# A PSYCHOLINGUISTIC LOOK AT SURVEY QUESTION DESIGN AND RESPONSE QUALITY

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von
Timo Lenzner, M.A.

Dekan der Fakultät für Sozialwissenschaften:
Prof. Dr. Thorsten Meiser, Universität Mannheim

1. Gutachter:
Prof. Dr. Michael Braun, GESIS – Leibniz-Institut für Sozialwissenschaften & Universität Mannheim

2. Gutachter:
Prof. Dr. Rüdiger Schmitt-Beck, Universität Mannheim

Tag der Disputation:
15. November 2011

# CONTENTS

# LIST OF TABLES

# ACKNOWLEDGEMENTS

Many people have contributed to the successful completion of this doctoral thesis. First of all, I would like to thank my supervisor Prof. Dr. Michael Braun for his critical advice and encouraging feedback on my research. I would also like to express my gratitude to Prof. Dr. Rüdiger Schmitt-Beck for co-refereeing this dissertation.

The experiments reported in this thesis could not have been conducted without the help of my co-authors. I especially thank Dr. Lars Kaczmirek for the technical and organizational support in performing the Web-based experiments and, of course, for the countless hours of valuable discussion. I am also grateful to Dr. Mirta Galesic for her co-operation in the eye-tracking study and for allowing me to use the eye-tracking equipment at the Max Planck Institute for Human Development.

I have also greatly benefited from the encouraging and friendly environment at GESIS – Leibniz Institute for the Social Sciences. Many thanks to Dr. Wolfgang Bandilla, Dr. Dorothée Behr, PD Dr. Henning Best, Prof. Dr. Ingwer Borg, Dr. Christoph Kemper, Julia Khorshed, Dr. Natalja Menold, Rolf Porst, Peter Prüfer, Prof. Dr. Beatrice Rammstedt, Ines Schaurer, Dr. Evi Scholz, Bella Struminskaya, Cornelia Züll, and all my former and current colleagues at the institute.

Finally, I would like to thank my family for their unconditional love and believing. Most importantly, I thank Dr. Alwine Lenzner for both her professional and emotional support throughout the years, for co-authoring the pilot study, and for co-authoring my life.

# 1 INTRODUCTION

Asking questions is the predominant method of gathering information about people's beliefs, values, attitudes, behaviors and states of affaires (e.g., Foddy, 1993; Schuman & Presser, 1981). Each year, political campaigns, governments, newspapers, market and social research agencies, companies, and university scholars conduct thousands of surveys to find out about such things as the public's view on political and social issues, people's confidence in political leaders, their preferences in upcoming elections, customer and employee satisfaction, and economic trends (cf. Fowler, 1995; Groves et al., 2004). On the basis of these data, many important decisions are being made, and hence it is fundamental that the answers to survey questions accurately reflect peoples' opinions, states of affairs, and whatever information the questions ask for.

To ensure that the data obtained through surveys are accurate, reliable, and lead to valid conclusions, respondents must understand the questions as intended by the survey designer and have access to the information being sought. More specifically, they must thoroughly carry out the following cognitive tasks: (1) comprehend the question, (2) retrieve relevant information, (3) use this information to make a judgment, and (4) select and report an answer (Strack & Martin, 1987; Tourangeau, 1984). Depending on various characteristics of the questionnaire, respondents may find it more or less difficult to perform these steps accurately. For example, question comprehension is impeded by questions containing imprecise terms or complex syntactic structures which make it difficult for respondents to identify the question focus or represent the logical form of the question (cf. Tourangeau, Rips, & Rasinski, 2000). Similarly, retrieving relevant information is particularly difficult if questions ask about events that happened a long time ago and which are difficult to recall (Canell, Miller, & Oksenberg, 1981; Loftus, Smith, Klinger, & Fiedler, 1992; Smith & Jobe, 1994). A suboptimal wording of survey questions can thus increase respondent burden and thereby negatively affect the accuracy of their answers.

This thesis focuses on survey question characteristics that affect the first cognitive task that respondents have to carry out, that is, understanding what the

question is about. Specifically, it takes a closer look at the ways in which the wordings of survey questions affect question comprehensibility and, in turn, how question comprehensibility affects the quality of the answers respondents provide.

It is generally acknowledged among survey researchers that asking clear questions that are easily and consistently understood by all respondents is a prerequisite for obtaining meaningful responses, and hence reliable and valid data. This notion has been so prominent in the questionnaire design literature that it can almost be conceived as being axiomatic (Fowler, 1992). Ideally, respondents find it easy to understand the meaning of a question and interpret it in the way the researcher intended. To achieve these goals, survey designers need to formulate questions that are (1) unambiguous and (2) require little processing effort.

Earlier research focused almost exclusively on the first of these two aspects and examined the effects of unclear (e.g., vague or ambiguous) question wordings on interpretation variability. For example, it has been shown that vague relative terms, such as *often* or *substantially*, are interpreted quite differently among respondents, depending on the content of the question and respondents' gender, age, education, and race (Bradburn & Miles, 1979; Schaeffer, 1991). Similar effects have been found for ambiguous or abstract terms, such as *exercise*, *welfare*, or *most people*, which can mean different things to different respondents (Fowler, 1992; Smith, 1987; Sturgis & Smith, 2010). Hence, vague or ambiguous terms can lower response quality and increase measurement error in the survey data.

The effort required to comprehend a survey question (i.e., its comprehensibility) may affect responses similarly. If questions are difficult to understand (because of their linguistic complexity), respondents may not be willing or able to invest the additional effort required to overcome these difficulties, and thus may not provide meaningful answers. Instead, they are likely to provide inaccurate responses (Schober & Conrad, 1997) and/or apply response strategies that reduce data quality and induce measurement error (e.g., breakoff, Galesic, 2006; satisficing, Krosnick, 1991). For example, if respondents experience difficulties in determining the meaning of an attitudinal question, they may decide that they have no strong opinion about this issue and then provide a non-substantive (e.g., "don't know") or neutral

(e.g., “neither/nor”) response. Similarly, if confronted with an ambiguous question, respondents may not simply interpret it idiosyncratically, which would be easy for them to do but result in high interpretation variability. Instead, if they perceive the ambiguity and the difficulty involved in answering the question meaningfully, they may decide that resolving this ambiguity is too wearisome and then may offer a non-substantive response. In both cases, the difficulty with understanding the meaning of a question would lead to inaccurate answers. Moreover, given that question comprehension is the first step respondents have to perform, it is very likely that cognitive overload occurring at this stage will translate to later stages as well. Consequently, designing questions to minimize the cognitive effort required to process them is an important strategy for reducing comprehension difficulties and thus response error. The ways in which the cognitive effort required to comprehend a survey question affects response quality have received comparatively little attention to date and have rarely been examined experimentally (see Krosnick, 1991, for a theoretical discussion of this issue).

This thesis aims at filling this gap by applying a psycholinguistic perspective to survey question design. Psycholinguistics is an interdisciplinary field of research at the intersection of psychology, linguistics, and neuroscience which studies “the mental processes and skills underlying the production and comprehension of language, and […] the acquisition of these skills” (Levelt, 1992, p. 290). Simplified, it comprises three main research areas: (1) language production, (2) language comprehension, and (3) language acquisition. Of particular interest for survey designers is certainly the research evidence obtained in the second area. For example, typical research questions in psycholinguistic studies on language comprehension are: How are the words stored in the brain and how do humans access their meaning when they need them? What is the role of memory in the processing of linguistic input? How do listeners and readers make sense of the strings of letters and words that they encounter? And on the applied side: What insights can psycholinguistics gain that might assist writers in formulating comprehensible texts? (cf. Bloomer, Griffiths, & Merrison, 2005). The answers provided to the last one of these questions may be particularly relevant for survey designers when they try to write questions

that are easy to understand and to answer. In short, a psycholinguistic perspective on questionnaire design may provide an answer to one of the fundamental questions that survey designers are concerned with: "Given the information that we want, what's the best way of structuring the question as a linguistic object to get at those facts and opinions?" (Tourangeau et al., 2000, p. 27).

The remainder of this thesis is arranged as follows. In Chapter 2, I take a closer look at the cognitive and linguistic processes involved in survey question comprehension and discuss two theories which indicate that the cognitive effort required to perform these processes has an important impact on respondents' answer behavior (*Relevance Theory*, Sperber & Wilson, 1986; *Satisficing Theory*, Krosnick, 1991). Chapter 3 then reviews what is known in the survey methodological literature about designing comprehensible survey questions that minimize respondent burden. Arguing that the present guidelines of asking questions are often vague and underspecified, I provide an overview of specific text features that have been found to undermine comprehension in psycholinguistic research studies. Chapter 4 is dedicated to the main research questions of this thesis and to the objectives of the three empirical studies that were conducted to answer these questions. The first study (Chapter 5) was a pilot study to examine whether the seven problematic text features identified in Chapter 3 reduce question comprehensibility, increase respondent burden, and lower response quality. Using response times as measures of the cognitive effort required to answer survey questions, the results indicated that most text features induce comprehension difficulties. Moreover, the text features were found to reduce response quality as indicated by an increase in midpoint responses compared to questions without these text features. The second study (Chapter 6) extended upon these findings in two ways. First, eye tracking was used as a more direct method to examine whether comprehension is indeed impeded by these text features. Second, the study examined whether the text features have different effects for different types of questions (attitudinal, factual, and behavioral). Finally, a third study (Chapter 7) was conducted to look at the measurement consequences of these text features in more detail. The study examined whether the low question comprehensibility induced by the text features reduces response quality, and if so,

whether these effects are moderated by respondent characteristics such as their verbal intelligence and their motivation to answer survey questions. Chapter 8 closes with a summary of the major findings, a discussion of the implications of these results for survey question design and evaluation, and an outlook on future research directions.

# 2 MAKING SENSE OF SURVEY QUESTIONS

This chapter takes a detailed look at the various cognitive and linguistic processes that respondents have to perform when making sense of survey questions. Furthermore, it presents two theories on communication and survey responding which imply that the effort involved in carrying out these processes has a crucial impact on the quality of respondents' answers.

## 2.1 The question comprehension process

Understanding a survey question requires respondents to perform a series of cognitive and linguistic tasks. These can be divided into two broad categories: (1) decoding the semantic meaning and (2) inferring the pragmatic meaning of the question. Semantic processes include determining the lexical meaning of the individual words, linking these meanings to relevant concepts in memory, identifying the sentence form (e.g., whether it is an interrogative question or statement), and parsing the sentence into its grammatical parts (such as subjects and objects). Pragmatic operations include assigning the contextual meaning to the individual words and inferring the questions' point, that is, identifying the information sought (Carroll, 2004; Graesser, McMahen, & Johnson, 1994; Tourangeau et al., 2000). Consider the following example:

(1) In your free time, how often do you go to the theater?

In comprehending this question, respondents first parse the different signs and strings into words, assign meanings to these words (e.g., *theater* denotes a building in which plays, shows, and other performances are presented), and link these meanings to concrete representations in their memories (e.g., to their mental representations of specific theaters). Moreover, they identify the subject (*you*) and object (*theater*) of the question and determine its grammatical and logical structure. When processing the wh-word *how*, they identify the question as an interrogative and recognize that they are supposed to provide an answer (which is tacitly agreed upon in ordinary conversation, cf. Grice, 1989). Furthermore, they infer the intended (or pragmatic)

meaning of the question and recognize that they are only supposed to report how often they attend theatrical performances, excluding occasions on which they enter the building for other purposes (such as delivering a parcel). Finally, if the question offers a set of response options, such as *several times a week, several times a month,* and *several times a year*, respondents use these to specify their understanding of the question's point (e.g., that they are expected to provide an average account and not an exact numerical response). The sequential order in which these processes are presented here might suggest that respondents perform them in a successive way. However, these operations are interdependent and are usually carried out simultaneously (Sudman, Bradburn, & Schwarz, 1996).

Ideally, respondents find it easy to perform these processes and eventually interpret a question in the way intended by the survey designer. In reality, however, comprehending a question often poses more of a challenge for respondents. For example, individual words may be ambiguous or unfamiliar so that respondents have difficulties to access their meaning. The question's syntactical structure may be complicated and parsing the question into its grammatical components may exceed people's working memory capacity. Or the question may be hypothetical, requiring respondents to build a mental representation of the hypothetical situation and hold it in memory while processing the rest of the question. In short, there are several psycholinguistic aspects of survey questions which make them difficult to comprehend. Consider question (2), for example:

(2) In a typical week, how much time, in total, do you spend on the Internet?

