Survey Research Methods (2013)
Vol. 7, No. 3, pp. 181-189
ISSN 1864-3361
http://www.surveymethods.org

Positive, negative, and bipolar questions: The effect of question polarity on ratings of text readability

Naomi Kamoen
Utrecht Institute of Linguistics OTS Utrecht University and
Department of Communication and Information Sciences Tilburg University

Bregje Holleman
Utrecht Institute of Linguistics OTS
Utrecht University

Huub van den Bergh
Utrecht Institute of Linguistics OTS
Utrecht University

For decades, survey researchers have known that respondents give different answers to attitude questions worded positively ("X is good. Agree-Disagree"), negatively ("X is bad. Agree-Disagree") or on a bipolar scale ("X is bad-good"). This makes survey answers hard to interpret, especially since findings on exactly how the answers are affected are conflicting. In the current paper, we present twelve studies in which the effect of question polarity was measured for a set of thirteen contrastive adjectives. In these studies, the same adjectives were used so the generalizability of wording effects across studies could be examined for each word pair. Results show that for five of the word pairs an effect of question wording can be generalized. The direction of these effects are largely consistent: respondents give the same answers to positive and bipolar questions, but they are more likely to disagree with negative questions than to agree with positive questions or to choose the positive side of the bipolar scale. In other words, respondents express their opinions more positively when the question is worded negatively.
Keywords: positive, negative, and bipolar questions

1 Introduction

How do people judge the seriousness of the global warming problem, the adequacy of the government's smoking policies, or the quality of the national education system? Social scientific research tries to understand people's opinions and attitudes towards such issues, as these are assumed to be predictors of human behavior (e.g., Ajzen, 1988; Eagly & Chaiken, 1993; Oskamp & Schultz, 2005).

To acquire insight into people's opinions and attitudes, the survey is often used as a measurement instrument. In survey questions about opinions and attitudes, respondents are usually asked to position themselves somewhere on a continuum of a cognitive or an evaluative dimension with respect to the attitude object, such as *favourable/unfavourable*, *good/bad*, or *interesting/boring*. This continuum is often worded in one of three ways:

1. This story is interesting
   Disagree □□□□□ Agree
2. This story is uninteresting
   Disagree □□□□□ Agree
3. This story is
   uninteresting □□□□□ interesting

Naomi Kamoen, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, e-mail: n.kamoen@uu.nl

The three questions above differ with respect to question polarity. Question 1 may be referred positive question: the respondent is asked to rate to what extent an attitude object ("this story") possesses an evaluatively positive characteristic ("interestingness"). Question 2 may be referred to as a negative question, as the respondent is asked to rate the attitude object onto a negative scale ("uninterestingness"). Lastly, question 3 may be called a bipolar question, because the respondent is asked to rate the attitude object on a scale ranging from negative ("uninteresting") to positive ("interesting").

At first sight, these positive, negative and bipolar wordings seem to represent identical attitude questions: a respondent agreeing with a positive question is expected to disagree with a negative question and to choose the positive side of the bipolar scale. But is it indeed the case that similar responses are given to positive, negative, and bipolar questions? A definitive answer to this question is still lacking, because experimental studies report equivocal results about whether and how survey answers are affected. Therefore, the aim of the current study is to investigate for a set of contrastive questions whether a wording effect can be generalized across studies. Insight into the generalizability of wording effects is necessary for survey practice, as it provides information about how large answering differences are expected to be in a new study. In addition, an investigation of the generalizability is a necessary first step to decide whether future research into the validity of contrastive questions is warranted.

1.1 Positive versus negative questions

In previous research on the effect of question polarity, comparisons have been made between positive and negative questions or between unipolar and bipolar questions. Perhaps the most cited study in which positive and negative questions have been compared, was conducted in the 1940s. Donald Rugg (1941) asked one group of respondents the positive question "Do you think the government should allow public speeches against democracy? Yes/No", and another group of respondents the negative question "Do you think the government should forbid public speeches against democracy? Yes/No". Results of this study showed that respondents are 21% more likely to answer "no" to the negative *forbid* question, than to answer "yes" to the equivalent positive *allow* question.

