.g., because one person may agree more with Question A, whereas another person agrees more with Question B). We refer to Appendix A for a formalization of this model. Note that by using this model, the hierarchical structure of the data is taken into account (cf. Quené & Van den Bergh, 2004, 2008; Yan & Tourangeau, 2008). The mean answers to positive and negative questions are compared in a contrast test, which yields a chi-square distributed test statistic (Goldstein, 2003; Snijders & Bosker, 1999).

Measures of processing time.[3] There are four measures of processing time: the first pass reading time for the question, the first pass reading time for the answers, the remaining total fixation time for the question, and the remaining total fixation time for the answers. For some trials—that is, for some Respondent × Item combinations—only a single reading of the question and the answers is involved. Consequently, for these "single-reading trials," the remaining total fixation time for the question and the remaining total fixation time for the answers are nonexistent. That is why, for these trials, only an average first pass fixation time for the question and the answers is estimated for positive

[2]Because wording effects vary largely over questions, we need some method to generalize over questions. As the residual variance would increase hugely by pooling all questions in this survey (scores will then vary from 1–5, because, for instance, a smoker is generally against the prohibition of smoking in the hospitality business, but in favor of the protection of children against smoke), we decided on this cluster approach. An additional advantage is that in survey practice, answers are often pooled over clusters of questions making it useful to know how the mean answer toward one topic is biased by the choice of wording.

[3]Fixations were determined by an algorithm that restricts fixations to data points within 35 pixels; the minimal fixation length was set at 100 ms. The viewing time in a region was computed as the time from the beginning of the first until the end of the last successive fixation in a region. For neither one of the clusters, the distribution of the raw and untransformed data was comparable to the normal distribution. This is a common phenomenon for processing data (e.g., Yan & Tourangeau, 2008). Therefore, we took the natural log of the fixation times and checked the normality again. Two trials—that is, two Respondent × Item combinations—showed unusually short total average fixation times; these cases were removed from the data (<1%). We think that, in these cases, respondents accidently clicked an answer before reading the question. After removing these cases, the residuals of the total reading time, as well as the residuals of each of the individual four time measures, showed a normal distribution.

and negative questions. This results in 2 (Relevant Type Processing Measures) × 2 (Positive or Negative) average processing times. For trials in which the question or the answers are reread, all four processing measures are relevant. For these "rereading trials," another 4 (Type of Processing Measure) × 2 (Positive or Negative) average fixation times are estimated. Taken together, this means that 12 mean processing times are estimated. The question variance, between-person variance, and residual variance are also simultaneously estimated.

In a subsequent contrast test, all four measures of processing time can be compared between positive and negative questions using this model. In addition, the first pass fixation time for the question and the answers can be compared between the single-reading trials and the rereading trials; this is important, as it helps us learn more about the differences between these trials and the kinds of cognitive processing rereading reflects. Additional information on these statistical models is provided in Appendix B.

Dichotomous registration of rereading. For the analysis of whether rereading occurs, separate multilevel models are constructed for each cluster of questions. These models predict the chance that rereading occurs for positive and negative questions and answers. This results in 2 (Region: Question or Answers) × 2 (Positive or Negative) predictions of whether rereading occurs. Again, question variance and between-person variance are estimated. However, the residual variance cannot be estimated because the data are binomial; for binomial data, the residual variance cannot be estimated separately from the mean scores (Goldstein, 2003; Snijders & Bosker, 1999). Mean rereading occurrences for positive and negative questions are compared in a contrast test (Goldstein, 2003; Snijders & Bosker, 1999). Additional information on the models used for analyzing the dichotomous registration of rereading is provided in Appendix C.

RESULTS

Prerequisites[4]

If the survey measured actual attitudes on smoking policies, it is to be expected that the clusters of questions measured respondents' opinions in a reliable way. Cronbach's alphas for each of the four clusters are shown in Table 3.

[4]Most of the results reported in this section are based on multilevel models comparable to those described in an earlier section (see Analyses section); this explains why χ^2 tests are reported and why the N values are left out of the reported test statistics. Analyses of answering data for individual questions (age, gender, etc.) are based on "ordinary" χ^2 tests; in these cases, the N values are reported.

TABLE 3
Cronbach's Alphas of the Four Clusters of Questions

Cluster	*Scale Type*	*N Questions*	α
New smoking policies	2-point	10	0.83
Other measures to prevent smoking	2-point	7	0.61
Role of politics in smoking prevention	5-point	14	0.84
How to protect the weak against smoke	5-point	14	0.75

A Cronbach's alpha of 0.6 is usually regarded as the lower bound of an acceptable reliability for experimental purposes. All four clusters measure the underlying construct in a reliable way (see Table 3). This indicates that the questions within each cluster tap the same attitude. This also implies that clusters of questions can indeed be analyzed, rather than separate questions.[5]

If this survey measured attitudes toward smoking policies in a valid way, it is to be expected that each cluster of questions can distinguish between smokers and non-smokers. At the time of administration, smokers and non-smokers debated heavily on such issues as the new smoking policies (see Table 2), which would prevent smoking in the hospitality industry—something that has always been permitted in the Netherlands. Results show that smokers hold more negative opinions toward these new measures to restrict smoking, $\chi^2(1) = 9.92$, $p < .01$; as well as toward other measures to prevent smoking, $\chi^2(1) = 4.58$, $p = .03$. Moreover, smokers are less in favor of governmental interference in issues concerning smoking, $\chi^2(1) = 5.61$, $p = .02$. However, compared to non-smokers, smokers are more willing to take measures to protect the weak in society against their smoke, $\chi^2(1) = 38.39$, $p < .001$. All in all, these results confirm our own intuitions and can, hence, be regarded as a validity check.