In this question, the adjective *typical* is an imprecise relative term with no clear empirical referent. Hence, respondents may find it difficult to link the meaning of the word to a concrete representation in their memories. *Week* is an ambiguous noun which can either mean *workweek* (i.e., Monday to Friday) or *seven-day-week* (including the weekend). The question's syntax is left-embedded requiring respondents to process many prepositional phrases and qualifiers before they encounter the main verb of the main clause. Moreover, respondents may have difficulties to identify the pragmatic meaning of the pronoun *you* (which can either

refer to *you yourself* or to *you and your housemates*) and the pragmatic meaning of the question as a whole (i.e., whether it asks about time spent on the Internet at home, at work, or both). Finally, respondents may wonder whether *spending time on the Internet* refers only to times in which they actively surf the Internet or whether this also includes time when the computer is on and connected to the Internet, but they are engaged in other activities (e.g., while downloading large files). In sum, questions like (2) may require respondents to invest a considerable deal of cognitive effort in order to overcome these comprehension difficulties and to make sense out of the question.

Now, the crucial point is whether respondents can be expected to invest this effort, and hence to overcome these difficulties, or not. If respondents (or some subgroups of respondents) perceive the additional effort as burdensome and have difficulties or no interest in investing it, then formulating comprehensible questions that are *easy to understand* becomes an important objective in survey question design. In the following, I discuss two theories which address the role of cognitive effort in information processing and communication in general (*Relevance Theory*) and in survey responding in particular (*Satisficing Theory*). Both theories imply that the ease with which respondents are able to perform the semantic and pragmatic processes involved in question comprehension has an important impact on how well they understand questions and how likely they are to provide meaningful responses.

## 2.2 Relevance theory

Relevance theory is recognized within linguistics as a major theoretical approach to understanding communication and human cognition. About two decades ago, it was introduced into the survey methodological literature and has since helped to illuminate the pragmatic aspects of the survey situation (e.g., Clark & Schober, 1992; Schwarz, 1996; Sudman et al., 1996). For example, it has been noted that survey interviews, like all other forms of conversation, are governed by the *principle of relevance*, according to which every "communicated information comes with a guarantee of relevance" (Sperber & Wilson, 1986, p. VII). As a consequence of this principle, survey respondents expect the questions in a survey to be personally

relevant, that is, to pertain to their lives and ask about issues they have information about or opinions on. This helps to explain, for example, why respondents even answer questions about completely fictitious issues, such as a non-existent "Monetary Control Bill" (Bishop, Tuchfarber, & Oldendick, 1986): it is simply highly unlikely that a surveyor would ask a completely irrelevant question that would provide no meaningful information. Hence, respondents try to figure out how the survey designer may have intended the question and what the obscure "Monetary Control Bill" might refer to.

Relevance theory does not only help to explain how several response effects emerge from the pragmatic rules that govern communication. It also has important implications for the formulation of survey questions and its impact on question comprehensibility. This aspect of relevance theory has not received wide recognition in empirical studies on survey question responding to date. In addition to the principle of relevance, an essential claim of relevance theory is that humans are very efficient information processing devices, consistently balancing efforts and rewards:

> "Our claim is that all human beings automatically aim at the most efficient information processing possible. This is so whether they are conscious of it or not; […]. Information processing involves effort; it will only be undertaken in the expectation of some reward" (Sperber & Wilson, 1986, p. 49).

People are constantly surrounded by a huge array of information competing for their attention, and hence it is essential to distinguish between relevant and irrelevant information. One factor that determines whether information is deemed relevant or not is the contextual effect expected as a result of processing the information. Information that is unlikely to have any contextual effect, that is, any added informative value, is not worth processing, and hence considered irrelevant in the given context. In survey situations, respondents usually perceive some contextual effect of the questions. After all, the surveyors are interested in information respondents should be able to provide, and thus they are unlikely to ask for information that is totally unrelated to the respondents' representation of the world. As mentioned above, asking completely irrelevant questions would violate the

principle of relevance and benefit neither the respondents nor the surveyors. However, having some contextual effect is not sufficient for a piece of information (and hence a survey question) to be relevant.

A second determinant of relevance is the effort required to bring about the contextual effect. According to relevance theory, "[p]rocessing effort is a negative factor: other things being equal, the greater the processing effort, the lower the relevance" (Sperber & Wilson, 1986, p. 124). Hence, information, or more specifically, survey questions that are difficult to process, are less relevant than those that are easier to process. Thus, a prerequisite for survey questions to be relevant for respondents is that they require comparatively little processing effort. To illustrate how the relative relevance of a survey question is affected by processing effort, compare (3) and (4) in the context of (5):

(3) During the last four weeks, how often did you suffer from *somatic* pain?
(4) During the last four weeks, how often did you suffer from *physical* pain?
(5) Context: Respondent experienced two headaches during the last four weeks.

In the context of (5), both questions (3) and (4) aim at exactly the same contextual effect: respondents are asked to report the two instances of headaches they experienced during the last four weeks. However, the questions differ in the amount of processing effort required to answer the questions: (3) contains the low-frequency term *somatic*, which is unfamiliar to many respondents, and thus more difficult to comprehend than the more frequent synonym *physical* used in (4). Hence, from a relevance theory perspective, question (3) should be less relevant than question (4), which achieves the same contextual effect with less processing effort. Now, when confronted with question (3), respondents may conclude that answering this question in a thoughtful way requires more processing effort than is warranted by the contextual effect. After all, why should they expend energy on trying to figure out what exactly the unfamiliar term *somatic* may mean? Being efficient information processing devices, they may stop thinking after making a best guess and continue with the next question. This kind of response behavior has been documented and discussed in terms of *satisficing* in the survey methodological literature.

## 2.3 Satisficing theory

Satisficing theory (Krosnick & Alwin, 1987; Krosnick, 1991) is based on the presumption that optimally answering a survey question often requires respondents to invest a great deal of cognitive effort. For example, respondents are asked to recall past behaviors such as the number of visits to a doctor during the past year (e.g., Adams, Hendershot, & Marano, 1999), a question which may require a quite demanding memory search. Or they are asked to select what they think are the three most desirable personal qualities for a child to have among a list of thirteen items (Kohn, 1969), which requires them to evaluate each quality in comparison to all others. At the same time, they are rarely compensated for their efforts in an apparent and straightforward way (e.g., in the form of an adequate monetary incentive or the conviction to influence policy-makers). This imbalance between respondents' costs and rewards associated with participating in a survey suggests that respondents may not always perform the cognitive processes of answering a question (comprehension, information retrieval, judgment, response selection) thoroughly and accurately (i.e., they do not always "optimize"). Instead, they may try to shortcut these processes and, for example, interpret questions only superficially, stop searching their memories after retrieving the first piece of relevant information, perform a judgment more carelessly, and select a response option more randomly. This response behavior has been termed "survey satisficing" (Krosnick, 1991). Satisficing respondents employ various response strategies that allow them to avoid strenuous cognitive work while still appearing as if they were completing the survey appropriately. For example, satisficing response strategies include saying "don't know" instead of reporting an opinion, selecting the first answer option that seems reasonable, or selecting the midpoint response option. All of these strategies are problematic in surveys as they increase measurement error and produce lower-quality data.

According to Krosnick (1991), the probability that respondents apply satisficing response strategies is a function of three factors: task difficulty (i.e., question difficulty), respondent ability, and respondent motivation. The more difficult a survey question is to answer, and the lower the respondents' cognitive abilities and

motivation, the more likely they are to satisfice in a survey. Survey questions can be difficult to answer for a number of reasons. First, they may be difficult to comprehend, for example, if they contain vague words or complex syntactic structures which make it difficult for respondents to identify the question focus or represent the logical form of the question. Second, questions may ask respondents to perform a difficult information retrieval task, such as reporting about events that happened a long time ago. Third, they may pose a challenge at the judgment stage, for example, if respondents are asked to indicate which one is the most serious problem facing the country today among a list of several problems. This question requires respondents to perform a comparative judgment which can be quite demanding. Finally, questions may be difficult to answer if they offer multiple-point scales with verbal labels for the end-points only. In these scales, the meaning of the mid-points is ambiguous, making it potentially difficult for respondents to select the appropriate category. All of these difficulties enhance the likelihood that respondents employ satisficing response strategies when answering these questions.

A second determinant of satisficing is respondent ability. People may find it more or less burdensome to optimally respond to survey questions depending on their question-relevant knowledge (e.g., whether they have a consolidated opinion on the issue in question), their experience with performing complex mental operations, their ability to process information, their vocabularies, and their ability to verbally express themselves. Respondents who are low in these abilities are more likely to satisfice because the effort required of them is greater than the effort required of people who are high in these abilities. For example, respondents with limited vocabularies and lower language processing abilities may have considerable difficulties in interpreting questions that include technical terms or complex logical structures.

Finally, respondents' tendency to satisfice is determined by their motivation to answer survey questions. For example, the topic of some questions (or the whole survey) may be personally important for respondents. Hence, they may feel excited to communicate their views on the issue and are unlikely to use cognitive shortcuts in answering the questions. Other respondents may generally believe in the usefulness of surveys for society and for politicians to make informed decisions. In their view,

the survey outcomes may warrant the effort. Moreover, respondents differ from each other in their need for cognition (Cacioppo & Petty, 1982) and their need to evaluate (Jarvis & Petty, 1996). While need for cognition (NFC) is an indicator of how much people enjoy thinking and performing effortful mental exercises, need to evaluate (NTE) is a measure of how opinionated people are and how willingly they engage in evaluation. People who are low in NFC and/or NTE are presumably more susceptible to satisfice in surveys than those high in these traits (cf. Krosnick, 1991, Toepoel, Vis, Das, & van Soest, 2009).

## 2.4 Implications of the theories for survey question design

Both relevance theory and satisficing theory suggest that the comprehensibility of survey questions has an important impact on respondents' answers. According to the theories, respondents should not be expected to be able or willing to overcome all kinds of complexities irrespective of the cognitive effort required to do so. Instead, they are likely to use their available processing resources efficiently, consistently monitoring efforts and rewards, and terminating the comprehension process if the rewards no longer warrant the efforts. An important claim of relevance theory is that people automatically do so when processing any kind of information, whether they are conscious of it or not. Hence, the greater the effort of decoding and inferring the meaning of a question required, the less likely respondents are to perform these processes accurately and thoroughly. According to satisficing theory, the probability that respondents invest the required processing effort is moderated by two respondent characteristics: ability and motivation. The more difficult a survey question is to answer, and the lower the respondents' cognitive abilities and motivation, the more likely they are to satisfice in a survey. Hence, the comprehensibility of survey questions is particularly relevant for respondents low in cognitive ability and motivation to respond to a survey. In sum, both theories imply that it is important to design survey questions in a way so that respondents find it easy to interpret and answer them. An important goal during question design should be to minimize respondent burden and thus enabling respondents to efficiently use their available processing resources.

# 3 DESIGNING COMPREHENSIBLE SURVEY QUESTIONS

The previous chapter argued that the cognitive effort required for comprehending survey questions has a considerable impact on the quality of respondents' answers. Consequently, attempts should be made to reduce this effort by formulating clear and comprehensible questions. This chapter reviews what is known in the survey methodological literature about designing comprehensible questions and presents additional insights from psycholinguistics on how to enhance text comprehensibility.

## 3.1 Guidelines of asking survey questions

Survey researchers have long been concerned about the comprehensibility of their questions (Belson, 1968, 1981; Cantril, 1944; Payne, 1951) and a great deal of research has been done on the impact of question wording on respondents' answers (e.g., Converse & Presser, 1986; Schuman & Presser, 1981; Sudman & Bradburn, 1982). These early studies demonstrated that even small changes in the words used in a survey question can completely change respondents' understanding, and hence the response distribution of that question (e.g., Smith, 1987). Moreover, they revealed that survey questions are often difficult for respondents to understand and to answer, and thus may not always elicit meaningful, reliable, and valid responses (e.g., Belson, 1981). Consequently, survey researchers developed so-called "guidelines," "standards," or "principles" of asking survey questions, which are aimed at minimizing flaws in the wording of the questions (e.g., Belson, 1981; Bradburn, Sudman, & Wansink, 2004; Fink, 1995; Fowler, 1995). These guidelines are general rules of thumb emphasizing, for example, the need to avoid long or complex questions, difficult or unfamiliar terms, and questions that call for a lot of respondent effort. Consider some of the guidelines compiled by Belson (1981, p. 389), for instance:

AVOID:

- loading up the question with a lot of different or defining terms;
- offering long alternatives (as possible answers to a question);
- the use of words that are not the usual working tools of the respondent;

- the use of words that mean something different if partly misheard;
- giving the respondent a difficult task to perform;
- giving the respondent a task that calls for a major memory effort;
- offering alternatives that could *both* be true.