After Rugg's study, others targeted at replicating the so-called *forbid/allow* asymmetry (e.g., Bishop, Hippler, Schwarz, & Strack, 1988; Glendall & Hoek, 1990; Hippler & Schwarz, 1986; Krosnick & Schuman, 1988; Loosveldt, 1997; Narayan & Krosnick, 1996; Schuman & Presser, 1981; Waterplas, Billiet, & Loosveldt, 1988). The main conclusion to be drawn from these studies is that the occurrence, size, and direction of the effect vary from question to question and from study to study. Apparently, whether and how the wording of the question affects survey answers, depends on all sorts of experimental characteristics, such as the mode of administration and the sample of respondents. Despite this variation, a meta-analysis by Holleman (1999a) showed that the *forbid/allow* asymmetry can be generalized beyond the question level. In general, respondents are more likely to answer "no" to negative *forbid* questions, than to answer "yes" to equivalent positive *allow* questions. In other words, there is a nay-saying bias: respondents are disproportionally more likely to answer "no" to both wordings. The average size of this effect is large, but the standard deviation of the effect is also considerable. Hence, if we would conduct a set of random *forbid/allow* experiments, most studies will show a response effect in the expected direction, but the size of the difference between "not forbid" and "yes allow" will vary greatly over studies and in some studies even response effects in the opposite direction will be observed.

For survey practice it is important to predict response effects for questions with other contrastive word pairs as well. If respondents are consistent in the way they answer positive and negative questions, a nay-saying bias similar to the effect found for *forbid* and *allow* questions would be expected. However, a yea-saying or acquiescence bias is the prevailing effect reported for contrasts other than *forbid/allow*. For example, in a study by O'Neill (1967) positive and negative comparatives were examined, such as "Advertising results in better products for the public" versus "Advertising results in poorer products for the public". For most items, he observed that respondents are more likely to agree with positive questions than to disagree with negative ones. In addition, Falthzik and Jolson (1974) compared the answers to positive questions like "Unit pricing is beneficial to a majority of consumers" with answers to questions with an explicit negation like "Unit pricing is not beneficial to a majority of consumers". For seven out of the twelve statements, respondents were more likely to agree with the positive wording than to disagree with the negative wording. Also, Javeline (1999) found evidence for yea-saying for five questions concerning political and economic issues.

The studies cited show that although a yea-saying bias is frequently reported for word pairs other than *forbid/allow*, this effect is not observed for every manipulated question in every experiment (see also Molenaar, 1982 for a summary of older work by Blankenship, 1940; Gallup, 1941; Roslow, Wallace, Wulfeck & Corby, 1940; Rugg & Cantrill, 1947; Schuman & Presser, 1977). There are at least three causes for the observed variation. First, studies differ with respect to all sorts of contextual characteristics, such as the topic of the survey, the type of answering scale, and the type of respondents. This variation may cause different results, as the effect of question wording may interact with these contextual characteristics. Second, within and between studies, different word pairs are used for the manipulations. Clearly, this may cause variation because different word pairs may sort different kind of effects. Third, definitions of what counts as a positive or negative wording are sometimes unclear, for example in cases where the object evaluated carries an evaluatively negative aspect (e.g., in Javeline's (1999) "It is a great fortune/misfortune the Sovjet Union no longer exists"). Hence, starting out from a clear definition of question polarity, the main question to be answered in the current study is whether for each word pair there is a wording effect that can be generalized beyond the large contextual variation.

1.2 Unipolar versus bipolar questions

In contrast to the large number of studies in which the answers to positive and negative questions are compared, the number of studies comparing the answers to unipolar and bipolar questions is far less. In a review article, Molenaar (1982, p. 60) draws the following conclusion about the difference between unipolar and bipolar questions: "a given specified alternative tends to be chosen more when presented alone, i.e. in an imbalanced yes/no question, than when presented along with one or more other contrasting alternatives, i.e. in a more balanced dichotomous or multiple choice question". Hence, respondents are more likely to evaluate an attitude object as *bad* when we ask: "X is bad. Yes/No" than when we ask "X is good - bad". Similarly, respondents are more likely to evaluate an attitude object as *good* when asking "X is good. Yes/No", than when asking "X is good - bad". This conclusion fits predictions for the acquiescence bias: if respondents are more likely to answer "yes" to a survey question independent of its polarity, we would expect respondents to express their opinions most positively to positive questions, and least positively to negative questions. The answers to bipolar questions would fall between those two: positive > bipolar > negative. As opposed to this line of reasoning, Menezes and Elbert (1979) show no significant differences in response distributions of twelve unipolar and bipolar questions. In addition, based on vari-

ous empirical studies, Schuman and Presser (1981) conclude with respect to the difference between unipolar and bipolar questions: "..., it appears to make little, if any, difference whether an item is formally balanced or imbalanced in the sense of adding 'oppose' or similar terms to the questions that already clearly imply the possibility of a negative alternative" (Schuman & Presser, 1981, p. 199). Thus, taking these studies as reference point, the choice for a unipolar or bipolar question seems irrelevant for the answers obtained.