Another prerequisite that needs to be met before the answers and the cognitive processes for positive and negative questions can be compared is that the experimental groups are equal. Respondents in survey Versions 1 and 2 were found to be comparable with respect to their age, $\chi^2(3, N = 56) = 5.09$, $p = .17$; gender, $\chi^2(1, N = 56) = 3.31$, $p = .07$; educational level, $\chi^2(3, N = 56) = 4.08$, $p = .25$; and smoking behavior, $\chi^2(3, N = 56) = 2.23$, $p = .53$. Moreover, they responded similarly to the filler questions: for the yes or no questions, $\chi^2(1) = 0.04$, $p = .53$; for the agree–disagree questions, $\chi^2(1) =$

[5]A third cluster of 2-point scale questions on *politics in general* was also administered, but the reliability of this cluster was too low. Hence, the measurement error for this cluster is unacceptably large. We decided not to report the results of this cluster. This explains why not all manipulated questions that were originally in the survey were taken into account in the analysis.

0.06, $p = .80$; and for the often–never questions, $\chi^2(1) = 2.51$, $p = .11$; and they took the same amount of time to process the filler questions: for the time spent on the question, $\chi^2(1) = 0.05$, $p = .83$; and for the time spent on the answers, $\chi^2(1) = 0.64$, $p = .43$. So, randomization problems or sampling error cannot account for differences between positive and negative questions in mean answers, processing time, or rereading occurrence.

Answers to Survey Questions: Response Effects for Contrastive Questions[6]

To establish for which clusters of questions response effects occur, the mean answers for positive and negative questions were compared in each cluster. The results are shown in Table 4.

Response effects for the 2-point scale questions. For both clusters of 2-point scale questions, an overall response effect in the expected direction was observed: Respondents are more likely to *disagree* with *negative* questions than to *agree* with equivalent *positive* ones. For the cluster with questions on *new smoking policies*, this effect can be classified as small, $\chi^2(1) = 4.25$, $p = .04$ (Cohen's $d = 0.28$). Although the mean difference in answers can be classified as small in relation to the standard deviation, the direction of the wording effect is in line with previous studies, and the mean difference is quite substantial in terms of percentages. Respondents are 4% more likely to answer *no* to a *negative* question than to answer *yes* to a *positive* question. For the cluster on *other measures to prevent smoking*, the difference in mean answers is 12%. This difference is large as compared to the standard deviation, $\chi^2(1) = 6.05$, $p = 0.01$ (Cohen's $d = 1.16$).

Response effects for 5-point scale questions. The cluster on *the role of politics in the prevention of smoking* showed an overall response effect in the expected direction, $\chi^2(1) = 5.41$, $p = .02$. The mean difference between positive and negative questions is 0.2 points on a scale ranging from 1 to 5. This difference can be classified as small in relation to the standard deviation (Cohen's $d = 0.33$), but again reflects a substantial answering difference of about 4% of the scale. For the other cluster with 5-point scale questions on *how to protect the weak against smoke*, no overall response effect was observed, $\chi^2(1) = 0.29$, $p = .59$.

All in all, response effects are observed for three out of the four clusters of questions, with usual variation in effect size. This means that a necessary

[6]All results discussed in this section are observed for both smokers and non-smokers.

TABLE 4
Wording Effects per Cluster of Questions

Wording	*M*	S^2 *Questions*	S^2 *Persons*	S^2 *Residual*[a]	*Effect Size*[b]
New smoking policies (2-point scale)					
Positive	0.78 (1.27)*	1.29[c]	1.25	—[d]	0.28
Negative	0.82 (1.55)		0.71		
Other measures to prevent smoking (2-point scale)					
Positive	0.61 (0.46)*	1.48	0.11	—	1.16
Negative	0.73 (0.98)		0.29		
The role of politics in smoke prevention (5-point scale)					
Positive	3.69*	0.39	0.30	0.94	0.33
Negative	3.89		0.44	0.93	
How to protect the weak against smoke (5-point scale)					
Positive	2.91	0.63	0.23	1.28	—
Negative	2.88		0.30	1.28	

Note. The mean score represents a number between zero and one for the 2-point scale questions and a number between one and five for the 5-point scale questions. In all cases, a higher mean score represents a more positive opinion toward the attitude object; thus, more agreement with the positive question and more disagreement with the negative question. For the sake of convenience, the answers for the binomial 2-point scale questions are given in proportions and in the logits used for the analysis (between parentheses).

[a]The residual variance consists of interaction variance—that is, one person agrees more with one item than with another—as well as random error variance. [b]The size of an effect is often classified in relation to the standard deviation (Cohen, 1988). The effect size we report here is based on the between-person standard deviation. [c]The question variance is estimated once, for positive and negative questions together. This is a constraint of the model. [d]In logit models, the residual variance cannot be estimated separately from the mean score (Goldstein, 2003; Snijders & Bosker, 1999).

*$p < .05$.

condition is met for investigating the cognitive processes underlying question answering. Because one cluster failed to show an overall effect, this enables us to investigate whether cognitive differences are always shown or only for clusters with a significant wording effect.

Processing Time

The results of the analyses on the processing time are shown in Table 5. The average fixation times are provided in milliseconds and in logs.