Even though the guidelines are useful in avoiding gross mistakes, their major drawback is that they are rather vague and often lack explicit definitions (cf. Porst, 2008). Hence, it is up to the survey designer's subjective interpretation to decide if specific words are "not the usual working tools of respondents" or if answering a particular question requires respondents "to perform a difficult task." Moreover, when survey designers have identified a problematic question, the principles rarely offer any advice on how to repair it. In addition, the guidelines are sometimes contradictory, so that making a question better in one respect makes it worse in some other respect (Fowler, 2001). For example, making a question more precise by adding explanations, clarifying phrases, and definitions can easily make it syntactically more complex, and thus more difficult to comprehend. All in all, it is important to conceive the guidelines only as a rough framework and a starting point from which more concrete principles need to be derived. In his 1981 book, Belson already cautioned his readers that "in the long term it is most important that we understand the principles and the processes that underlie such guidelines and that the guidelines themselves be subject to challenge through research" (p. 390). Even though many valuable findings have been obtained since Belson's statement, to date, "crafting survey questionnaires remains something less than a scientific enterprise (Tourangeau, 2004, p. 209).

Only recently have survey methodologists turned to the examination of specific text features (i.e., linguistic properties) of survey questions in order to explain why some questions are more difficult to comprehend than others (Graesser, Cai, Louwerse, & Daniel, 2006; Lessler & Forsyth, 1996; Saris & Gallhofer, 2007; Tourangeau et al., 2000). Evidence from psycholinguistics indicates that certain text features, such as low-frequency words or vague relative terms, cause comprehension difficulties and can thus have a strong impact on response quality. The next section

provides an overview of these text features and thereby aims at establishing a more sophisticated basis for the formulation of survey questions. A specification of these text features and their relation to question comprehensibility may help practitioners to systematically check and improve the comprehensibility of their questions. Manuals describing the text features in detail may supplement the existing guidelines of asking survey questions and lend further precision to these rules.

## 3.2 Psycholinguistic determinants of question comprehensibility

Theoretical and empirical evidence from psycholinguistics (e.g., Duffy, Morris, & Rayner, 1988; Haviland & Clark, 1974; Horning, 1979; Inhoff & Rayner, 1986; Kimball, 1973; Kintsch & Keenan, 1973; Mosier, 1941) indicates that survey designers can enhance the comprehensibility of their questions by avoiding several text features that make survey questions difficult to process and to understand:

- Low-frequency words
- Vague or imprecise relative terms
- Vague or ambiguous noun phrases
- Complex syntax
- Complex logical structures
- Low syntactic redundancy
- Bridging inferences

In general, these text features undermine reading comprehension by placing high demands on people's limited working memory capacity (Baddeley, 1986; Ericsson & Kintsch, 1995; Just & Carpenter, 1992). A first attempt to systematically link these text features to survey question comprehension has been made by Graesser, Cai, Louwerse, and Daniel (2006) who developed the computer tool *Question Understanding Aid* (QUAID). QUAID evaluates survey questions with regard to the first five text features listed above, and labels those questions as problematic that include one or more of these features. Evidence from reading research, however, suggests that there are at least two more variables that affect question

comprehensibility to a similar degree, namely low syntactic redundancy (Horning, 1979) and bridging inferences (e.g., Vonk & Noordman, 1990). Incorporating these into QUAID may enhance the validity of the tool and cover additional aspects of survey question comprehensibility.

Besides extending QUAID by discussing two further problematic text features, in the following I deviate to a certain degree from Graesser et al.'s (2006) terminology. First, the text feature that I call *low-frequency words* is termed *unfamiliar technical terms* in QUAID. In addition to word frequency, the QUAID variable computes semantic familiarity using the familiarity rating in Coltheart's (1981) MRC Psycholinguistic Database. I do not include familiarity in the first text feature because this concept is too vague in the psycholinguistic literature and is a rather subjective measure of word difficulty based on ratings of perceived familiarity (e.g., Colombo, Pasini, & Balota, 2006; Rayner & Pollatsek, 2006). Moreover, given that frequency and familiarity are highly correlated (Balota, Yap, & Cortese, 2006), in my view, familiarity can largely be subsumed under frequency, especially if objective up-to-date frequency lists are used in determining word difficulty (see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004).

Second, the feature *complex logical structures* corresponds to QUAID's *working memory overload*. In both instances, this text feature refers to question structures that require respondents to hold a lot of information in mind while processing other information. Thus, such questions quickly overload the working memory capacity of respondents. For example, these structures are involved in hypothetical questions and questions containing numerous logical operators (such as *or)*. I reject the term *working memory overload* because this overload is a result rather than a factor. Arguably, all of the seven text features presented above can result in memory overload because they require respondents to invest considerable cognitive effort to answer the questions. For example, left-embedded syntactic structures (i.e., sentences beginning with many subordinate clauses embedded in the main clause) often result in working memory overload, because they require respondents to hold information from various subordinate clauses in mind before encountering the main verb of the

main clause.[1] Consequently, *working memory overload* is the result of many other text features. Moreover, working memory overload as conceptualized by QUAID does also occur in questions which require the performance of quantitative mental calculations. However, quantitative mental calculations are rather burdensome at the retrieval stage and not at the comprehension stage. All other text features dealt with in this thesis affect the survey response process at the comprehension stage. Trying to establish a taxonomy of psycholinguistic text features which undermine survey question *comprehension,* I refer to this text feature as *complex logical structures*. In the remainder of this section I will discuss these text features in more detail to provide the theoretical background for the empirical studies.

### 3.2.1 Low-frequency words

Ample empirical evidence suggests that the frequency of a word (i.e., the number of times it occurs in large text corpora) has a considerable effect on the effort required to read and comprehend it. People are slower at accessing the meaning of low-frequency words and must invest more cognitive resources to comprehend them in comparison to higher frequency words (e.g., Inhoff & Rayner, 1986; Williams & Morris, 2004). This phenomenon is referred to as the word frequency effect, which has been identified in virtually every measure of word recognition (e.g., naming, Forster & Chambers, 1973; lexical decision, Whaley, 1978; phoneme monitoring, Foss, 1969; eye movements, Rayner & Duffy, 1986). Consequently, low-frequency words such as technical terms, abbreviations, acronyms, and rare words should be avoided in survey questions and replaced by higher frequency synonyms. The following question pair, contrasting a question including a low-frequency word (1a) with the same question including a higher frequency word (1b), illustrates this matter:

(1a) During the last four weeks, how often did you suffer from *somatic* pain?

(1b) During the last four weeks, how often did you suffer from *physical* pain?

[1] Particularly for young or unpracticed readers, this sentence, for example, which is characterized by many clauses encountered before the main verb of the main clause (i.e., *pose*), can pose considerably difficulty.

### 3.2.2 Vague or imprecise relative terms

Vague or imprecise relative terms are predicates whose meanings are relative rather than absolute, such as *often*, *substantially* or *recently* (e.g., Mosier, 1941). They implicitly refer to an underlying continuum, but the points on the continuum may be vague or imprecise. Hence, they can be interpreted in various ways, making it potentially difficult for respondents to extract the meaning intended by the survey designer. Of course, when vague or imprecise terms occur in the response options, their relative position in the list helps to interpret them. In these cases respondents use the pragmatic context, that is, the ordered list of answer options, to assign a meaning to each relative term (Fillmore, 1999). Nevertheless, whenever these terms are presented in the question stem, respondents are likely to have difficulties interpreting them. This is because vague predicates result in sentences which can neither be valued as true or false; they lack the content to allow for an absolute ascription of truth or falsity. For example, compare the vague wording in (2a) to the more concrete wording in (2b):

(2a) Have you *recently* seen a doctor? If yes, please provide the number of visits you paid to the doctor.

(2b) Have you seen a doctor during the last *four weeks*? If yes, please provide the number of visits you paid to the doctor.

Natural languages contain numerous relative terms that are vague or imprecise. The following list, adapted from QUAID (University of Memphis, n.d.) contains some examples:

- Vague frequency terms: *regularly, often, several, frequently, usually*
- Vague temporal terms: *currently, recently, earlier, now*
- Vague intensity terms: *big, large, little, moderate, severe*
- Vague quantification terms: *almost, substantially, worse, certain*

To avoid the vagueness and imprecision caused by relative terms, survey designers are advised to use more concrete terms whenever they ask respondents to place

themselves on a continuum. For example, when they are interested in the frequency of a given behavior, they are advised to specify the reference period and to ask respondents to report in an absolute metric (cf. Fowler, 1995).

### 3.2.3 Vague or ambiguous noun phrases

This term refers to noun phrases, nouns, or pronouns which have an unclear or ambiguous referent. First, noun phrases are potentially problematic if they are abstract (i.e., are hypernyms of numerous hyponyms). A hypernym is a word that encompasses one or more specific words (hyponyms). Consider the following hierarchy: *atmospheric phenomenon* → *storm* → *windstorm* → *hurricane*. In this hierarchy, the superordinate words are more abstract than the subordinate ones they encompass. Every word can be assigned a hypernym value, which is low for abstract words (i.e., abstract words have only few hypernyms) and high for concrete words (i.e., concrete words can be subsumed under numerous hypernyms). By definition, noun phrases with a low hypernym value are vague, and thus should be avoided in survey questions. Consider the following example contrasting a vague term with a low hypernym value (3a) with a more concrete term with a higher hypernym value (3b):

(3a) In your free time, how often do you attend *cultural events*?

(3b) In your free time, how often do you go to the *theater*?

Secondly, ambiguous noun phrases are polysemic, that is, they have a single orthographic form which is associated with multiple senses. Thus, when reading an ambiguous noun phrase, respondents may not immediately know which sense of the word is relevant to the question. Ambiguous words can be divided into balanced ambiguous words such as *straw*, which have two almost equally dominant meanings (1. straw of wheat, 2. straw to suck up a drink), and biased ambiguous words such as *bank*, which have one highly dominant meaning (1. financial institution, 2. river bank). Several studies found that if the preceding context of a biased ambiguous word supports the non-dominant interpretation of the word, then the reading process

is disrupted (*subordinate bias effect*, Duffy, Morris, & Rayner, 1988; Rayner, Pacht, & Duffy, 1994). This finding can be explained by the fact that the context activates the non-dominant meaning while the ambiguous word activates the dominant meaning. In conclusion, even though respondents may use the pragmatic context (i.e., the question text and the answer options) to disambiguate ambiguous noun phrases, biased ambiguous words used in their non-dominant meaning should be avoided in survey questions.

A third instance of ambiguous noun phrases are ambiguous pronouns. Given that the writer is usually not present during reading, there is basically no deictic use of pronouns or adverbs in written communication. Words such as *it, they, here, there* and *this* "always refer anaphorically, that is, to something the writer has previously introduced explicitely or implicitly" (Morgan & Green, 1980, p. 136). Hence, the task of connecting an anaphoric element to its antecedent in the text is central to reading comprehension. When readers come across a pronoun such as *it*, they must identify an antecedent that matches it (*antecedent search*). If there is a considerable distance between the anaphora and the antecedent, fixation durations are longer when the pronoun is encountered (Garrod, Freudenthal, & Boyle, 1994). Similarly, when there are multiple referents that could match the antecedent (4a), the pronoun is ambiguous and antecedent search might take longer. Consider the following example:

(4a) In general, would you say that people should obey the *law* without exception, or are there exceptional occasions on which people should follow their *conscience* even if it means breaking *it*?

(4b) In general, would you say that people should obey the *law* without exception, or are there exceptional occasions on which people should follow their conscience even if it means breaking *the law*?

### 3.2.4 Complex syntax

Syntax can become complex for two reasons: either the structures are ambiguous and lead to a wrong interpretation which has to be corrected by the reader; or they

overload the processing abilities of the reader. In general, readers make sense of the syntactic structure of a sentence by parsing it into its components, that is, by assigning the elements of the surface structure to linguistic categories. According to Just and Carpenter (1980), these processes are carried out immediately as people read a word, a principle they call the *immediacy principle*. As soon as readers see a word, they fit it into the syntactic structure of the sentence. This is due to working memory limitations: postponing the decision would sooner or later overload working memory. Although this strategy is generally useful, in the case of ambiguous syntactic structures it sometimes leads to errors and subsequent reanalyses of the sentences. If later information makes clear that the wrong decision was made, then some backtracking is necessary. This can explain the comprehension difficulties induced by garden path sentences. For example, consider the following garden path prototype:

(5) John hit the girl *with a book* with a bat.

The italicized phrase makes this sentence structurally ambiguous, because it must be attached differently from the reader's initial preference. Obviously, syntactic constructions like these should be avoided in survey questions.

Besides ambiguous structures, a complex syntax can result from propositionally dense sentences. The ease with which readers comprehend the syntactic structure of a sentence heavily depends on the number of propositions it contains (Forster, 1970; Graesser, Hoffman, & Clark, 1980; Kintsch & Keenan, 1973). Kintsch and Keenan (1973), for example, found that the number of propositions influences the time required to read a passage. Consider the following two sentences:

(6) Cleopatra's downfall lay in her foolish trust in the fickle political figures of the Roman world.

(7) Romulus, the legendary founder of Rome, took the women of the Sabine by force.

Even though both sentences have nearly the same number of words, Kintsch and Keenan (1973) have shown that sentence (6) takes longer to read than sentence (7). This result is explained by the fact that (6) is propositionally more complex (eight propositions) than (7), which contains only four propositions[2]. An overflow of propositions in a sentence results in dense noun phrases and dense clauses, which are both difficult to comprehend. A noun phrase is dense if it is supplemented by many adjectives and adverbs. It becomes hard to either understand how the adjectives restrict the noun or to narrow down the precise intended referent of the noun.