All in all, the question that arises also from these studies is: why do they show such conflicting results? The three previously identified factors probably also play a role here. First, the different studies vary with respect to all kinds of contextual characteristics, which may interact with the wording effect. Second, the various word pairs that are used across the different studies may yield different kinds of response effects. Third, the definition of what counts as a positive, a negative, or a bipolar question is not always constant across studies. In addition to these previously mentioned factors, and perhaps most importantly, most studies analyze both positive and negative questions as "unipolar" questions, overlooking differences between those two and between their relation to the bipolar question format. To overcome this problem, the current research aims to compare the answers to all three question formats at the same time.

1.3 Research question and hypotheses

Previous studies report equivocal results about whether and how the answers to positive, negative and bipolar questions differ. Therefore, we will investigate in the current research whether the answers to positive, negative and bipolar questions differ when generalizing across studies. To investigate this question, we will examine the effect of question polarity for a set of word pairs in various studies. In this way we can establish for each word pair whether there is an effect of question wording that can be generalized across studies.

2 Method

2.1 Set-up and materials

Twelve split-ballot experiments were conducted to compare the answers to positive, negative, and bipolar questions. In each experiment, respondents read a text. Afterwards, they filled out a survey in which they expressed their opinions and attitudes about the quality of the text. Three versions of the survey were constructed: questions that were worded positively in version 1 were worded negatively in version 2 and were posed on a bipolar scale in version 3. Each study also included a large number of filler questions which had an identical wording across the survey versions. In each study, respondents were randomly assigned to one of the three survey versions, and they participated in only one of the twelve studies. The experiments were all administered in Dutch, and hence, to native speakers of Dutch.

The choice to measure the influence of question wording in studies assessing attitudes towards texts, was motivated from the current debate in the field of Communication studies on how attitudes towards texts can best be measured (e.g., Anderson & Fornell, 2000; Brooke, 1996; Muylle, Moenaert, & Despontin, 2004; van Schaik & Ling, 2005). In these studies several theoretical constructs are defined and each construct is measured by either a mixed set of positive and negative questions, or a full set of bipolar questions.

In the twelve experiments in the current study, the survey questions were always constructed using the guidelines given by Maes, Ummelen, and Hoeken (1996). These authors distinguish *perceived text comprehension* and *perceived attractiveness of the text* as underlying constructs of text quality. Both clusters are measured with six questions (see Table 1). Besides the twelve questions related to these constructs, a separate question was asked about the image of the sender in most of the studies (see Appendix for an overview of exactly which questions were asked in the twelve studies).

Table 1 Questions used to assess the perceived comprehension and the perceived attractiveness of the text; the original Dutch wordings used in the different studies are presented in the left column, and an English translation is given in the right column

Comprehension: The text is ...	
Ingewikkeld/Eenvoudig	Simple/Complicated
Duidelijk/Onduidelijk	Clear/Unclear
Overzichtelijk/Onoverzichtelijk	Orderly/Chaotic
Logisch/Onlogisch opgebouwd	Logically/Illogically arranged
Bondig/Omslachtig	Concise/Wordy
Makkelijk/Moeilijk	Easy/Difficult
Attractiveness: The text is ...	
Aansprekend/Afstandelijk	Appealing/Distant
Uitnodigend/Afhoudend	Inviting/Reluctant
Boeiend/Saai	Fascinating/Boring
Persoonlijk/Onpersoonlijk	Personal/Impersonal
Afwisselend/Eentonig	Varied/Monotonous
Interessant/Oninteressant	Interesting/Uninteresting
Separate question: The sender is ...	
Deskundig/Ondeskundig	Expert/Amateur

Important to note is that the same word pairs were used for the manipulations in each study, and in addition, that all manipulated questions in all of the twelve studies are established conform the definition of positive, negative and bipolar questions set in the introduction. This means that the evaluative terms in the questions are always neutral ("the text"), free of any evaluatively positive or negative aspect. What counts as a positive, negative or bipolar wording is solely determined by the evaluative polarity (see Hamilton and Deese, 1971) of the manipulated evaluative term in the question.

While the same word pairs were used for the manipulations in each study, and while all surveys were self-administered, there was variation across studies in a large number of non-manipulated characteristics. This variation

can be considered random variation; the 12 experiments were conducted by MA-students in an advanced course on survey design, and they independently made the various choices about the design of the survey. Therefore, some studies phrased the questions in an objective way ("The text is ..."), whereas others used a more subjective phrasing ("I think the text is ..."). Moreover, the response scales varied with respect to the number of scalar points used: two studies used 5-points scales; the others used 7-points scales. In addition, the manipulated questions had a different position in the survey; some studies first measured *perceived text comprehension*, others first measured *perceived attractiveness*, and other studies mixed the questions about these two constructs. We refer to Tables 2 and 3 for a more detailed overview of several design characteristics of the various studies.