Processing time for 2-point scale questions. For both clusters of 2-point scale questions, and for both the single-reading trials and the rereading trials, none of the four time measures showed a significant difference between positive and negative questions: all χ^2s(1) < 3.84, $p > .05$. This implies that, based on

TABLE 5
Average Total Fixation Times in Milliseconds (Estimated In (times) Between Parentheses)

Variable	*First Pass Reading Time for the Question*	*First Pass Reading Time for the Answers*	*Remaining Total Fixation Time Question*	*Remaining Total Fixation Time Answers*
New smoking policies (2-point scale)				
Rereading trials				
Mean positive	2,392 (7.78)	508 (6.23)	416 (6.03)	433 (6.07)
Mean negative	2,416 (7.79)	459 (6.13)	498 (6.21)	424 (6.05)
Single-reading trials				
Mean positive	Similar to rereading trials	750 (6.62)*	Redundant	Redundant
Mean negative	Similar to rereading trials	757 (6.63)*	Redundant	Redundant
S^2 questions[a]	0.01	0.01	0.01	0.01
S^2 persons	0.09	0.03	0.24	0.18
S^2 residual[b]	0.26	0.18	1.28	0.85
Other measures to prevent smoking (2-point scale)				
Rereading trials				
Mean positive	2,101 (7.65)	469 (6.15)	944 (6.85)	437 (6.08)
Mean negative	2,165 (7.68)	478 (6.17)	639 (6.46)	567 (6.34)
Single-reading trials				
Mean positive	Similar to rereading trials	796 (6.68)*	Redundant	Redundant
Mean negative	Similar to rereading trials	685 (6.53)*	Redundant	Redundant
S^2 questions	0.02	0.02	0.02	0.02
S^2 persons	0.13	0.05	0.05	0.10
S^2 residual	0.23	0.29	1.30	0.75
The role of politics in smoke prevention (5-point scale)				
Rereading trials				
Mean positive	1,939 (7.57)	578 (6.36)	508 (6.23)	645 (6.47)
Mean negative	2,276 (7.73)	539 (6.29)	614 (6.42)	728 (6.59)
Single-reading trials				
Mean positive	2,345 (7.76)*	1,064 (6.97)*	Redundant	Redundant
Mean negative	Similar to rereading trials	1,054 (6.96)*	Redundant	Redundant
S^2 questions	0.01	0.01	0.01	0.01
S^2 persons	0.09	0.04	0.16	0.09
S^2 residual	0.28	0.35	1.13	0.68
How to protect the weak against smoke (5-point scale)				
Rereading trials				
Mean positive	1,901 (7.55)	679 (6.52)	534 (6.28)	633 (6.45)
Mean negative	1,826 (7.51)	596 (6.39)	493 (6.20)	614 (6.42)
Single-reading trials				
Mean positive	Similar to rereading trials	1,176 (7.07)*	Redundant	Redundant
Mean negative	Similar to rereading trials	1,188 (7.08)*	Redundant	Redundant
S^2 questions	0.03	0.03	0.03	0.03
S^2 persons	0.09	0.09	0.09	0.09
S^2 residual	0.22	0.39	1.33	0.84

Note. For the sake of convenience, the average processing times are given in milliseconds and in the logs used for the analysis (between parentheses). Of course, the variances are only given in logs.

[a]The question variance is constrained to be equal for positive and negative questions. [b]The residual variance consists of interaction variance, as well as random error variance.

*$p < .05$ significant difference between the rereading trials and the single-reading trials.

the time measures, there is no reason to assume that either the comprehension-retrieval process or the mapping process is affected by the choice of wording.

The first pass reading time for the question and the first pass reading time for the answers can also be compared between the single-reading trials and the rereading trials. For both clusters of questions, the first reading of the question takes the same amount of time for single-reading trials and rereading trials, $\chi^2(1) < 3.84$, $p > .05$. However, the first reading of the answers always takes longer for the single-reading trials than for the rereading trials, $\chi^2(1) > 3.84$, $p < .05$.

To give a general impression of the time course for the question-answering process for 2-point scale questions, one of the clusters (on *new smoking policies*) is used as an exemplar case. For the single-reading trials, the reading of the question takes about 2,400 ms, and the reading of the answering options takes about 750 ms. This means that the entire question-answering process takes about 3,150 ms, of which the first pass reading time for the question takes up about 75%.

For the rereading trials, the first pass reading time of the question takes about 2,400 ms, the first reading of the answers takes about 500 ms, and thereafter the question and the answers are reread for about 420 and 430 ms, respectively. This means that the entire question-answering process unfolds in about 3,750 ms, of which the largest part (64%) is spent on the first reading of the question. Note that results are comparable for the other cluster of 2-point scale questions (on *other smoking policies*).

Processing time for 5-point scale questions. Table 5 shows that for both clusters of 5-point scale questions, none of the four processing measures shows a difference between positive and negative questions: for all time measures, $\chi^2(1) < 3.84$, $p > .05$. These results apply to both the single-reading trials and the rereading trials. Hence, based on the processing measures, there is no reason to assume that either the comprehension-retrieval process or the mapping process is affected by the choice of wording.

For the single-reading trials, the question is read for about 1,900 ms, and thereafter the answers are read for 1,180 ms. This means that the entire question-answering process unfolds in about 3,080 ms, of which the largest part is spent on the reading of the question (62%). These results are based on the cluster on *how to protect the weak against smoke*, but the results for the other cluster of 5-point scale questions are comparable.

For rereading trials, the first reading of the question takes about 1,900 ms. This is similar to the single-reading trials. The first reading of the answering options takes about 680 ms. This is significantly less time than for the single-reading trials. The rereading of the question and the answers takes 530 and 630 ms, respectively. This means that, for the rereading trials, the question-

answering process unfolds in about 3,740 ms, of which about one half of the time involves the first reading of the question. Again, results for the other cluster of 5-point scale questions are largely comparable.

Rereading Occurrence

Table 6 shows the average rereading occurrence for positive and negative questions.