Finally, a complex syntax can also result from left-embedded sentences. Left-embedded syntax occurs when readers have to process many clauses, prepositional phrases and qualifiers before they encounter the main verb of the main clause. These constructions require readers to hold a large amount of partially interpreted information in memory before they receive the main proposition. In contrast, sentences with right-branching syntax are easier to process because they first present the main clause (e.g., assertion or question) and subsequently add clauses and phrases that qualify the first clause. Consider the following example:

(8a) How likely is it that if a law was considered by parliament that you considered to be unjust or harmful, you, acting alone or together with others, would *try to do* something against it?

(8b) How likely is it that you, acting alone or together with others, would *try to do* something against a law that was considered by parliament and that you believed to be unjust or harmful?

### 3.2.5 Complex logical structures

Questions with complex logical structures require respondents to remember a large amount of information while simultaneously processing other, new information. Thus, they quickly overload respondents' working memory capacity. For example, hypothetical questions are difficult to process because they are not grounded in the

[2] took[Romulus, women, by force], found[Romulus, Rome], legendary[Romulus], Sabine[women].

real world, requiring respondents to build a mental representation of the hypothetical situation (as unlikely as it may be) and hold it in memory while processing the rest of the question.

Another instance of complex logical structures are questions containing numerous logical operators such as *or*. These questions quickly overload working memory because respondents need to keep track of different options and possibilities. Consider the following example:

(9a) There are many ways people *or* organizations can protest against a government action *or* a government plan they strongly *or* at least somewhat oppose. In this regard, do you think the following should be allowed?
Organizing public meetings to protest against the government.

(9b) There are many ways people *or* organizations can protest against a government action they strongly oppose. In this regard, do you think the following should be allowed?
Organizing public meetings to protest against the government.

A final example of complex logical structures are questions that contain negatives. Answering negative questions is quite difficult for many respondents because they require an exercise in logical thinking (e.g., Fink, 1995; Foddy, 1993; Fowler, 2001). This is particularly obvious in questions that ask respondents to agree or disagree with a negative statement. In these questions, respondents have to disagree with the negative statement in order to express an affirmative opinion. Consider the following question pair:

(10a) Do you agree or disagree with the following statement?
Poorer countries *should not be expected* to make less effort than richer countries to protect the environment.

(10b) Do you agree or disagree with the following statement?
Poorer countries *should be expected* to make less effort than richer countries to protect the environment.

In question (10a), respondents have to disagree that poorer countries should not be expected to make less effort to say that they should be expected to make less effort. According to the standard verification model for statements (Akiyama, Brewer, & Shoben, 1979; Clark & Chase, 1972), negative statements require respondents first to translate the negative into a positive statement, and then to decide whether they agree or disagree with the proposition. Consequently, negative statements require more processing effort than affirmative statements.

### 3.2.6 Low syntactic redundancy

Syntactic redundancy refers to the predictability of the grammatical structure of a sentence (Horning, 1979). The higher the level of syntactic redundancy of a sentence, the quicker and easier one can process and comprehend it. For example, syntactic redundancy can be increased by changing passive sentences to active sentences. In passive constructions the object of an action is turned into the subject of the sentence. Thus, passives emphasize the action rather than the agent responsible for the action, making it harder for the reader to predict the course of action (Forster & Olbrei, 1974). For example:

(11a) Do you agree or disagree with the following statement?

Too much money *is being spent* by the government on assisting immigrants.

(11b) Do you agree or disagree with the following statement?

Government *spends* too much money on assisting immigrants.

Another way to increase syntactic redundancy is by avoiding nominalizations. Nominalizations are verbs that have been transformed into nouns. Spyridakis and Isakson (1998) examined the effect of nominalizations in texts on readers' recall and comprehension and found that those nominalizations that are critical to the meaning of the text should be denominalized to improve readers' recall of the information provided in the document. Even though nominalizations do not necessarily undermine comprehension, there is some evidence that whenever possible, they

should be replaced by active verbs (Coleman, 1964; Duffelmeyer, 1979). The following question alternatives may illustrate this point:

(12a) Do you agree or disagree with the following statement?
Trade unions are important for the job *security* of employees.

(12b) Do you agree or disagree with the following statement?
Trade unions are important to *secure* the jobs of employees.

### 3.2.7 Bridging inferences

Writers do not make explicit everything that they want to communicate in a text. Thus, a text always contains implicit information that readers need to infer from the text. Drawing inferences is generally assumed to be a time-consuming process (Vonk & Noordman, 1990) and numerous psycholinguistic experiments demonstrated that reading time increases with the number of inferences readers need to generate (e.g., Haviland & Clark, 1974; Just & Carpenter, 1980; Myers, Cook, Kambe, Mason, & O'Brien, 2000). In questionnaires, inferences of this sort usually come in the form of bridging inferences, which are required in order to establish coherence between current information and previous information. For example, bridging inferences are required when an introductory sentence precedes the actual question and information from both sentences needs to be connected. The following example illustrates this matter:

(13a) All systems of justice *make mistakes*. What do you think is worse, to convict an innocent person; or to let a guilty person go free?

(13b) All systems of justice *make wrong verdicts*. What do you think is worse, to convict an innocent person; or to let a guilty person go free?

While respondents answering (Q13a) need to infer that by 'making mistakes' the questionnaire designer refers to making wrong judgments, the wording in (Q13b) makes clear that the questions focuses on this one particular instance of judicial error.

# 4 OBJECTIVES OF THE EMPIRICAL STUDIES

The main goals of this thesis are twofold. First, it examines whether several psycholinguistic text features reduce question comprehensibility and increase the cognitive effort required to answer survey questions. Second, it takes a closer look at the ways in which the cognitive burden imposed by less comprehensible survey questions affects response quality and measurement error. If these text features are found to undermine question comprehensibility and to increase measurement error, survey designers may benefit from manuals describing how to avoid them when writing questions.

The two central research questions were examined in three consecutive studies. These studies and their objectives are briefly described in the following subsections.

## 4.1 Study 1

The first experiment was a pilot study designed to obtain first results on whether the seven problematic text features presented in the previous chapter reduce question comprehensibility and increase respondent burden. Therefore, a Web survey experiment was conducted in which respondents ($N$ = 985) were randomly assigned to one of two questionnaire versions: one group received questions that contained one text feature (text feature condition) and the other group received control questions which did not contain the text feature (control condition). As a measure of cognitive burden we collected the response times for the individual questions. Previous research has shown that response times are good indicators of question difficulty: the time it takes respondents to answer a survey question is generally assumed to reflect the cognitive effort that is necessary to arrive at an answer (Bassili, 1996; Draisma & Dijkstra, 2004).

There are two main advantages of response latency analysis in comparison to other question evaluation methods, such as cognitive interviews, interviewer debriefings, question appraisal systems, expert reviews, or behavior codings. First, the measurement of response times is unobtrusive, and hence not affected by the researcher (and the ways in which he or she tests the questions) and the research

context. Second, they are natural byproducts of the question answering process, and hence their validity neither depends on the respondents' ability to verbally express themselves and to communicate their own thoughts nor on the interviewers' or the researchers' experience and expertise in evaluating questions.

If the text features were found to be associated with longer response times, a second objective of the first study was to examine how this additional burden affects the quality of respondents' answers. According to relevance theory (Sperber & Wilson, 1986) and satisficing theory (Krosnick, 1991), survey respondents may not always be able or willing to invest the additional cognitive effort required to comprehend burdensome questions (i.e., to "optimize"). Instead they may try to shortcut the response process and apply response strategies that simplify the survey endeavor (i.e., to "satisfice"). For example, these strategies include saying "don't know" instead of reporting an opinion, selecting the first answer option that seems reasonable, or selecting the midpoint response option. All of these strategies are problematic in surveys as they increase measurement error and produce lower-quality data. Hence, study 1 examined several indicators of survey satisficing, such as very short response times, neutral responses, acquiescence, and primacy effects.

## 4.2 Study 2

In study 2, the findings of study 1 were extended in several ways. First, eye tracking was used as a more direct method to examine whether comprehension is indeed impeded by the seven problematic text features. While response latencies are valuable indicators of the overall cognitive effort required to answer survey questions, they do not enable us to distinguish between the time required to read and understand a question (comprehension stage) and the time it takes to provide an answer (including retrieval, judgment, and response selection). By contrast, recording respondents' eye movements while answering a Web survey allows us to identify the specific parts of the question they struggle with during the comprehension stage. Of course, this does not imply that respondents always perform these cognitive tasks in a sequential order. Sometimes respondents may start to retrieve relevant information while reading and comprehending the question, for

example. Nevertheless, given that eye tracking allows us to examine respondents' fixation times and counts on specific parts of the question, this technique enables us to identify comprehension difficulties with much greater precision than does the collection of response times.

Again, participants ($N$ = 44) were randomly assigned to either a text feature questionnaire or a control questionnaire and we examined whether people fixated longer on questions that include a text feature and required more fixations to process these questions in comparison to control questions. In addition, study 2 examined whether the text features have different effects for different types of questions. While the earlier findings were almost exclusively limited to attitudinal questions, we included a considerable number of behavioral and factual questions in this study. Finally, we analyzed the participants' response times to test whether we are able to replicate the findings of the first study with a considerably smaller sample.

## 4.3 Study 3

Study 3 examined in more detail whether and to what extent the cognitive effort imposed by less comprehensible questions reduces the quality of respondents' answers. In doing so, it extended upon the findings of study 1 in two ways. First, the effects of question comprehensibility on response quality were examined under more realistic conditions. While the questionnaires in the first experiment entirely consisted of either text feature or control question, the questionnaires in study 3 included a large number of filler questions as well. Hence, the ecological validity of the experiment was enhanced in comparison to study 1.

Second, study 3 examined whether the text feature effects are moderated by respondent characteristics such as verbal intelligence and motivation. According to satisficing theory (Krosnick, 1991), the probability that respondents provide low-quality responses is a function of three factors: question difficulty, respondent ability, and respondent motivation. The more difficult a survey question is to understand and to answer, and the lower the respondents' cognitive abilities and motivation, the more likely they are to satisfice in survey. Therefore, study 3 did not only examine whether questions including the problematic text features reduce response quality,

but also whether these effects are more pronounced among respondents with limited verbal skills and among respondents with low motivation to answer surveys.

In this study, respondents were asked to complete two Web surveys at a two-week interval ($N_1$ = 825, $N_2$ = 515). In the first survey, respondents were randomly assigned to either the text feature version or the control version of a questionnaire. Dependent variables in this first survey were number of drop-outs, number of non-substantive responses ("don't know" or skipped questions), and number of neutral (midpoint) responses as indicators of response quality. The second survey asked exactly the same questions as the first one, making it possible to assess the reliability of the responses in both conditions as an additional data quality indicator. In addition to examining the main effect of the text features on response quality, study 3 examined whether there are interaction effects of question comprehensibility with verbal intelligence and/or motivation.

## 4.4 Summary

Table 4.1 provides an overview of the relationship between the two overall research questions, the various dependent and moderator variables of the experiments and the three empirical studies.

The three studies have been published or accepted for publication as:

1. Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24,* 1003-1020. doi: 10.1002/acp.1602

2. Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23,* 361-373. doi: 10.1093/ijpor/edq053

3. Lenzner, T. (in press). Effects of survey question comprehensibility on response quality. *Field Methods*.

Table 4.1. Relationship between research questions, dependent variables, and studies

| | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| Question comprehensibility / Cognitive effort | | | |
| Dependent variables: | | | |
| Response times | X | X | |
| Fixation times | | X | |
| Fixation counts | | X | |
| Response quality / Measurement error | | | |
| Dependent variables: | | | |
| Drop-out rate | X | | X |
| Very short response times | X | | |
| Neutral (midpoint) responses | X | | X |
| Non-substantive responses | | | X |
| Acquiescence | X | | |
| Primacy effects | X | | |
| Over-time consistency | | | X |
| Moderator variables: | | | |
| Verbal intelligence | | | X |
| Motivation | | | X |

# 5 STUDY 1: COGNITIVE BURDEN OF SURVEY QUESTIONS AND RESPONSE TIMES: A PSYCHOLINGUISTIC EXPERIMENT[3]

## 5.1 Research questions

The goal of this study was to test whether seven psycholinguistic text features, which have been found to undermine reading comprehension, reduce the comprehensibility of survey questions and increase the cognitive burden on respondents. Furthermore, the study examined how the cognitive burden imposed by less comprehensible questions affects respondent behavior and the quality of the survey data.