Importantly, the variation in contextual characteristics provides an opportunity to investigate the generalizability of wording effects. As most of the characteristics that were varied between studies by itself are shown to affect survey answers (e.g., Tourangeau, Rips, & Rasinski, 2000; Weijters, Cabooter, & Schillewaert, 2010), the variation in these characteristics will create a considerable variance; hence, this variation causes a strict test for deciding whether wording effects can be generalized.

2.2 Respondents

In each of the twelve experiments, about 200 respondents took part. They were always the target group of the text. As the type of text read varied across studies, there was also variation in the type of respondents participating. For example, in one of the studies, members of the Dutch Epilepsy Foundation, mainly elderly people, rated the quality of the institution's yearly magazine. In another study, the quality of a new teaching method was evaluated among high school pupils. In a third study, a heterogeneous sample of adults rated the quality of a persuasive text about organ donation. As respondent characteristics may also affect the size and occurrence of wording effects (e.g., Narayan & Krosnick, 1996; Schuman & Presser, 1981), this variation again allows for an opportunity to generalize: do wording effects for contrastive questions arise in spite of the variation between studies in the type of respondents taking part in the survey? For a more detailed description of the type of respondents participating in each study, we refer to Table 2.

2.3 Analysis

To examine whether effects of question wording for each word pair can be generalized beyond the study level, multilevel analysis was used. This is a statistical technique capable of separately estimating the variance at the different levels of the sampling hierarchy. For the data obtained in the current study it is important to estimate the variance at different levels of the sampling hierarchy: there may be an overall effect of question wording for a specific word pair, but the size of this effect may vary between studies. By estimating the between study variance separately from the error variance, the chance of making a type-1 error (rejecting H0 while H0 is true) is properly controlled for.

For each word pair, a separate multi-level model was constructed in which three mean scores are estimated: one for positive, one for negative, and one for bipolar questions. These scores may vary between studies, as well as between persons within studies. Equation 1 formalizes the model we used. In this equation, Y_{ij} is the answer of person $i, i = 1 \ldots I_j$ in study $j, j = 1 \ldots 12$. Moreover, there are three dummies (D) one for positive (D_Positive), one for negative (D_Negative), and one for bipolar questions (D_Bipolar), which can be turned on if the observation matches the prescribed type. Using these dummies, three mean scores are estimated ($\beta_1, \beta_2, \beta_3$), which may vary between persons ($e_{1ij}, e_{2ij}, e_{3ij}$), and studies ($u_{10j}, u_{20j}, u_{30j}$). Regression weights are compared in a subsequent contrast test, which yields a χ^2-distributed test statistic (Snijders & Bosker, 1999; Goldstein, 2003).

$$\begin{aligned} Y_{ij} \;=\; & \text{D\_Positive}_{ij}(\beta_1 + e_{1ij} + u_{10j}) \\ & + \text{D\_Negative}_{ij}(\beta_2 + e_{2ij} + u_{20j}) \\ & + \text{D\_Bipolar}_{ij}(\beta_3 + e_{3ij} + u_{30j}) \end{aligned} \quad (1)$$

3 Results

Table 4 shows the parameter estimates for the models that were used to assess whether response effects can be generalized across studies.

Positive versus negative questions. As Table 4 shows, respondents are in general more likely to disagree with negative questions than to agree with equivalent positive questions for five out of the thirteen word pairs. Thus, for these word pairs, there is an overall nay-saying bias (appealing/distant $\chi^2 = 5.52$, $df = 1$, $p = 0.02$; simple/complicated $\chi^2 = 5.61$, $df = 1$, $p = 0.02$; personal/impersonal $\chi^2 = 28.68$, $df = 1$, $p < 0.001$; inviting/reluctant $\chi^2 = 49.82$, $df = 1$, $p < 0.001$; expert/amateur $\chi^2 = 35.26$, $df = 1$, $p < 0.001$).