Rereading occurrence for 2-point scale questions. Results show that negative questions and answers to negative questions are reread more often

TABLE 6
Average Occurrence of Rereading (Logits Between Parentheses)

Region	*Wording*	*Occurrence of Rereading*	S^2 *Questions*	S^2 *Persons*	*Effect Size*[a]
New smoking policies (2-point scale)					
Question	Positive	0.39 (−0.44)*,+	0.00	0.18	1.21
	Negative	0.52 (0.07)			
Answer	Positive	0.22 (−1.29)*,+	0.00	0.07	2.31
	Negative	0.34 (−0.69)			
Other measures to prevent smoking (2-point scale)					
Question	Positive	0.41 (−0.35)**,+	0.00	0.81	1.10
	Negative	0.66 (0.65)			
Answer	Positive	0.31 (−0.78)+	0.00	0.54	—
	Negative	0.43 (−0.29)			
The role of politics in smoke prevention (5-point scale)					
Question	Positive	0.51 (0.02)	0.85	0.00	—
	Negative	0.53 (0.13)			
Answer	Positive	0.35 (−0.63)	0.85	0.00	—
	Negative	0.40 (−0.41)			
How to protect the weak against smoke (5-point scale)					
Question	Positive	0.44 (−0.24)*,+	0.16	0.46	0.56
	Negative	0.54 (0.18)			
Answer	Positive	0.33 (−0.71)+	0.16	0.48	—
	Negative	0.39 (−0.43)			

Note. For the sake of convenience, the average number of times a respondent rereads the question or the answers are given in proportions and in the logits used for the analysis (between parentheses). Of course, the question variance, as well as the person variance, are only given is logits.

[a]The size of an effect is often classified in relation to the standard deviation (Cohen, 1988). The effect size we report here is based on the between-person and the between-question standard deviation; only effect sizes of individual effects are reported.

*$p < .05$. **$p < .01$. +$p < .05$ for the difference between positive and negative question and answers together in a combined analysis.

compared to positive questions and answers to positive questions for both clusters of 2-point scale questions. For *new smoking policies*, positive questions are reread in 39% of the cases, whereas negative questions are reread 52% of the time, $\chi^2(1) = 4.43$, $p = .04$. This difference can be classified as large (Cohen's $d = 1.21$). In addition, 22% of the positive trials involve rereading the answers versus 34% of the negative ones, $\chi^2(1) = 4.93$, $p = .03$. This also represents a large effect (Cohen's $d = 2.31$).

For the cluster on *other measures to prevent smoking*, positive questions are reread in 41% of the cases, whereas negative questions are reread 61% of the time, $\chi^2(1) = 10.09$, $p < .001$. This is a large effect (Cohen's $d = 1.10$). The answering options are reread in 31% of the positive trials and 43% of the negative trials. By itself, this difference is not significant, $\chi^2(1) = 2.71$, $p = .10$; but, in a combined analysis, results show that rereading a question or an answer happens more often for negative questions, $\chi^2(1) = 7.86$, $p < .01$.

The estimated variances help to further interpret differences in the occurrence of rereading: Is rereading a characteristic of certain questions, of certain persons, or mainly an interactional phenomenon? Table 6 shows there is no systematic between-question variance for both clusters of questions. Hence, each question is reread roughly just as often as the others. Rereading does appear to be a strategy certain persons adopt more often than other persons; for both clusters of 2-point scale questions, between-person variance can be shown. Hence, some respondents are more likely than others to reread a question or an answer to a question. However, most of all, rereading is something one person does for one question and another person for another question.[7]

Rereading occurrence for 5-point scale questions. Differences between positive and negative questions in rereading occurrence can be shown for only

[7]The residual variances are not given in Table 6 because they cannot be estimated separately from the mean occurrences of rereading. Yet, when the mean scores are known, the residual variance can be approximated using the formula $p \times (1 - p)$. In this formula, p represents the estimated probability. Hence, to approximate the residual variance for the occurrence of rereading for positive questions for the cluster on *new smoking policies* (see Table 6), we get $.39 \times (1 - .39) = .24$. This is the residual variance on a proportional scale ranging from zero to one. Hence, the residual standard deviation is $\sqrt{(0.24)}$, which gives 0.49. This means that in an 80% confidence interval, the occurrence of rereading for positive questions ranges from $.39 - (1.28 \times 0.49)$ to $.39 + (1.28 \times 0.49)$. Hence, roughly speaking, the occurrence of rereading ranges between zero and one. This variation is always larger than the variation due to between-person differences. For the cluster on new smoking policies, the between-person variance for rereading positive questions is 0.18 on a logit scale. This means that the between-person standard deviation is $\sqrt{(0.18)}$, which gives 0.42. In an 80% confidence interval, the between-person variation for rereading positive questions lies within $-.44 - (1.28 \times .42)$ and $-.44 + (1.28 \times .42)$ on a logit scale. If we transform this logit confidence interval ranging from -0.98 to 0.10 to a proportional scale, rereading occurs between 27% and 54% of the time.

one of the clusters with 5-point scale questions. These differences concern the cluster on *how to protect the weak against smoke.* This is the cluster with 5-point scale questions that *did not* show a significant difference between positive and negative questions in mean answers (see the Answers to Survey Questions section). Yet, for this cluster, negative questions are reread in 54% of the cases, whereas positive questions are reread only 44% of the time, $\chi^2(1) = 4.36$, $p = .04$. This effect is medium in size (Cohen's $d = 0.56$). Answers to negative questions are reread 39% of the time versus 33% of the positive cases. By itself, this difference is not significant, $\chi^2(1) = 1.76$, $p = .18$. In a combined analysis, however, results show that both the question and the answers are reread more frequently for negative questions, $\chi^2(1) = 4.85$, $p = .03$. Hence, these results suggest that differences in the occurrence of rereading can be shown irrespective of differences in the mean answers of positive and negative questions.