## 5.2 Method

We conducted an online experiment in which respondents were randomly assigned to one of two questionnaire versions. One group ($n$ = 490) received survey questions which contained one of the seven problematic text features (text feature condition) and the other group ($n$ = 495) answered control questions which did not contain the text features (control condition). The presence or absence of a text feature in a question was the main factor in the experiment and operationalized the comprehensibility of survey questions. Dependent variables were response times as measures of cognitive burden and drop-out rate and survey satisficing as measures of response quality.

Response times have received increasing attention in the survey research literature over the last decade (Yan & Tourangeau, 2008) and have been found to be good indicators of question difficulty (Bassili, 1996; Bassili & Scott, 1996; Draisma & Dijkstra, 2004). The time it takes respondents to answer a survey question is generally assumed to reflect the cognitive effort that is necessary to arrive at an

[3] This chapter is based on:
Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24,* 1003-1020. doi: 10.1002/acp.1602
Parts of this chapter were presented at the 7th International Conference on Social Science Methodology (RC33), September 1-5, 2008, Naples, Italy, and at the the 64th Annual AAPOR Conference, May 14-17, 2009, Hollywood, Florida.

answer, that is, it measures the cognitive burden of a question. Consequently, we hypothesized that the text feature questions produce longer response times than the control questions (Hypothesis 1).

The drop-out rate denotes the proportion of the respondents who answer some questions of the survey but do not complete it. In online surveys the drop-out rate can become a substantial problem, especially if the questions are complex or the questionnaire is long (Ganassali, 2008). Survey questions which induce greater cognitive burden are supposed to reduce respondent motivation. Therefore, we hypothesized that the drop-out rate would be higher in the text feature condition than in the control condition (Hypothesis 2).

According to satisficing theory (Krosnick, 1991), the likelihood that respondents provide low-quality data is a function of three factors: question difficulty, respondent ability and respondent motivation. The more difficult a survey question is to understand and to answer, and the lower the respondent's ability and motivation, the more likely satisficing is to occur. We examined several indicators of satisficing in the two conditions (very short response times, neutral responses, acquiescence, and primacy effects) and expected to find more satisficing behavior in the text feature condition (Hypothesis 3).

### 5.2.1 Participants

Participants were randomly drawn from the online access panel Sozioland (Respondi AG). In total, 5000 people were invited and 1445 respondents (28.9%) started the survey. Some participants were ineligible because either German was not their native language ($n$ = 72), problems occurred with their Internet connection ($n$ = 31), they reported having been interrupted or distracted during answering ($n$ = 124), they dropped from the study before answering any substantial questions ($n$ = 71), technical problems prevented the collection of their response times ($n$ = 6), or they did not complete the survey ($n$ = 136)[4]. For response times the upper and lower one percentiles were defined as outliers (Ratcliff, 1993), excluding another 20 respondents and leaving 985 respondents in the analysis. The participants were

[4] Respondents who dropped out before completing the survey were solely considered in the analysis of drop-out rates and excluded from the other analyses.

between 14 and 75 years of age with a mean age of 32 (*SD* = 11.7). After random assignment, the two groups consisted of 244 males and 246 females (text feature condition, *n* = 490) vs. 257 males and 238 females (control condition, *n* = 495). Of these, 65% had received twelve or more years of schooling, 20% had received ten years, and 15% had received nine or less years of schooling. Educational achievement between the two randomized groups did not differ significantly.

### 5.2.2 Questions

With the exception of four questions that were designed by the author (Q5, Q6, Q7, Q21), the questions used in this study were adapted from the International Social Survey Programme (ISSP). The ISSP is a cross-national collaborative programme of social science survey research. Every year a questionnaire for social science research is fielded in 30 to 40 countries. Using ISSP topics allowed us to ask ecologically valid questions which are common in social science research.

In total, the questionnaire contained 28 experimental questions (four questions per text feature) on a variety of topics such as social inequality, national identity, environment, and changing gender roles. Of these questions, 23 were attitudinal questions, 3 were factual questions (Q7, Q12, Q18) and another 2 were behavioral questions (Q5, Q13). The language of the questionnaire was German. We created two versions of each question by manipulating the complexity of one text feature, holding the other linguistic properties constant. The text feature and control questions can be found in the Appendix (Chapter 5.5). The concrete rewriting rules for the text feature questions were as follows:

*1. Low-frequency words:* Replace a higher-frequency word with a low-frequency synonym (Q1, Q3, Q4). Replace a noun with its acronym (Q2).

*2. Vague or imprecise relative terms:* Raise an imprecise relative term out of the response options into the question stem (Q5, Q6). Delete information (such as date) that clarifies a vague temporal term (Q7). Add a vague intensity term to the question (Q8).

*3. Vague or ambiguous noun phrases:* Replace a noun with a pronoun with multiple referents (Q9). Replace a concrete noun with an abstract noun (Q10, Q11). Replace an unambiguous pronoun with an ambiguous pronoun (Q12).

*4. Complex syntax:* Create a left-embedded syntactic structure by moving a subordinate clause from the end of the sentence to the beginning (Q13, Q16). Create a syntactically ambiguous structure (garden-path, Q14). Make a noun phrase dense by modifying it with numerous adjectives (Q15).

*5. Complex logical structures:* Create a hypothetical question (Q17, Q19). Rewrite the question so that it requires a quantitative mental calculation (Q18). Add numerous logical operators such as "or" (Q20).

*6. Low syntactic redundancy:* Nominalize the verb in the question (Q21, Q22). Change an active sentence to a passive sentence (Q23, Q24).

*7. Bridging inferences:* Rewrite the question so that respondents need to draw a bridging inference between an introductory sentence and the actual question (Q25, Q26, Q27, Q28).

An important requirement for the comparability of the questions through response times was to keep them virtually equal in length. Given that more syllables per question require more processing time (Baddeley & Hitch, 1974; McCutchen, Dibble, & Blount, 1994), the question alternatives were constructed so that they did not differ in more than two syllables from each other. The only exception to this rule were questions, in which the control version was longer than the text feature version (Q7, Q10), thus not affecting the response time in favor of our hypotheses.

### 5.2.3 Procedure

The software used in this online study was EFS Survey (Globalpark, 2007), a software for conducting Web-based surveys. We used JavaScript to measure response times. The response time was defined as the time from presenting the question on the screen to the time the final answer was selected using the computer

mouse. The accuracy of this response time measurement was found to be very robust and superior to other possible forms of implementation (Kaczmirek & Faaß, 2008).

Participants were personally invited by e-mail. The first page in the online questionnaire informed about the topics of the survey (politics, society, and environment). Respondents were instructed to read each question in the given order and not to skip questions or to go back to an earlier question. Moreover, they were asked to shut down other applications running in parallel in order to avoid long page loading times. After clicking on a next-button, the first question was presented.

Only one question per screen was displayed and participants had to use the computer mouse to mark their answers. Once an answer was given, participants had to click on a next-button and the next question was presented. The experiment was a randomized trial and participants were randomly assigned to either the questionnaire with text feature questions or to the condition with control questions. First, respondents answered a series of background questions on sex, age, and native language. Then they received the 28 experimental questions in a random sequence to control for question order effects. Finally, they answered additional background questions on education, work status, and the speed of their Internet connection.

## 5.3 Results

### 5.3.1 Response times

Response times were analyzed as indicators of cognitive burden. Because response times do not follow a normal distribution (Ratcliff, 1993) a logarithmic transformation was calculated on the response times to reduce the skewness of the distribution (cf. Fazio, 1990; Yan & Tourangeau, 2008). To control for differences in reading rate between participants, three identical questions were answered by all respondents in the beginning of the survey. The reading rate was computed as an aggregate of these three questions. We analyzed response times on three levels: the overall effect, the effects for each text feature, and the effects for each question.

The overall effect for all text features was analyzed with a one-factor (question comprehensibility: text feature vs. control) analysis of covariance (ANCOVA) with

reading rate[5] as a covariate. The total response time for a respondent during the treatment was the sum over all 28 questions. The total mean response time was 370.3 seconds (*SD* = 150.2) in the text feature condition and 341.5 seconds (*SD* = 146.5) in the control condition. Respondents were significantly faster in responding to clearer formulated questions, $F(1,982) = 20.56$, $p < .001$.

After having confirmed an overall effect of text features, the second level of analysis assesses the relevance of each text feature with regard to the comprehensibility of a question. Each text feature was operationalized with a set of four questions for each group. The impact of each text feature was therefore analyzed in separate general linear models with the corresponding set of four questions each as repeated measures and reading rate as a covariate. The results in Table 5.1 show that six of seven text features significantly account for longer response times: low-frequency words (LFRW), vague or imprecise relative terms (VIRT), complex syntax (CSYN), complex logical structures (CLOG), low syntactic redundancy (LSYR), and bridging inferences (BINF). Only vague or ambiguous noun phrases (VANP) had no effect on response times. Because several tests were conducted, we controlled for α-inflation with the conservative Bonferroni-correction. Here, the threshold of significance for the p-values for two-tailed tests ($\alpha = .05$) is $p \leq .007$.

On the lowest level of analysis, that is, the single questions in the survey, Table 5.2 identifies which items had the highest impact within each text feature. Considering a Bonferroni-correction, 12 out of 28 questions show a significant difference in response times. Summarizing, the interpretation and implications for the construction of questions with regard to response times is as follows. Text features which should be avoided in survey questions are:

- acronyms
- low-frequency terms
- vague quantification terms

[5] We control for the reading rate because it accounts for most of the differences between respondents' response times. The correlation between reading rate and total response time is $r = .49$. The reading rate in this study was measured so that it also includes the time respondents need to read and answer the question (reading rate + response rate). However, to avoid confusion caused by the term "response rate", which can either refer to speed or percentage of survey completions, we use the term reading rate.

Table 5.1. Analysis of response times per text feature

| Text feature | LFRW | VIRT | VANP | CSYN | CLOG | LSYR | BINF |
|---|---|---|---|---|---|---|---|
| Between groups | | | | | | | |
| *F*(df1=1) | 22.97 | 17.05 | 0.18 | 49.03 | 197.57 | 18.00 | 9.40 |
| df2 | 966 | 969 | 973 | 973 | 972 | 972 | 948 |
| *P* | <.001 | <.001 | .668 | <.001 | <.001 | <.001 | .002 |

*Note*. The seven analyses used a general linear model with the corresponding set of four questions each as repeated measures and reading rate as a covariate.

- left-embedded syntactic structures
- ambiguous syntactic structures
- dense noun phrases
- quantitative mental calculations
- hypothetical questions
- numerous logical operators
- nominalizations
- passive constructions
- bridging inferences

Overall, significantly longer response times were found in the text feature condition with regard to all five text features in QUAID (Graesser et al., 2006) except for vague or ambiguous noun phrases. The additionally proposed text features low syntactic redundancy and bridging inferences were also found to increase response times. Furthermore, the analysis per question shows specifically which text features undermine question comprehensibility.

### 5.3.2 Drop-out rates

Drop-out rates were analyzed as a first indicator of response quality. As mentioned above, 136 participants (11.9%) dropped out before completing the survey. The drop-out rates were 13.2% ($n$ = 77) in the text feature condition and 10.6% ($n$ = 59) in the control condition. Drop-out rates did not differ significantly between conditions, $\chi^2 = 2.38$, df = 1, $p = .123$.