The word pair *personal/impersonal* is one of the pairs for which an overall nay-saying bias is observed. Let us take this word pair to examine the difference between positive and negative questions more closely. For *personal/impersonal*, the nay-saying bias can be summarized as follows: a text is evaluated as being more *personal* when we ask how *impersonal* the text is. From the parameter estimates in Table 4, it can be calculated that the mean difference between those two wordings is 0.59 scalar points on (approximately) a seven-points scale (see Table 3). Hence, in a random study the expected difference between "disagree with impersonal" and "agree with personal" will be 0.59 scalar points. This, however, does not imply that in every future study a wording effect is expected to be found: there is considerable between study variation. The study standard deviation for this specific word pair, *personal/impersonal*, is 0.81 ($\sqrt{(0.73 + 0.60)/2}$). This means that the overall effect of question wording is medium in size as compared to the differences between studies (Cohen's $d = 0.73$). This also means that in an 80% con-

Table 2 Overview of study characteristics (1)

Study	N	Topic	Author	Respondents	Sampling procedure	Group/individual setting
1	151	Convince people to quit smoking	European Union	Heterogeneous	In trains	Individual & small groups
2	155	Convince people to join the asthma fund	Non-profit organization	Students	At the university	Individual
3	120	Convince people to take a credit card	Bank	Students	At the university	Individual & small groups
4	145	Inform people on teaching methods	Government	High school pupils	In high school classes	Group (school classes)
5	127	Inform people about alternative healing	Government	High school pupils	In high school classes	Group (school classes)
6	110	Convince people to book a holiday	Profit organization	Students on methodology	During a course	Group (school classes)
7	130	Convince people to be a donor	Government	Heterogeneous	In trains	Individual & small groups
8	150	Convince people that working for people aged 65+ is good	Government	Students	At the university	Individual
9	130	Convince women to buy breast growing pills	Profit organization	Heterogeneous	Lingery stores	Individual (women)
10	80[a]	Convince people to help the poor	Non-profit organization	Heterogeneous	In compagnies	Individual & small groups
11	210	Inform people about the new healthcare system	Government	Students	HBO students	Individual & small groups
12	111	Inform members about the Dutch Epilepsy foundation	Non-profit organization	Elderly	Epilepsy foundations' members	Individual

[a]In this study only the answers to positive and negative questions were compared.

Table 3 Overview of study characteristics (2)

Study	*N*	Type of sample	N Scale points[a]	Unipolar scale	Bipolar scale	Question wording
1	151	Convenience	7	Disagree-Agree	Unbalanced (pos-neg)	Subjective (I think ...)
2	155	Convenience	7	Agree-Disagree	Unbalanced (pos-neg)	Subjective
3	120	Convenience	7	Disagree-Agree	Balanced	Subjective
4	145	Convenience	7	Disagree-Agree	Unbalanced (pos-neg)	Subjective
5	127	Convenience	7	Disagree-Agree	Balanced	Subjective
6	110	Convenience	7	Agree-Disagree	Unbalanced (pos-neg)	Subjective
7	130	Convenience	7	Agree-Disagree	Unbalanced (pos-neg)	Objective (The text is ...)
8	150	Convenience	7	Disagree-Agree	Unbalanced (neg-pos)	Subjective
9	130	Convenience	7	Agree-Disagree	Unbalanced (pos-neg)	Objective
10	80	Convenience	5	Disagree-Agree	Does not apply	Objective + Subjective
11	210	Convenience	7	Disagree-Agree	Balanced	Subjective
12	111	Random sample (response rate: 20%)	5	Agree-Disagree	Unbalanced (Pos-Neg)	Objective
Total	1619					

[a]In each study the meaning of the scale points was explained in the introduction to the survey; the middle category of the positive and negative questions was explained to mean "neither agree nor disagree", whereas it was labeled as "neither X nor Y" in the bipolar version.

fidence interval the expected difference between "disagree with impersonal" and "agree with personal" will lie between −0.45 and 1.63 scalar points ($0.59 \pm 1.28 \cdot 0.81$), on, approximately, a seven-points scale. Hence, in a set of random studies respondents will usually be observed to express their opinions more positively when the question is worded negatively, but the size of this effect will vary across studies and in some studies there will be no effect at all, or we will even find an effect in the opposite direction. This same tendency applies to all five word pairs for which an overall wording effect is observed (see Table 4).

Apart from the five word pairs for which an overall effect is observed, there are eight word pairs for which no overall difference between positive and negative questions can be shown (all $\chi^2 < 3.84, df = 1, p > 0.05$). Judging from the means and the large between-study standard deviations

Table 4 Parameter estimates for the three question types (S_s^2: variance between studies; S_r^2: residual variance)