For the cluster of 5-point scale questions that *did* show an answering difference (see the Answers to Survey Questions section), *no* differences in the occurrence of rereading can be shown. For this cluster (on *the role of politics in the prevention of smoking*), rereading a question happens in 53% of the negative cases versus 51% of the positive ones, $\chi^2(1) = 0.23$, $p = .63$. Rereading an answer occurs in about 40% of the negative cases versus 35% of the positive ones, $\chi^2(1) = 0.88$, $p = .35$. Also, in a combined analysis, these differences fail to reach significance, $\chi^2(1) = 0.93$, $p = .33$. These latter results are unexpected. However, they might be related to the fact that, for this cluster, a relatively large between-question variance is observed. A large question variance implies that certain questions are reread particularly often, irrespective of the positive or negative question wording. In addition, as a large question variance results in a large total variance, large differences in rereading occurrence are needed to obtain a significant difference.

CONCLUSION AND DISCUSSION

The survey is a device often used to measure people's opinions and attitudes. Yet, a large body of research shows that the question polarity, a seemingly unimportant linguistic question characteristic, influences how respondents express their attitudes: Respondents are more inclined to disagree with negative questions than to agree with equivalent ones (e.g., Bischop et al., 1988; Holleman, 2000; Kamoen et al., 2007; Waterplas et al., 1988). This study aimed to explain this effect by relating it to the cognitive processes underlying question answering. Response effects and cognitive processes were investigated for a large set of questions that can be grouped into two clusters of 2-point scale questions and two clusters of 5-point scale questions.

Results show a response effect for both clusters of 2-point scale questions: Respondents are more likely to disagree with negative questions than to agree with equivalent positive ones. One of these effects is large compared to the standard deviation; the other effect is small in size. A small response effect in the expected direction was also observed for one of the clusters of 5-point scale questions; the other 5-point scale cluster showed no effect. These results are in line with previous studies (e.g., Bischop et al., 1988; Holleman, 2000; Kamoen et al., 2007; Waterplas et al., 1988). First, in line with earlier research, respondents generally express their opinions more positively when the question is worded negatively. Second, as in previous studies, large variation in the size and occurrence of response effects can be shown.

Variation in wording effects is usually attributed to interaction of the effect of question wording with a broad range of experimental characteristics and question characteristics, such as the type of administration, the type of word pair used, and the average opinion of respondents about the topic of the question (e.g., Bischop et al., 1988; Holleman, 2000; Kamoen et al., 2007). Because there are so many possible causes for variation in survey wording effects, it is impossible to isolate one post hoc explanation for why one of the clusters of questions in this study failed to show an overall wording effect or why some effects were larger than others. More important, three clusters showed an overall response effect when pooling over questions and, hence, by pooling over this large variation. Although these effects were not always large in comparison to the standard deviation, they were always substantial in terms of percentages. Therefore, an explanation of survey wording effects is called for, and insight into the validity of survey questions is required.

Analyses of the respondents' eye-movement patterns show that survey questions are usually answered in two distinct ways. In some trials—that is, for some Respondent × Item combinations—the question and the answering options are read once, and then a second time before an answer is given (rereading trials); in other trials, the question and the answering options are read only once (single-reading trials). For the single-reading trials, answering an attitude question takes about 3 s; whereas for the rereading trials, the question-answering process lasts roughly 800 ms longer (see Tourangeau et al., 2000; for comparable results, see Chessa & Holleman, 2007).

Comparisons of question-answering processes for positive and negative questions indicate that, for all four clusters of questions, the initial time needed to read the question and the answers is the same for positive and negative questions; these observations apply to both the rereading trials and the single-reading trials. In addition, once rereading a question or an answer occurs, this process also takes the same amount of time for positive and negative questions. However, the questions and response options of negative questions are reread more frequently than positive ones: For three out of the four clusters of questions, results show a

difference in rereading occurrence. These effects are generally medium or large in size.

What do these results mean in terms of cognitive processes? Is the comprehension-retrieval process or the mapping process affected by the choice of wording? In our view, our results seem to support the mapping hypothesis: Respondents experience difficulty more often when mapping their attitude to the response scale for a negative question than for a positive one; that is why rereading occurs more often for negative questions than for positive ones (cf. Chessa & Holleman, 2007; Holleman, 1999a). In our view, the alternative cognitive load account (Clark, 1976), which predicts that the comprehension-retrieval stage is affected by the choice of wording, does not provide an adequate explanation for the results obtained in this study. Therefore, in the following, we present three arguments that clarify why the mapping interpretation seems preferable.

First, the first pass reading time for the question never indicates a difference between positive and negative questions. As this measure provides an unbiased measure of the comprehension-retrieval process, the systematic non-occurrence of an effect provides a strong argument against the cognitive load account.

Second, the three other measures of processing time are both unlikely to reflect mainly comprehension-retrieval processes, and fail to show a difference between positive and negative questions. Initially, we assumed that after the first reading of the question, additional comprehension-retrieval processes would probably not be necessary because the survey contains short and easy questions, and these questions would be read by relatively skilled readers. Our results support this assumption, as the first pass reading time for the question always constitutes a substantial part of the question-answering process. In addition, our data show that the first pass reading time for the question always takes the same length of time for single-reading trials and rereading trials. This can be seen as an argument that rereading is, in general, not a sign that the question was read inadequately the first time. If, despite these arguments, one would still consider all or some time measures after the respondent's initial reading of the question to reflect delayed comprehension-retrieval processes, our results provide no evidence for the predicted inherent difficulty of negative questions. The increased difficulty associated with the interpretation of negative terms would then be expected to become apparent in longer processing times for negative questions, but a difference between positive and negative questions could not be shown for any of the measures of processing time in this study.

Third, differences in the occurrence of rereading are unlikely to reflect the kind of comprehension-retrieval effects predicted by the cognitive load account. First of all, the cognitive load account predicts differences in processing time and no differences in the occurrence of certain processes. Of course, in some sense, differences in rereading occurrence can be interpreted as differences in rereading time: If rereading occurs more frequently for negative questions, the

average rereading time will be higher for negative questions when pooling over all trials—that is, both single-reading trials and rereading trials.[8] However, even if we consider the differences in rereading occurrence as alternative measures of rereading time, the cognitive load account cannot fully account for the results. If the differences in rereading occurrence would be due to increased cognitive load, we would have expected rereading to occur more frequently for certain questions (e.g., the more difficult questions) or for certain respondents (e.g., the less skilled respondents). However, results show that whether rereading occurs is an interactional phenomenon: One respondent rereads one question and another respondent rereads another question.