Table 5.2. Mean response times between conditions for each question: text feature questions (TF) vs. control questions (Control)

| | Means for raw data in seconds | | Means for log-transformed data | | | | |
|---|---|---|---|---|---|---|---|
| Item | TF | Control | TF | Control | *F*-value (df1=1) | df2 | *p* |
| Low-frequency words | | | | | | | |
| Q 01 low-frequency term | 12.17 | 12.46 | 4.01 | 3.98 | 4.410 | 980 | .036 |
| Q 02 acronym | 11.19 | 9.77 | 3.98 | 3.92 | 22.042 | 979 | <.001* |
| Q 03 low-frequency term | 8.85 | 7.94 | 3.85 | 3.83 | 3.846 | 977 | .050 |
| Q 04 low-frequency term | 22.58 | 17.77 | 4.24 | 4.19 | 16.921 | 975 | <.001* |
| Vague/imprecise relative terms | | | | | | | |
| Q 05 vague quantification term | 7.04 | 5.84 | 3.79 | 3.69 | 63.973 | 981 | <.001* |
| Q 06 imprecise relative term | 9.06 | 10.10 | 3.90 | 3.90 | .133 | 977 | .715 |
| Q 07 vague temporal term | 9.89 | 8.45 | 3.88 | 3.87 | 2.066 | 979 | .151 |
| Q 08 vague intensity term | 5.86 | 6.20 | 3.70 | 3.69 | .870 | 977 | .351 |
| Vague/ ambiguous noun phrases | | | | | | | |
| Q 09 pronoun multiple referents | 12.21 | 12.51 | 4.03 | 4.02 | 1.838 | 980 | .176 |
| Q 10 abstract noun/hypernym | 12.54 | 12.49 | 4.02 | 4.01 | .110 | 981 | .740 |
| Q 11 abstract noun/hypernym | 10.78 | 10.62 | 3.93 | 3.94 | .927 | 976 | .336 |
| Q 12 ambiguous pronoun | 21.67 | 19.75 | 4.26 | 4.23 | 3.491 | 980 | .062 |
| Complex syntax | | | | | | | |
| Q 13 left-embedded syntax | 15.69 | 13.03 | 4.14 | 4.04 | 55.852 | 977 | <.001* |
| Q 14 ambiguous syntax | 10.28 | 8.92 | 3.94 | 3.85 | 40.747 | 980 | <.001* |
| Q 15 dense noun phrase | 12.57 | 11.73 | 4.03 | 3.97 | 16.457 | 981 | <.001* |
| Q 16 left-embedded syntax | 14.87 | 15.05 | 4.11 | 4.10 | .421 | 979 | .517 |
| Complex logical structures | | | | | | | |
| Q 17 hypothetical question | 20.94 | 19.81 | 4.25 | 4.20 | 9.672 | 979 | .002 |
| Q 18 quant. mental calculation | 12.86 | 7.82 | 4.02 | 3.78 | 251.901 | 980 | <.001* |
| Q 19 hypothetical question | 15.41 | 11.29 | 4.10 | 3.95 | 118.967 | 982 | <.001* |
| Q 20 numer. logical operators | 17.18 | 15.27 | 4.18 | 4.10 | 46.244 | 977 | <.001* |
| Low syntactic redundancy | | | | | | | |
| Q21 nominalization | 8.69 | 8.49 | 3.85 | 3.82 | 5.177 | 981 | .023 |
| Q22 nominalization | 10.95 | 9.16 | 3.94 | 3.89 | 14.656 | 979 | <.001* |
| Q23 passive | 10.84 | 10.36 | 3.96 | 3.93 | 6.854 | 977 | .009 |
| Q24 passive | 9.34 | 8.04 | 3.86 | 3.82 | 10.337 | 978 | .001* |
| Bridging inferences | | | | | | | |
| Q25 bridging inference required | 14.48 | 12.83 | 4.10 | 4.02 | 35.127 | 977 | <.001* |
| Q26 bridging inference required | 10.40 | 12.10 | 3.96 | 3.99 | 3.864 | 979 | .050 |
| Q27 bridging inference required | 23.13 | 20.69 | 4.27 | 4.22 | 8.940 | 975 | .003 |
| Q28 bridging inference required | 22.32 | 23.07 | 4.28 | 4.27 | .367 | 958 | .545 |

*Note*. The ANCOVAs were calculated for the logarithmic response times.
* A p-value of .00179 or lower indicates a significant difference for a two-tailed test with Bonferroni correction with $\alpha =$. 05 (.05/28=.00179).

### 5.3.3 Survey satisficing

Survey satisficing was analyzed as a second indicator of response quality. We examined four indicators of satisficing: very short response times, neutral responses, acquiescence, and primacy effects. These analyses were performed on all eligible questions; however, because the study tested a wide range of different question types, the analyses could not be calculated for every question. In some cases, the answers to the two question alternatives were not comparable between conditions because of differences in the response options. For example, one question (Q5) asked respondents to indicate the frequency with which they usually eat meat on the five-point scale *Always-Often-Sometimes-Seldom-Never*. In the alternative, the vague term "seldom" was raised out of the response categories into the question text and consequently, the response options had to be modified (see Appendix). In total, modifications like these occurred in five questions, leaving 23 questions for the analysis of either acquiescence or primacy effects.

The tendency to provide very short response times was estimated by examining the lower five percentile (fastest response times) for the total response time. Among the five percent of participants who provided the shortest total response times ($n$ = 49), 30 respondents were in the control condition and 19 respondents answered text feature questions. The direction of the effect is contrary to hypothesized satisficing behavior, showing more respondents with very short response times in the control condition. This difference was not statistically significant, $\chi^2 = 2.47$, df = 1, $p = .116$.

The propensity to give neutral answers was estimated by calculating item non-response rates[6] and by counting the number of midpoint responses to the 12 question pairs offering a middle category. The item non-response rate was very low with only 125 questions (0.45%) being left unanswered and there was no significant difference in item non-response between the two conditions, $\chi^2 = .15$, df = 1, $p = .696$. However, respondents gave more neutral responses to the 12 questions offering a middle category when answering text feature questions (1445 counts out of 5880

[6] The questions did not include an explicit "don't know" answer category. Respondents who were unwilling to provide an answer were expected to proceed to the next question without clicking on an answer category.

responses) than when answering control questions (1323 counts out of 5940 responses), $\chi^2 = 8.73$, df = 1, $p = .003$.

Acquiescence was analyzed by counting the answers for "somewhat agree" and "strongly agree" with the statements in 12 attitudinal questions. Respondents in the text feature condition did not provide more answers in the acquiescent direction (3134 counts out of 5880 responses) than respondents in the control condition (3212 counts out of 5940 responses), $\chi^2 = .71$, df = 1, $p = .398$.

Finally, to estimate primacy effects for questions without an agree-disagree-scale, we compared the number of responses in which response choices presented in the first half of the answer options were selected. This way, another 11 questions were examined for primacy effects. Again, we found no primacy effects in the text feature condition (1305 counts out of 5390 responses) compared to the control condition (1390 counts out of 5445 responses), $\chi^2 = 2.51$, df = 1, $p = .113$.

## 5.4 Discussion

This study examined how seven psycholinguistic text features affect the cognitive burden of survey questions and the quality of the data these questions obtain. Using response times as measures of the cognitive effort required to answer survey questions, we compared two versions of similar questions in a Web experiment. Additional dependent variables were drop-out rate and survey satisficing behavior to examine the effects of cognitive burden on response quality.

The present findings show a strong support for the relevance of the psycholinguistic text features on respondent burden. First, the overall effect of text features on total response times was highly significant. Secondly, six text features differed significantly between conditions: respondents answering the text feature questions required longer response times. The highest impact (i.e., the most significant effects out of a set of four questions and the largest mean overall differences in response times) was shown for complex syntax and complex logical structures. In addition, survey designers should also avoid asking questions with low frequency words, vague or imprecise relative terms, low syntactic redundancy, and questions that require bridging inferences. The analyses per question show which

instantiations of text features are the most relevant to consider when crafting survey questions and allow for specific guidelines for question wording.

However, for some questions the differences in response times did not reach a significant level. For two questions (Q7, Q10) this might be due to the fact that the control questions contained three and four syllables more than the text feature questions. Question length may thus have suppressed the impact of question comprehensibility on the response times. However, the relevance of vague or ambiguous noun phrases was not confirmed. A possible explanation for this might be that the words associated with this text feature are usually interpreted idiosyncratically by respondents, and thus, do not necessarily require more processing effort.

Response quality was only partially found to be affected by the text features. Contrary to our expectations, higher cognitive burden did not result in higher drop-out. Even though more respondents refused to complete the survey in the text feature condition, the decision to quit answering the survey was not explicitly related to the cognitive burden imposed by the questions. Hence, other features of the questionnaire (e.g., length) may have a stronger influence on survey drop-out than question comprehensibility. Insofar as drop-out is mediated by respondent motivation, it is likely that our sample consisted of highly motivated respondents who would try to complete the survey irrespective of the cognitive burden it imposes. Evidence for this high motivation is, for example, the low initial response rate (28.9%), suggesting that only a small proportion of highly interested respondents started the survey in the first place (cf. Couper, Tourangeau, Conrad, & Crawford, 2004). Moreover, respondents did not receive any incentives and thus agreed to answer the questions for no apparent reward.

Several indicators of response quality were assessed which concern survey satisficing behavior among respondents. Examining four indicators of satisficing (very short response times, neutral responses, acquiescence, and primacy effects), the text feature questions only resulted in more neutral (i.e., midpoint) responses. Again, we believe that this is due to the characteristics of our sample. According to Krosnick (1991), question difficulty may not necessarily instigate satisficing if respondents are

highly motivated or high in cognitive ability. As was mentioned above, the low initial response rate (28.9%) and the low drop-out rate (11.9%) indicate that our sample consisted of highly motivated individuals. Moreover, item non-response was extremely rare in our data, suggesting that most respondents were willing to optimize through the survey. With regard to cognitive ability, 66.9% of the respondents received 12 or more years of schooling, suggesting that higher educated individuals were overrepresented in our sample. Moreover, participants were drawn from an online access panel and were experienced in answering questionnaires (and presumably also in answering poor questionnaires). All in all, we assume that our respondents were both relatively high in motivation and cognitive ability, so that the cognitive burden induced by the survey questions did not affect response quality. Instead, respondents were willing and able to cope with the higher demands of text feature questions while still requiring longer response times.

There are several limitations to this study. First, response times do not enable us to distinguish between the time required to read and understand a question (comprehension stage) and the time it takes to provide an answer (including retrieval, judgment, and response selection). Further research is needed to examine whether the longer response times are indeed induced by comprehension difficulties. Second, participants in this study either received a questionnaire with only poorly formulated questions or a questionnaire with only (relatively) well-formulated questions. While this design allowed us to maximize the treatment effect on our dependent variables and to analyze the text feature effects on survey drop-out, mixing text feature questions and control questions across conditions or including a set of (unproblematic) filler items may have improved the experiment in terms of ecological validity. Third, the text features had only small effects on the quality of responses. Besides adding to respondent burden it is still unclear whether these text features substantially reduce response quality. Further studies are needed to examine this issue and to investigate in more detail whether the question comprehensibility effects on response quality are moderated by respondent characteristics such as cognitive ability and motivation.

## 5.5 APPENDIX: QUESTION WORDINGS

**Questions to compute reading rate (covariate)**

(Q A) Wie stark interessieren Sie sich für Politik?

Answer options:
Sehr stark; Stark; Mittel; Wenig; Überhaupt nicht

(Q B) Es müsste verbindliche internationale Abkommen für den Umweltschutz geben, an die sich Deutschland und andere Länder halten müssen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q C) Wenn die Regierung die Wahl hätte, entweder die Steuern zu senken oder mehr für Sozialleistungen auszugeben, wofür sollte sie sich Ihrer Meinung nach entscheiden?

Answer options:
Die Steuern zu senken, selbst wenn dies bedeutet, dass weniger für Sozialleistungen ausgegeben wird; Mehr für Sozialleistungen auszugeben, selbst wenn dies höhere Steuern bedeutet.

**Low-frequency words (LFRW)**

(Q1) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Nationale **Minoritäten** sollten vom Staat Unterstützung erhalten, damit sie ihre Sitten und Gebräuche erhalten können.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Nationale **Minderheiten** sollten vom Staat Unterstützung erhalten, damit sie ihre Sitten und Gebräuche erhalten können.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q2) Text feature version:
Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **AKW** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Control version:
Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **Atomkraftwerk** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Answer options:
Sehr wahrscheinlich; Wahrscheinlich; Unwahrscheinlich; Sehr unwahrscheinlich

(Q3) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die sozialen **Diskrepanzen** in Deutschland bleiben in Zukunft sicherlich bestehen.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die sozialen **Unterschiede** in Deutschland bleiben in Zukunft sicherlich bestehen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q4) Text feature version:
Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **übervorteilen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Control version:
Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **betrügen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Answer options:
fast immer versuchen, mich zu übervorteilen/betrügen; meistens versuchen, mich zu übervorteilen/betrügen; meistens versuchen, sich mir gegenüber fair zu verhalten; fast immer versuchen, sich mir gegenüber fair zu verhalten

**Vague or imprecise relative terms (VIRT)**

(Q5) Text feature version:
Es kommt **selten** vor, dass ich auf Fleisch beim Essen verzichte.
Stimme zu; Stimme nicht zu

Control version:
Wie häufig verzichten Sie beim Essen auf Fleisch?
Immer; Oft; Manchmal; **Selten**; Nie

(Q6) Text feature version:
Der Einfluss der Gewerkschaften in unserem Land ist **gering**.
Stimme zu; Stimme nicht zu

Control version:
Der Einfluss der Gewerkschaften in unserem Land ist...
Zu groß; Genau richtig; Zu **gering**

(Q7) Text feature version:
Wenn Sie sich einmal an den Börsencrash der **„New Economy“** erinnern. Hatten Sie damals einen privaten Internetzugang?

Control version:
Wenn Sie sich einmal an den Börsencrash der „New Economy“ **im Jahr 2001** erinnern. Hatten Sie damals einen privaten Internetzugang?

Answer options:
Ja; Nein

(Q8) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Deutschland sollte die Einfuhr ausländischer Produkte **wesentlich** beschränken, um seine eigene Wirtschaft zu schützen.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Deutschland sollte die Einfuhr von ausländischen Produkten beschränken, um seine eigene Wirtschaft zu schützen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

**Vague or ambiguous noun phrases (VANP)**

(Q9) Text feature version:
Ganz allgemein gesprochen, würden Sie sagen, dass man **Gesetze** ohne Ausnahme befolgen muss, oder gibt es Ausnahmesituationen, in denen man seinem Gewissen folgen sollte, auch wenn dies bedeutet, **sie** zu übertreten?