Word pair	Mean negative (S_s^2; S_r^2)[a]	Mean positive (S_s^2; S_r^2)	Mean bipolar (S_s^2, S_r^2)	Effect[b]
Personal/Impersonal	3.90 (0.6; 2.3)	3.31 (0.7; 2.0)	3.55 (0.4; 1.6)	N>B>P
Inviting/Reluctant	4.28 (0.4; 1.5)	3.49 (0.7; 1.8)	3.84 (0.3; 1.8)	N>B>P
Appealing/Distant	4.15 (0.5; 1.7)	3.68 (0.6; 1.7)	3.83 (0.2; 1.8)	N>P; N>B
Simple/Complicated	5.26 (0.8; 1.7)	4.59 (1.2; 1.7)	4.60 (0.9; 1.8)	N>P; N>B
Expert/Amateur	4.61 (0.5; 1.7)	4.07 (0.3; 2.0)	4.29 (0.1; 1.7)	N>P
Easy/Difficult	5.30 (0.7; 1.8)	4.98 (1.4; 1.5)	4.80 (1.0; 1.6)	N>B
Interesting/ Uninteresting	4.12 (0.4; 2.5)	3.93 (0.4; 2.2)	3.56 (0.5; 2.3)	P>B
Fascinating/Boring	3.82 (0.4; 2.4)	3.68 (0.4; 1.9)	3.54 (0.3; 2.1)	n.s.
Orderly/Chaotic	4.99 (0.3; 2.3)	4.68 (0.6; 1.7)	4.55 (0.4; 2.0)	n.s.
Logical/Illogical	4.55 (0.6; 2.3)	4.35 (0.8; 1.9)	4.26 (0.7; 1.9)	n.s.
Clear/Unclear	5.21 (0.3; 2.0)	4.94 (0.7; 1.5)	4.75 (0.8; 1.8)	n.s.
Concise/Wordy	4.46 (0.9; 2.1)	4.20 (0.7; 1.8)	4.08 (0.9; 1.6)	n.s.
Varied/Monotonous	3.86 (0.3; 2.3)	3.64 (0.2; 1.8)	3.60 (0.1; 1.8)	n.s.

Note. A higher mean score represents a more positive opinion towards the attitude object.
[a]The residual variance for each of the word pairs is relatively large. This is because the variation between persons is also at this level; the between-person variance cannot be estimated separately from the residual variance in this model.
[b]The significant differences between P (positive), N (negative) and B (bipolar) questions are given; > indicates a significantly more positive evaluation of the attitude object.

for these word pairs, there may be individual studies that do show an effect in the direction we have seen before (negative > positive). However, these are no systematic differences that are stable across experimental contexts.

Negative versus bipolar questions. When we compare the answers to the negative and the bipolar questions, results show that, for five out of the thirteen questions, respondents express their opinions more positively on negative questions than on bipolar questions (*appealing/distant* $\chi^2 = 9.31$, $df = 1$, $p < 0.01$; *simple/complicated* $\chi^2 = 4.98$, $df = 1$, $p < 0.03$; *personal/impersonal* $\chi^2 = 13.19$, $df = 1$, $p < 0.001$; *inviting/reluctant* $\chi^2 = 19.00$, $df = 1$, $p < 0.001$; *easy/difficult* $\chi^2 = 4.21$, $df = 1$, $p < 0.04$). Interestingly, four of these five word pairs also showed a difference between positive and negative questions. The expected mean differences for the negative and bipolar questions are generally somewhat smaller than mean differences for positive and negative questions. The study standard deviations, however, are comparable to what we have seen before. This suggests that, for the five word pairs showing an overall effect, we expect negative questions to be answered more positively than bipolar questions in a random future study. In practice, this prediction will come true most of the times, but there will also be studies in which no effect at all will be found or even an effect in the opposite direction.

For the remaining eight word pairs, no significant difference between the negative and the bipolar wording is observed (all $\chi^2 < 3.84$, $df = 1$, $p > 0.05$). Hence, for these word pairs, the effect of question wording cannot be generalized beyond the question level.

Positive versus bipolar questions. A comparison between the answers of positive and bipolar questions is not that straightforward. For the word pair *interesting/uninteresting* respondents express their opinions more positively when the question is worded positively ($\chi^2 = 19.28$, $df = 1$, $p < 0.001$). However, for the word pairs *personal/impersonal* and *inviting/reluctant* the opposite holds: respondents express their opinions more positively when asked bipolar questions as compared to positive questions ($\chi^2 = 5.41$, $df = 1$, $p < 0.02$ and $\chi^2 = 4.87$, $df = 1$, $p < 0.03$ respectively). For the remaining 10 word pairs, no significant differences between positive and bipolar questions are observed (all $\chi^2 < 3.84$, $df = 1$, $p > 0.05$).

4 Conclusion and Discussion

The current study reports on twelve split-ballot studies that investigate the choice for a positive, a negative, or a bipolar wording. In each of these studies, the same thirteen contrastive adjectives were used for the manipulations. Using this set-up, we investigated whether for each word pair the effect of question wording can be generalized across studies. Knowledge about the generalizability of response effects is important to predict wording effects for contrastive questions in future studies. Moreover, there is only need for a theoretical explanation of an underlying mechanism that causes the effect when wording effects for various contrastive questions can be generalized.