All in all, based on the results of this study, there are good reasons to prefer the mapping account over the cognitive load account. However, caution should be taken when concluding that the comprehension-retrieval stage is not affected at all by the choice of wording. One reason is that, in this study, the only unbiased measure for the comprehension-retrieval process, the first pass reading time of the question, provides a very broad measure. A possible effect on a smaller region, such as the manipulated word alone, may, therefore, have gone unnoticed. It would be interesting to define more fine-grained regions to evaluate the comprehension-retrieval stage for contrastive questions. Unfortunately, these data did not enable us to investigate smaller regions per cluster of questions because, for each individual survey question, the critical regions had a different position in the question.

The conclusions drawn from this study can be generalized to 2-point scale questions because both clusters showed differences in the occurrence of rereading. As significant answering differences were also shown for both of these clusters, there seems to be a direct link between answering differences and the underlying question-answering process. For the 5-point scale questions, results are equivocal. The cluster that *did not* show an answering difference *did* show differences in the occurrence of rereading. This seems to imply that the mapping process always takes more time for negative questions, irrespective of whether answering differences can be shown. However, the cluster with 5-point scale questions that *did* show a significant answering difference *did not* show a difference in rereading occurrence. A closer inspection of the results for this cluster reveals how this unexpected observation can be explained. For this cluster, there are some questions that are reread particularly often, irrespective of the question wording. For example, this concerns the question, "I consider it to be a good thing if the government interferes with the private life

[8]We performed these analyses and indeed found an effect on the rereading time for the question and the answers when pooling over all trials. These analyses, however, produced uninformative average rereading times and large variances because of the inclusion of zero observations.

of its citizens." Respondents probably experience mapping difficulties for this question because it is formulated in terms that are too general. Such a poorly worded question, therefore, causes smaller average differences between positive and negative questions in rereading occurrence. In addition, the large question variance caused by the mapping difficulties all respondents experience increases the total variance; hence, larger differences in the occurrence of rereading are required to obtain a significant difference. All in all, this implies that mapping differences probably only occur for 5-point scale questions when well-worded questions are used, irrespective of whether a significant response effect occurs for a certain cluster of questions. To obtain a better understanding of the extent to which the response scale influences the occurrence of rereading, it is necessary to investigate the question-answering process for exactly the same questions in versions with different response scales.

Future research should also investigate to what extent the results of this study can be generalized to populations other than highly educated, mainly female respondents. As gender has been shown not to influence cognitive processes underlying question answering (Holbrook, Cho, & Johnson, 2006), results can probably be generalized to both men and women. Educational level has previously been shown to affect the size of response effects for contrastive questions, as larger response effects were found for lower-educated respondents (Narayan & Krosnick, 1996). This might result in larger differences in the underlying question-answering processes for this group.

Despite these generalizability issues, results of this study increase the likelihood that only the mapping process is affected by the choice for a positive or a negative wording of a survey question. Thus, it follows that response effects occur because the answering options to contrastive questions are not merely simple categories, but are interpreted in relation to the evaluative term in the question (cf. Holleman, 2000). This also suggests that, although the answers to positive and negative questions differ, contrastive questions are equally valid. An important step to further develop this cognitive explanation of survey response effects is to investigate the cognitive processes underlying question answering in an interactional design, taking not only the question polarity but also the polarity of the answer into account (cf. Chessa & Holleman, 2007). This way, the effects of negativity and denial can be investigated at the same time.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and the editor of *Discourse Processes* for their detailed comments. Their feedback has helped to significantly improve the quality of the manuscript.

REFERENCES

Ayidiya, S. A., & McClendon, M. J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229–247.

Bassili, J. N. (1996). The how and why of response latency measurement in surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 319–346). San Francisco, CA: Jossey-Bass.

Bassili, J. N., & Scott, S. B. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*, 390–399.

Belli, R. F. (2005). Editorial. Announcing a special issue on cognitive aspects of survey methodology. *Applied Cognitive Psychology, 19*, 245–247.

Bischop, G., Hippler, H.-J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 273–282). New York, NY: Wiley.

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21*, 203–225.

Clark, H. H. (1976). *Semantics and comprehension*. The Hague, Netherlands: Mouton.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in Web surveys. *Public Opinion Quarterly, 71*, 623–634.

Dillman, D. A. (2000). *Mail and Internet Surveys. The tailored design method.* New York, NY: Wiley.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Review of personality and social psychology: Vol. 11. Research methods in personality and social psychology* (pp. 74–97). Newbury Park, CA: Sage.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data. New insights on response effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72*, 892–913.

Glendall, P., & Hoek, J. (1990). A question of wording. *Marketing Bulletin, 1*, 25–36.

Goldstein, H. (2003). *Multilevel statistical models.* London, UK: Edward Arnold.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A Web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*, 3–22.

Hippler, H.-J., & Schwarz, N. (1986). Not forbidding isn't allowing: The cognitive basis of the forbid/allow asymmetry. *Public Opinion Quarterly, 50*, 87–96.

Hippler, H.-J., & Schwarz, N. (1987). Response effects in surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 102–122). New York, NY: Springer-Verlag.

Holbrook, A., Cho, Y. I., & Johnson, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly, 70*, 565–595.

Holleman, B. C. (1999a). The nature of the forbid/allow asymmetry: Two correlational studies. *Sociological Methods & Research, 28*, 209–244.

Holleman, B. C. (1999b). Response effects in survey research: Using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics, 6*, 29–40.