Control version:
Ganz allgemein gesprochen, würden Sie sagen, dass man **Gesetze** ohne Ausnahme befolgen muss, oder gibt es Ausnahmesituationen, in denen man seinem Gewissen folgen sollte, auch wenn dies bedeutet, **Gesetze** zu übertreten?

Answer options:
Gesetze ohne Ausnahme befolgen; In Ausnahmesituationen seinem Gewissen folgen

(Q10) Text feature version:
Was meinen Sie: Kann man **Menschen** vertrauen oder kann man im Umgang mit Menschen nicht vorsichtig genug sein?

Control version:
Was meinen Sie: Kann man **fremden Menschen** vertrauen oder kann man im Umgang mit fremden Menschen nicht vorsichtig genug sein?

Answer options:
Fast immer vertrauen; Normalerweise vertrauen; Normalerweise nicht vorsichtig genug sein im Umgang mit fremden Menschen; Fast nie vorsichtig genug sein im Umgang mit fremden Menschen

(Q11) Text feature version:
Glauben Sie, dass es schlimm ist oder nicht schlimm ist, wenn Jugendliche unter 16 Jahren **sexuellen Kontakt** haben?

Control version:
Glauben Sie, dass es schlimm ist oder nicht schlimm ist, wenn Jugendliche unter 16 Jahren **Geschlechtsverkehr** haben?

Answer options:
Immer schlimm; Fast immer schlimm; Nur manchmal schlimm; Nie schlimm

(Q12) Text feature version:
Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die Sie bitten können, wichtige Entscheidungen, zu **Ihren** Gunsten zu beeinflussen?

Control version:
Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die Sie bitten können, wichtige Entscheidungen, zu **deren** Gunsten zu beeinflussen.

Answer options:
Ja, viele; Ja, einige; Ja, aber nur wenige; Nein, niemand

**Complex syntax (CSYN)**

(Q13) Text feature version:
Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, etwas gegen ein Gesetz, das der Bundestag berät und das Sie für ungerecht oder schädlich halten, zu **unternehmen versuchen**?

Control version:
Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, **versuchen** etwas gegen ein Gesetz zu **unternehmen**, das der Bundestag berät und das Sie für ungerecht oder schädlich halten?

Answer options:
Sehr wahrscheinlich; Einigermaßen wahrscheinlich; Nicht sehr wahrscheinlich; Überhaupt nicht wahrscheinlich

(Q14) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen, die in Deutschland geboren sind, **werden** von Zuwanderern Arbeitsplätze **weggenommen**.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zuwanderer **nehmen** den Menschen, die in Deutschland geboren sind, die Arbeitsplätze **weg**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q15) Text feature version:
Ist es gerecht oder ungerecht, dass Menschen mit höherem Einkommen ihren Kindern eine bessere, **berufsbezogene Hochschulausbildung** zukommen lassen können als Menschen mit niedrigerem?

Control version:
Ist es gerecht oder ungerecht, dass Menschen mit einem höheren Einkommen ihren Kindern eine bessere **Ausbildung** zukommen lassen können als Menschen mit einem niedrigeren Einkommen?

Answer options:
Sehr gerecht; Eher gerecht; Weder gerecht noch ungerecht; Eher ungerecht; Sehr ungerecht

(Q16) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Selbst wenn die deutsche Regierung manche Entscheidungen nicht für richtig hält, sollte Deutschland als Mitglied internationaler Organisationen deren Entscheidungen im Allgemeinen **befolgen**.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Deutschland sollte im Allgemeinen als Mitglied internationaler Organisationen deren Entscheidungen **befolgen**, selbst wenn die deutsche Regierung die Entscheidung nicht für richtig hält.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

**Complex logical structures (CLOG)**

(Q17) Text feature version:
**Stellen Sie sich bitte vor** Sie hätten eine erwachsene Tochter, die mit ihrem Partner ein Kind bekommen möchte, aber nicht heiraten will. Was meinen Sie, würden Sie ihr trotzdem dazu raten, zuerst zu heiraten?
Ja, auf jeden Fall; Eher ja; Weder noch; Eher nein; Nein, auf keinen Fall

Control version:
Wie sehr stimmen Sie der folgenden **Aussage** zu: Menschen, die Kinder wollen, sollen vorher heiraten. Bitte antworten Sie auf einer Skala von „Stimme voll und ganz zu“ bis „Stimme überhaupt nicht zu.“
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q18) Text feature version:
Wie viele Stunden verbrachten Sie **im letzten Jahr** ungefähr mit Hausarbeit?
Keine; 1 bis 50 Stunden; 51 bis 100 St.; 101 bis 150 St.; Mehr als 150 St.

Control version:
Wie viele Stunden **pro Woche** verbringen Sie durchschnittlich mit Hausarbeit?
Weniger als 1 Stunde; 1 bis 2 Stunden; 2 bis 3 Stunden; Mehr als 3 Stunden

(Q19) Text feature version:
**Angenommen** Sie wären Bundeskanzler/in und stünden vor dem Problem, dass deutsche Interessen in einer Streitfrage mit denen von anderen Ländern nicht vereinbar sind. Würden Sie sich dafür einsetzen, dass die deutschen Interessen verfolgt werden, auch wenn dies zu Konflikten mit anderen Ländern führt?
Ja, auf jeden Fall; Ja; Nein; Nein, auf keinen Fall

Control version:
Wie sehr stimmen Sie der folgenden **Aussage** zu: Deutschland sollte seine eigenen Interessen verfolgen, selbst wenn dies zu Konflikten mit anderen Ländern führt. Bitte antworten Sie auf der folgenden Skala von „Stimme voll und ganz zu“ bis „Stimme überhaupt nicht zu.“
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q20) Text feature version:
Es gibt viele Möglichkeiten, mit denen einzelne **oder** Gruppen gegen eine Regierungsmaßnahme **oder** ein -vorhaben protestieren können, wenn sie diese Maßnahme entschieden **oder** zumindest ein wenig ablehnen. Sollte in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein?
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren

Control version:
Es gibt viele Möglichkeiten, mit denen einzelne **oder** Vereinigungen gegen eine Regierungsmaßnahme protestieren können, wenn sie diese Maßnahme entschieden ablehnen. Geben Sie bitte an, inwieweit in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein sollte.
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein

**Low syntactic redundancy (LSYR)**

(Q21) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Heutzutage ist es die Aufgabe des Staates, eine **Beschränkung** der Gehälter von Top-Managern zu erwirken.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
In der heutigen Zeit ist es die Aufgabe des Staates, die Gehälter von den Top-Managern zu **beschränken**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu.

(Q22) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind für die **Sicherung** der Arbeitsplätze von Arbeitnehmern wichtig.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind wichtig um die Arbeitsplätze von Arbeitnehmern zu **sichern**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q23) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Von Menschen in reichen Ländern sollte eine zusätzliche Steuer **entrichtet werden**, um Menschen in armen Ländern zu helfen.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen in reichen Ländern sollten eine zusätzliche Steuer **entrichten**, um den Menschen in armen Ländern zu helfen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q24) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Um Zuwanderer zu unterstützen, **wird** vom Staat zu viel Geld **ausgegeben**.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Der Staat **gibt** zu viel Geld **aus**, um die Zuwanderer zu unterstützen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

**Bridging inferences (BINF)**

(Q25) Text feature version:
Auch Gerichte **können sich irren**. Was halten Sie dann für schlimmer...

Control version:
Auch Gerichte **fällen falsche Urteile**. Was halten Sie dann für schlimmer...

Answer options:
eine unschuldige Person zu verurteilen; eine schuldige Person freizusprechen?

(Q26) Text feature version:
In der letzten Zeit wurde in der Öffentlichkeit viel über **Eva Hermanns Buch „Das Eva-Prinzip“** diskutiert. Inwieweit stimmen Sie in diesem Zusammenhang der folgenden Aussage zu oder nicht zu?
Eine berufstätige Mutter kann ein genauso herzliches und vertrauensvolles Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist.

Control version:
In der letzten Zeit wurde in der Öffentlichkeit viel über die **Berufstätigkeit von Frauen** diskutiert. Inwieweit stimmen Sie in diesem Zusammenhang der folgenden Aussage zu oder nicht zu?
Eine berufstätige Mutter kann ein genauso herzliches und vertrauensvolles Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q27) Text feature version:
Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden. **Denken Sie einmal** an Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte.
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Control version:
Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden, **wie zum Beispiel** Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte.
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein

(Q28) Text feature version:
Zurzeit wird in Deutschland viel über die **alternde Gesellschaft und die Altersvorsorge** diskutiert. In diesem Zusammenhang finden Sie unten drei mögliche Maßnahmen, um die Probleme der gesetzlichen Rentenversicherung zu lösen. Wenn Sie sich für eine davon entscheiden müssten, welche würden Sie wählen?
Um die Probleme der gesetzlichen Rentenversicherung zu lösen,...

Control version:
Zurzeit wird in Deutschland viel über **Rente, Rentenfinanzierung und Rentenalter** diskutiert. In diesem Zusammenhang finden Sie unten drei mögliche Maßnahmen, um die Probleme der gesetzlichen Rentenversicherung zu lösen. Wenn Sie sich für eine davon entscheiden müssten, welche würden Sie wählen?
Um die Probleme der gesetzlichen Rentenversicherung zu lösen,...

Answer options:
sollte das Rentenalter erhöht werden; sollten die Rentenbeiträge erhöht werden; sollten die gesetzlichen Renten gekürzt werden

# 6 STUDY 2: SEEING THROUGH THE EYES OF THE RESPONDENT: AN EYE-TRACKING STUDY ON SURVEY QUESTION COMPREHENSION[7]

## 6.1 Research questions

The main goal of this eye-tracking study was to examine whether the seven psycholinguistic text features (see Chapter 3.2) indeed reduce question comprehensibility as suggested by the response time results of the first study (Chapter 5). While response times are valuable indicators of the overall cognitive effort required to answer survey questions, they do not enable us to distinguish between the time required to read and understand a question (comprehension stage) and the time it takes to provide an answer (including retrieval, judgment, and response selection). By contrast, eye movements are direct measures of where and for how long respondents look at while processing survey questions (Galesic & Yan, 2011), and hence they are useful indicators of the on-line processing of information (Rayner, 1998). In addition, the study tested whether the text features differently affect the comprehensibility of different question types (attitudinal, factual, and behavioral questions). Finally, we analyzed participants' response times to examine whether we can replicate the findings of the first study with a much smaller sample.

## 6.2 Method

We conducted an eye-tracking experiment to examine the effects of the seven psycholinguistic text features on survey question comprehension during Web survey completion. If the text features do indeed undermine the survey response process at the comprehension stage, then this should show up in the eye-tracking record in the form of longer fixations and larger numbers of fixations. Our reasoning is based on

[7] This chapter is based on:
Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23,* 361-373. doi:10.1093/ijpor/edq053
Parts of this chapter were presented at the 65th Annual AAPOR Conference, May 13-16, 2010, Chicago, Illinois, at the 12th General Online Research Conference (GOR10), May 26-28, 2010, Pforzheim, Germany, and at the 4th MESS Workshop, August 27-28, 2010, Noordwijk, Netherlands.

two common assumptions about eye movements. The first assumption is that the eye remains fixated on a word as long as it is being processed (*eye-mind assumption*, Just & Carpenter, 1980). Thus, there is a direct link between the time spent fixating on a word and its comprehensibility: difficult words require longer fixations. Second, when larger regions of text such as phrases, clauses, or sentences are difficult to understand, readers are likely to re-fixate earlier words in order to re-read unclear parts of the text, resulting in more fixations on the text (*selective reanalysis hypothesis*, Frazier & Rayner, 1982). Adopting these two assumptions, we examined whether people fixate longer on questions that include a text feature and require more fixations to process the questions.

Respondents were randomly assigned to one of two versions of a Web survey: One group ($n$ = 22) received questions which contained one text feature (text feature condition) and the other group ($n$ = 22) received control questions which did not contain the text feature (control condition). Dependent variables were word/phrase fixation time, question fixation count, and question fixation time as indicators of question comprehensibility. Assuming that the problematic text features induce comprehension difficulties, we hypothesized that respondents would fixate longer on the specific text feature words/phrases, require more fixations to process the questions, and fixate longer on the whole questions in the text feature condition compared to the control condition (Hypothesis 1). While the first study almost only included attitudinal questions, we expected to identify these effects in factual and behavioral questions as well and thus independent of question type (Hypothesis 2).