Results show an overall wording effect for about half of the word pairs investigated; for these word pairs the effect of question wording can be generalized across different types of respondents, different kinds of texts, different scale lengths, objective or subjective question wordings, and so forth. Important to note is that these generalizable wording effects are largely consistent in their direction: respondents are more likely to disagree with negative questions than to agree with positive questions or to choose the positive side of the bipolar scale. In other words, where respondents usually give comparable answers to positive and bipolar questions, respondents express their opinions more positively to negative questions than to positive and bipolar questions. Importantly,

these results indicate that a similar effect is to be expected in future individual studies in which these word pairs are used. Yet, the large between study variance indicates that the size of the wording effect is likely to fluctuate in such future studies, and that there may also be cases in which no effect at all will be observed.

For the other half of the word pairs, results of individual studies could not be generalized. This indicates that for these word pairs no wording effect can be expected in future studies. However, this does not mean that a wording effect may never be found; the current study provided a pretty strict test for the generalizability of wording effects, and the large between study standard deviation clearly showed that wording effects may sometimes be found these word pairs too. Judging from the parameter estimates, it can be concluded that when an individual future study shows a wording effect it is likely to be in the same direction as we have seen previously: respondents express their opinions more positively when the question is worded negatively.

All in all, results of the current study are in line with previous studies on wording effects for contrastive questions showing variation in the size and direction of wording effects (e.g., Falthzik & Jolson, 1974; Holleman, 1999a; Menezes & Elbert, 1979; O'Neill, 1967; Schuman & Presser, 1981). The current study adds to the existing literature that for some word pairs wording effects are mostly random error effects, while for other word pairs the wording effects are substantial when generalizing over variation in all sorts of contextual study characteristics. The direction of the wording effects is in line with the wording effect observed for *forbid* an *allow* questions (Holleman, 1999a): respondents are more likely to disagree with negative questions than to choose the positive side of the bipolar scale.

For survey practice, results of the current study have at least two important implications. First, the results once again show that answers given to survey questions can only be interpreted with respect to the exact wording of the question. That is, if a respondent disagrees with the statement *this book is fascinating*, survey researchers must not conclude that the respondent thinks that the book is *boring*. In other words, no absolute meaning can be attached to survey answers; survey answers are measured on an interval scale at best.

Second, results indicate that multiple questions should be used for measuring theoretical constructs. Based on the results of the current study, we advise against measuring attitudes with mixed sets of positive and negative questions: as respondents express their opinions differently to positive and negative questions, mixing positive and negative questions will necessarily create additional between item variance (error variance). As the reliability is a function of the between item variance (Lord & Novick, 1968), the reliability will decrease if the between item variance increases.

The current study raises several issues for future research. For one, although this study focuses on the generalizability of wording effects, some generalization issues are left unaddressed. Most importantly, the current research integrates studies targeted at measuring people's opinions and attitudes towards texts. As the reading of the text was the respondent's first encounter with the attitude object, these opinions and attitudes were always computed on the spot. Therefore, in a future study, it would be interesting to measure the effect of the same word pairs in attitude questions with respect to more deeply rooted attitudes and other attitude objects.

Another aim for future studies might be to better understand the variation surrounding the response effects observed here. First, variation has shown to arise on a study level. This variation probably exists because the wording effect interacts with all kinds of contextual characteristics, such as the number of scale points, the position of the question in the survey etcetera. In a meta-analysis by Holleman (1999a), an exploration is presented about how certain contextual factors explain part of the between study variance for *forbid* and *allow* questions. However, as Holleman notes, such a post hoc "explanation" is not unproblematic, because many experimental characteristics are confounded, and because the interactions between the different contextual variables can hardly be modeled because there are too many of them. Therefore, if more insight into these factors is required, it would be best to systematically vary those characteristics across multiple smaller studies.

Second, the current study also shows that it is likely that between word pair variation in survey wording effects exists in addition to the between study variation; while the object of evaluation was identical for all word pairs in all studies (i.e., "the text"), we did observe variation in the occurrence of wording effects across word pairs. In semantics, there is a growing interest in the way gradable adjectives can be classified (Kennedy & McNally, 2005; Kennedy, 2007; Rotstein & Winter, 2004). A distinction is made between words for which the reference point is fixed, and words for which the reference point is more context dependent. In a future study it would be interesting to examine if such a classification explains the between word pair variation that was observed here.