Holleman, B. C. (2000). *The forbid/allow asymmetry. On the cognitive mechanisms underlying response effects in surveys*. Amsterdam, the Netherlands: Rodopi.

Holleman, B. C., & Murre, J. M. J. (2008). Getting from neuron to checkmark. Models and methods in cognitive survey research. *Applied Cognitive Psychology, 22*, 709–732.

Javeline, D. (1999). Response effects in polite cultures. A test of acquiescence in Kazakhstan. *Public Opinion Quarterly, 63*, 1–28.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

Kamoen, N., Holleman, B. C., & Van den Bergh, H. H. (2007). Hoe makkelijk is een niet moeilijke tekst? Een meta-analyse naar het effect van vraagformulering in tekstevaluatieonderzoek [How easy is a text that is not difficult? A meta-analysis about the effect of question wording in text evaluation research]. *Tijdschrift voor Taalbeheersing, 29*, 314–332.

Krosnick, J., & Schuman, H. (1988). Attitude intensity, importance, and centrality and susceptibility to response effects. *Journal of Personality & Social Psychology, 54*, 940–952.

Loosveldt, G. (1997). Interaction characteristics in some question wording experiments. *Bulletin de Méthodologie Sociologique, 56*, 20–31.

Molenaar, N. J. (1982). Response-effects of "formal" characteristics of questions. In W. Dijkstra & J. van der Zouwen (Eds.), *Response behaviour in the survey interview* (pp. 49–89). London, UK: Academic.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58–88.

Quené, H., & Van den Bergh, H. H. (2004). On multilevel modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*, 103–121.

Quené, H., & Van den Bergh, H. H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413–442.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rugg, D. (1941). Experiments in wording questions II. *Public Opinion Quarterly, 5*, 91–92.

Schuman, H. (2008). *Method and meaning in polls and surveys.* Cambridge, MA: Harvard University Press.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys. Experiments on question form, wording and context*. New York, NY: Academic.

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

Sudman, S., & Bradburn, N. M. (1982). *Asking questions. A practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass.

Tanur, J. (1999). Looking backwards and forwards at the CASM Movement. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 13–16). New York, NY: Wiley.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*, 368–393.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretative heuristics for response scales. *Public Opinion Quarterly, 71*, 91–112.

Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299–314.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Waterplas, L., Billiet, J., & Loosveldt, G. (1988). De verbieden versus niet toelaten asymmetrie. Een stabiel formuleringseffect in survey-onderzoek? [The forbid/allow asymmetry. A stable response effect in survey research?]. *Mens en Maatschappij, 63*, 399–415.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on Web survey response times. *Applied Cognitive Psychology, 22*, 51–68.

APPENDIX A
Multilevel Models for the Answering Data

As an example of a multilevel model for the response data, the model used for establishing if response effects occur for the cluster with 5-point scale questions on *the role of politics in the prevention of smoking* (N question = 14) is discussed. This model estimates the mean answering score for positive and negative questions in this cluster, as well as question variance, between-person variance, and residual variance.

In Equation 1, $Y_{(jk)}$ is the answer of individual j ($j = 1, 2, \ldots 28$) on question k ($k = 1, 2, \ldots 14$). In addition, there are two dummies—one for positive, $\text{D\_Pos}_{(jk)}$, and one for negative, $\text{D\_Neg}_{(jk)}$ questions—which can be turned on if the observation matches the prescribed question type. Using these dummies, two means are estimated (β_1, β_2), which are allowed to vary between questions, v_{0k}; persons, u_{1j0}, u_{2j0}; and due to residual factors, $e_{1(jk)}$, $e_{2(jk)}$:

$$Y_{(jk)} = \text{D\_Pos}_{(jk)}(\beta_1 + e_{1(jk)} + u_{1j0}) + \text{D\_Neg}_{(jk)}(\beta_2 + e_{2(jk)} + u_{2j0}) + v_{0k}. \quad (1)$$

Please note that a cross-classified model is in operation, as the answers are nested both within-questions and within-subjects (Quené & Van den Bergh, 2008). All residuals are normally distributed with an expected value of zero and a variance of, respectively, $S^2_{e1(jk)}$, $S^2_{e2(jk)}$, S^2_{u1j0}, S^2_{u2j0}, and S^2_{v0k}. In addition, please note that, in this model, the question variance (S^2_{v0k}) is estimated only once for positive and negative questions and all measures of processing time together. This is a constraint of the model. Furthermore, the residual variances consist of interaction variance and of (random) error variance.

The model described in this Appendix is based on a cluster of 5-point scale questions; for the clusters with 2-point scale questions, logits were analyzed. Note that for binomial data, the residual variance cannot be separately estimated from the mean score (Goldstein, 2003; Snijders & Bosker, 1999).

APPENDIX B
Multilevel Models for the Processing Time Data

As an example of a multilevel model for analysis of the processing time, the model used for the cluster with 5-point scale questions on *the role of politics in the prevention of smoking* (N question = 14) is discussed. Eight mean processing times are estimated—one for each combination of the question polarity (positive or negative) and the specific time measure (the first pass reading time for the

question, the first pass reading time for the answering options, the remaining total fixation time for the question, and the remaining total fixation time for the answers). These means are estimated based on only those respondents who actually look back at the question or the answers for a certain item. In addition, for the first pass reading time for the question and the answers, a deviation is estimated based on those trials in which the question and the answers are eventually not reread. This results in 12 mean processing times in total. The multilevel model allows for question variance, between-person variance, and residual variance.