The most important step for future research, however, is to obtain insight into the validity if contrastive questions; the current study has shown that in spite of between word study and between word pair variation, wording effects for contrastive questions exist. Therefore, insight into the validity of contrastive questions is required to know what we are actually measuring with positive, negative, and bipolar questions. Such insight can be obtained by comparing the three wording alternatives with respect to various quality criteria, such as the divergent validity, and the stability over time (cf. Friborg, Martinussen & Rosenvinge, 2006; Saris, Revilla, Krosnick & Shaeffer, 2010). Another possibility would be to investigate data quality from a more theoretical perspective: are contrastive questions equally valid, i.e., measuring the same underlying attitude, or not? For *forbid* and *allow* questions, this question has been investigated by looking into the cognitive processes underlying question-answering (Chessa & Holleman 2007; Holleman 1999b; for a cognitive model see Tourangeau, Rips & Rasinski 2000). Such an approach provides theoretical insight into the representation of attitudes, as well as practical knowledge about the validity of survey

questions. Hence, both of these reasons provide ample justification for such a future study.

References

Ajzen, I. (1988). *Attitudes, personality and behavior.* Chicago: Dorsey.

Anderson, E. W., & Fornell, C. (2000). Foundations of the american customer satisfaction index. *Total Quality Management*, *11*, 869-882.

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

Brooke, J. (1996). Sus: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, B. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (p. 189-194). London: Taylor and Francis.

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

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Orlando: Harcourt Brace Jovanovich.

Falthzik, A. M., & Jolson, M. A. (1974). Statement polarity in attitude studies. *Journal of Marketing Research*, *11*, 102-105.

Friborg, O., Martinussen, M., & Rosenvinge, J. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: a psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, *40*, 873-884.

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

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

Hamilton, H. W., & Deese, J. (1971). Does linguistic marking have a psychological correlate? *Journal of Verbal Learning and Verbal Behavior*, *10*, 707-714.

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.

Holleman, B. C. (1999a). Wording effects in survey research. using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics*, *6*, 29-40.

Holleman, B. C. (1999b). The nature of the forbid/allow asymmetry. Two correlational Studies. *Sociological Methods and Research*, *28*, 209-244.

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

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*, 1-45.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification and the semantics of gradable predicates. *Language*, *81*, 345-381.

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

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

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores.* Addison-Wesley: Reading Mass.

Maes, A., Ummelen, N., & Hoeken, H. (1996). *Instructieve teksten. Analyse, ontwerp en evaluatie. [instructive texts. analysis, design and evaluation.].* Bussum: Uitgeverij Coutinho.

Menezes, D., & Elbert, N. (1979). Alternative semantic scaling formats for measuring store image: an evaluation. *Journal of Marketing Research*, *16*, 80-87.

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* (p. 49-89). London: Academic Press.

Muylle, S., Moenaert, R., & Despontin, M. (2004). The conceptualization and empirical validation of web site user satisfaction. *Information and Management*, *41*, 543-560.

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

O'Neill, H. W. (1967). Response style influence in public opinion surveys. *Public Opinion Quarterly*, *31*, 95-102.

Oskamp, S., & Schultz, W. P. (2005). *Attitudes and opinions. 3rd edition.* Mahwah: Lawrence Erlbaum Associates.

Rotstein, C., & Winter, Y. (2004). Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics*, *12*, 259-288.

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

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*, 61-79.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys. experiments on question form, wording and context.* London: Academic Press.

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

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

van Schaik, P., & Ling, J. (2005). Five psychometric scales for online measurement of the quality of human-computer interaction in web sites. *International Journal of Human-Computer Interaction*, *18*, 309-322.

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.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*, 236-247.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research.* Chicago: The University of Chicago Press.

Appendix

Table A1 Overview of studies and subjects per word pair

Word Pair	Study numbers	*N*
Logical/Illogical	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	1619
Easy/Difficult	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	1619
Clear/Unclear	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12	1539
Orderly/Chaotic	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12	1539
Fascinating/Boring	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12	1499
Interesting/Uninteresting	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12	1499
Concise/Wordy	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12	1469
Personal/Impersonal	1, 2, 4, 5, 7, 8, 9, 10, 11, 12	1389
Varied/Monotonous	1, 2, 4, 5, 6, 7, 8, 9, 10, 11	1388
Inviting/Reluctant	1, 2, 4, 5, 6, 8, 9, 10, 11, 12	1369
Simple/Complicated	1, 2, 4, 5, 6, 7, 8, 9, 10, 11	1357
Appealing/Distant	1, 4, 5, 9, 10, 11, 12	958
Expert/Amateur	1, 2, 3, 8, 9, 10, 12	897