The model is formalized in Equation 2. In this equation, $Y_{(jk)}$ is the processing time of individual j ($j = 1, 2, \ldots 28$) on question k ($k = 1, 2, \ldots 14$). In addition, there are 12 dummies, $D_{(jk)}$, which can be turned on if the observation matches the prescribed type. These "types" represent a combination of whether, for that trial, rereading is eventually involved, of the specific time measure, and of the question polarity. In Equation 2, these types are indicated as follows: the first letter D indicates that the predictors are dummy variables; the second set of letters indicates whether the trial involves looking back at the question or the answers (LB), or not (NLB); the third set of letters indicates the processing measure: first pass reading time for the question (1st_Q), first pass reading time for the answers (1st_A), the remaining total fixation time for the question (R_Q), and the remaining total fixation time for the answers (R_A); the fourth set of letters indicates the polarity of the item, positive (POS) or negative (NEG). Using these dummies, 12 mean processing times are estimated (β_1, β_2, etc.), which are allowed to vary between questions, v_{0k}; persons, u_{1j0}, u_{2j0}, and so forth; and due to residual factors, $e_{1(jk)}$, $e_{2(jk)}$, and so forth:

$$
\begin{aligned}
Y_{(jk)} = {} & \text{D\_LB\_1st\_Q\_POS}_{(jk)}(\beta_1 + e_{1(jk)} + u_{1j0}) \\
& + \text{D\_NLB\_1st\_Q\_POS}_{(jk)}(\beta_2) \\
& + \text{D\_LB\_A\_Q\_POS}_{(jk)}(\beta_3 + e_{3(jk)} + u_{3j0}) \\
& + \text{D\_NLB\_1st\_A\_POS}_{(jk)}(\beta_4) \\
& + \text{D\_LB\_R\_Q\_POS}_{(jk)}(\beta_5 + e_{5(jk)} + u_{5j0}) \\
& + \text{D\_LB\_R\_A\_POS}_{(jk)}(\beta_6 + e_{6(jk)} + u_{6j0}) \\
& + \text{D\_LB\_1st\_Q\_NEG}_{(jk)}(\beta_7 + e_{1(jk)} + u_{1j0}) \\
& + \text{D\_NLB\_1st\_Q\_NEG}_{(jk)}(\beta_8) \\
& + \text{D\_LB\_1st\_A\_NEG}_{(jk)}(\beta_9 + e_{3(jk)} + u_{3j0})
\end{aligned}
$$

$$+ \text{D\_NLB\_1st\_A\_NEG}_{(jk)}(\beta_1 0)$$

$$+ \text{D\_LB\_R\_Q\_NEG}_{(jk)}(\beta_{11} + e_{5(jk)} + u_{5j0})$$

$$+ \text{D\_LB\_R\_A\_NEG}_{(jk)}(\beta_{12} + e_{6(jk)} + u_{6j0}) + v_{0k}. \qquad (2)$$

Please note that a cross-classified model is in operation, as the processing times are nested both within-questions and within-subjects (Quené & Van den Bergh, 2008). All residuals are normally distributed with an expected value of zero and a variance of, respectively, $S^2_{e1(jk)}$, ..., S^2_{u1j0}, ..., S^2_{v0k}. In addition, note that, in this model, the question variance, S^2_{v0k}, is estimated only once for positive and negative questions together. This is a constraint of the model. Furthermore, the residual variances consist of interaction variance and of (random) error variance.

APPENDIX C
Multilevel Models for the Dichotomous Registration of Rereading

As an example of a multilevel model for analyzing the rereading occurrence, the model used for the cluster with 5-point scale questions on *the role of politics in the prevention of smoking* (N question = 14) is discussed. In this model, the rereading occurrence is estimated for positive and negative questions and for the answers to positive and negative questions. Hence, the model estimates four rereading occurrences—one for each combination of the specific rereading measure (Question or answers), as well as the question polarity (Positive or Negative). In addition, question variances and between-person variances are allowed. The residual variance is not estimated separately because it is fixed if the mean processing occurrence is known. In other words, the residual variance cannot be estimated separately from the mean score (Goldstein, 2003; Snijders & Bosker, 1999). The residual variance can be approximated by the formula, $p \times (1 - p)$; in this formula p represents the estimated probability (in this case, the estimated occurrence of rereading). Please also note that this formula gives an approximation of the residual variance in proportions, and not on a logit scale.

In Equation 3, the model used for analyzing rereading occurrence has been formalized. In this model, $Y_{(jk)}$ indicates whether individual j (j = 1, 2, ... 28) rereads question k (k = 1, 2, ... 14). In addition, there are four dummies (D)—one for rereading the question for positive questions, $\text{Q\_Pos}_{(jk)}$; one for rereading the answers for positive questions, A_Pos; one for rereading the question for negative questions, $\text{Q\_Neg}_{(jk)}$; and one for rereading the answers for negative questions, A_Neg—which can be turned on if the observation matches the prescribed type. Using these dummies, four rereading frequencies

are estimated (β_1, β_2, etc.), which may vary between questions (v_{0k}) and persons (u_{1j0}, u_{2j0}):

$$\begin{aligned}\text{Logit}(Y_{(jk)}) = {} & \text{D\_Q\_Pos}_{(jk)}(\beta_1 + u_{1j0}) + \text{D\_A\_Pos}_{(jk)}(\beta_2 + u_{2j0}) \\ & + \text{D\_Q\_Neg}_{(jk)}(\beta_3 + u_{1j0}) + \text{D\_A\_Neg}_{(jk)}(\beta_4 + u_{2j0}) \\ & + v_{0k}. \end{aligned} \tag{3}$$

Again, a cross-classified model is in operation (Quené & Van den Bergh, 2008). All residuals are normally distributed with an expected value of zero and a variance of, respectively, S^2_{u1j0}, ..., S^2_{v0k}. Please note that the individual variances (S^2_{u1j0}, etc.) are estimated for positive and negative questions together. In addition, note that, in this model, the question variance, S^2_{v0k}, is estimated only once for positive and negative questions and for both the question and the answers. These are constraints of the model.