### 6.2.1 Participants

The study was conducted in June and July 2009 at the Max Planck Institute for Human Development in Berlin, Germany. In total, 49 participants were recruited from the respondent pool maintained by the institute. Technical difficulties made it impossible to accurately record the eye movements of one respondent wearing very thick glasses and another respondent dropped out from the study because of illness. In addition, the recordings of three respondents were of unsatisfactory quality displaying a systematic shift to the line below the one that was fixated. These three

recordings were excluded from the analyses, leaving 44 respondents (22 in each condition) in the experiment. Of these, 61% ($n$ = 27) were female and all were between 19 and 34 years of age with a mean age of 26 ($SD$ = 3.7). All participants had at least 12 years of schooling and 68% ($n$ = 30) were currently enrolled as university students. The native language of all participants was German (the language in which the questionnaires were designed).

### 6.2.2 Eye-tracking equipment

Participants' eye movements were recorded by a Tobii T120 Eye Tracker. In the T120 system, the eye-tracking cameras are integrated into a 17" TFT monitor allowing for unobtrusive recording of respondents' eye movements. The documentation of the T120 describes its accuracy to be within 0.5° with less than 0.3° drift over time and less than 1° due to head motion. It allows for head movement within a 30 × 22 × 30 cm volume centered up to 70 cm from the camera. The sampling rate is 120 Hz, meaning that 120 gaze data points per second are collected for each eye. The accuracy of the T120 was found to be generally sufficient to determine on which words respondents fixate. However, to make sure that all fixations were unequivocally allocated to the words respondents had actually read, we used a larger font size of 18 pixels and double-spaced text with a line height of 50 pixels (see Figure 6.1). Screen resolution was set to 1280 × 1024 pixels. In our analyses, we included all fixations that lasted at least 100 milliseconds and encompassed 20 pixels (about four characters of text) in the Web surveys (see Galesic, Tourangeau, Couper, and Conrad 2008, for similar methodology).

Before analyzing the eye-tracking data, we used the Tobii Studio 2.0.3 software to define so-called "areas of interest" (AOIs). These AOIs were created by drawing rectangles over the specific text feature words/phrases and over the question stems (see Figure 6.1 and Appendix A) to quantify the gaze data on these regions and to obtain our dependent variables (i.e., word/phrase fixation time, question fixation count, and question fixation time).

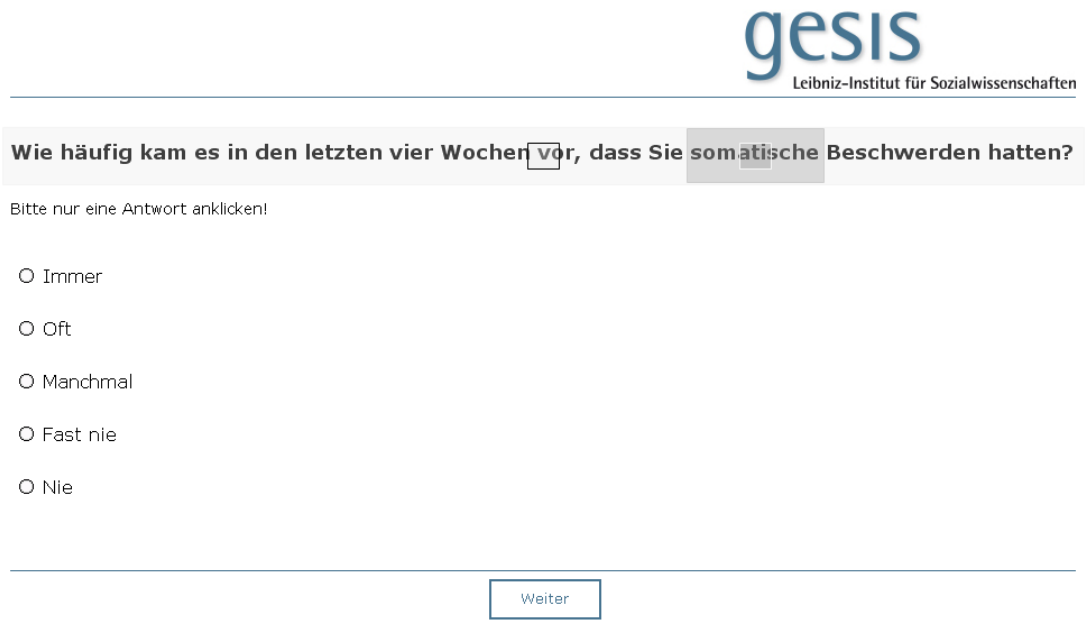


Figure 6.1. Screenshot of a question (Q1) in the text feature condition showing the areas of interest (AOIs) for specific text feature words/phrases (dark grey) and the question stem (light grey).

### 6.2.3 Questions

The questionnaires in both conditions included 28 experimental questions on a variety of topics such as social inequality, environment, health, leisure time, and citizenship (see Appendix A, Chapter 6.5). With the exception of one question that was designed by the author (Q10) the questions were adapted from various existing surveys such as the International Social Survey Programme (ISSP), the German General Social Survey (ALLBUS), and the German Socio-Economic Panel (GSOEP). Each of the seven text features was operationalized by a set of four questions (two attitudinal, one factual, and one behavioral question). Two versions of each question were created by manipulating the characteristic of one text feature according to the following rewriting rules:

*1. Low-frequency words:* Replace a higher-frequency word with a low-frequency synonym (Q1, Q3, Q4). Replace a noun with its acronym (Q2).

*2. Vague or imprecise relative terms:* Raise a vague frequency term out of the response options into the question text (Q5). Replace a concrete temporal term by a vague temporal term (Q6). Add a vague intensity term to the question (Q7). Add a vague quantification term to the question (Q8).

*3. Vague or ambiguous noun phrases:* Replace a noun with a pronoun with multiple referents (Q9). Replace an unambiguous noun by an ambiguous noun (Q10). Replace a concrete noun with an abstract noun (Q11). Replace an unambiguous pronoun with an ambiguous pronoun (Q12).

*4. Complex syntax:* Create a left-embedded syntactic structure by moving a subordinate clause from the end of the sentence to the beginning (Q13, Q16). Create a syntactically ambiguous structure (a so-called garden-path, Q14). Make a noun phrase dense by modifying it with numerous adjectives (Q15).

*5. Complex logical structures:* Create a hypothetical question (Q17, Q19). Add numerous logical operators such as *or* (Q18, Q20).

*6. Low syntactic redundancy:* Nominalize the verb in the question (Q22, Q23). Change an active sentence to a passive sentence (Q21, Q24).

*7. Bridging inferences:* Rewrite the question so that respondents need to draw a bridging inference between an introductory sentence and the actual question (Q25, Q26, Q27, Q28).

The language of the questionnaire was German. The exact wording of the questions is documented in Appendix A (Chapter 6.5).

### 6.2.4 Procedure

The randomized experiment was part of a larger study with several unrelated experiments. The whole study took about two hours of which one hour was devoted to eye tracking. Respondents in this experiment completed the Web survey in about 10 minutes during the first hour of the study. As calibration could decrease in accuracy over time, respondents were recalibrated every 10 to 15 minutes. This was done by a technical assistant who was present in the same room as the respondent

during data collection and ensured adherence to the procedure. The technical assistant was seated at a table next to the respondent and was monitoring his or her eye movements on a separate computer monitor. Respondents were seated in front of the eye tracker so that their eyes were 60 cm from the screen. They were instructed to read at a normal pace while trying to understand the questions as well as they could. After participants had successfully completed a standardized calibration procedure (in which they were asked to fixate on red dots appearing in different regions of the computer screen), they were presented with the welcome page of the Web survey.

Only one question at a time was displayed on the screen. First, participants answered three questions of different length which were identical in both conditions (see Appendix A, Chapter 6.5). These were used to compute the individual reading rate and the fixation rate for every respondent, which were later used as covariates in the analyses to control for interindividual differences. Second, they received the 28 text feature questions or control questions in a random sequence to control for context effects and effects of the position of the questions in the questionnaire. Finally, they answered a series of background questions on sex, age, education, and their native language. After they had completed the survey, the technical assistant recalibrated the eye tracker and started the next experiment. For their participation in the eye tracking part of the study, respondents received a compensation of €10. The study was approved by the Ethics Committee at the Max Planck Institute for Human Development in Berlin, Germany.

## 6.3 Results

The eye-tracking results of this experiment will be reported in terms of word/phrase fixation time, question fixation count, and question fixation time. *Word/phrase fixation time* refers to the total duration of fixations on a specific text feature (e.g., a low-frequency word or an ambiguous noun phrase), including re-readings of these features. *Question fixation count* refers to the sum of fixations respondents made on the question stem (excluding the answer options), again including re-readings. *Question fixation time* corresponds to the total duration respondents fixated on the question stem (excluding the answer options). These three measures are commonly

used to investigate processing difficulty in both word recognition and higher-order comprehension processes (cf. Rayner & Pollatsek, 2006).

Question fixation counts and question fixation times only included fixations on the question stem (excluding the answer options), because we were primarily interested in examining the comprehension stage of the response process. During the comprehension stage, respondents usually fixate on the question stem to find out what the question is about, and hence any comprehension difficulties should show up in the form of longer and higher numbers of fixations in this region. In contrast, while carrying out the remaining tasks of the response process (information retrieval, judgment, formatting, and editing), respondents are more likely to fixate on the answer options. Given that longer fixation times on the answer options can either reflect difficulties in performing these tasks or an optimizing response style, we excluded all fixations on the answers from our analyses.

Ideally, it would have been the case that the specific text feature words or phrases as well as the questions consisted of the same number of characters and the same number of words, respectively. However, in our experiment this was not possible without constructing very artificial questions that respondents would not normally encounter in the real world. Following the recommendations of Ferreira and Clifton (1986), we corrected for differences of word/phrase length and question length between the two question versions by dividing all three eye-tracking parameters by the number of characters in the words/phrases and questions (including character spaces and punctuation marks). Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported in our results.

### 6.3.1 Text features

The effect for each text feature was analyzed in separate general linear models with repeated measures on the four questions per text feature and reading rate or fixation rate as a covariate, respectively. Reading rate and fixation rate were computed from respondents' fixations on three introductory questions. Reading rate refers to the average question fixation time for these three questions; fixation rate refers to the average question fixation count for the three questions.

We controlled for reading rate and fixation rate because both account for most of the differences between respondents' fixation times and numbers of fixations. The correlation between reading rate and the total fixation time for all 28 questions was $r = .80$. The correlation between fixation rate and total number of fixations for all 28 questions was $r = .72$. Reading rate was used as a covariate in analyses of word/phrase fixation times and question fixation times; fixation rate was used as a covariate in analyses of question fixation counts.

Supporting our Hypothesis 1, six out of seven text features were found to undermine survey question comprehension as indicated by the three eye-tracking measures (Table 6.1). First, *word/phrase fixation times* were longer in the text feature condition than in the control condition, indicating that these words were difficult for respondents to comprehend. Statistically significant effects were found for low-frequency words [$F(1, 41) = 21.25$, $p = .0001$, partial $\eta^2 = .34$], vague or imprecise relative terms [$F(1, 41) = 14.19$, $p = .001$, partial $\eta^2 = .26$], vague or ambiguous noun phrases [$F(1, 41) = 8.60$, $p = .005$, partial $\eta^2 = .17$], complex syntax [$F(1, 41) = 8.42$, $p = .006$, partial $\eta^2 = .17$], complex logical structures [$F(1, 41) = 14.90$, $p = .0001$, partial $\eta^2 = .27$], and low-syntactic redundancy [$F(1, 41) = 8.40$, $p = .006$, partial $\eta^2 = .17$]. No significant effects were found for bridging inferences [$F(1, 41) = 0.07$, $p = .787$, partial $\eta^2 = .00$].

Similarly, the *question fixation count* was higher when respondents answered text feature questions, indicating that understanding the question text required re-reading some parts of the question. Again, statistically significant effects were found for low-frequency words [$F(1, 41) = 14.14$, $p = .001$, partial $\eta^2 = .26$], vague or imprecise relative terms [$F(1, 41) = 14.58$, $p = .0001$, partial $\eta^2 = .26$], vague or ambiguous noun phrases [$F(1, 41) = 8.96$, $p = .005$, partial $\eta^2 = .18$] complex syntax [$F(1, 41) = 8.91$, $p = .005$, partial $\eta^2 = .18$], complex logical structures [$F(1, 41) = 12.01$, $p = .001$, partial $\eta^2 = .23$], and low syntactic redundancy [$F(1, 41) = 5.74$, $p = .021$, partial $\eta^2 = .12$]. There was no significant effect of bridging inferences [$F(1, 41) = 0.08$, $p = .783$ partial $\eta^2 = .00$] on the number of fixations respondents made on the question text.

Table 6.1. Mean word/phrase fixation time, question fixation count, and question fixation time for text feature versions and controls

| | Word/phrase fixation time | Question fixation count | Question fixation time |
|---|---|---|---|
| Low-frequency words | 164 | 0.23 | 53 |
| Control | 39 | 0.15 | 30 |
| Vague or imprecise relative terms | 60 | 0.22 | 47 |
| Control | 35 | 0.15 | 30 |
| Vague or ambiguous noun phrases | 103 | 0.18 | 38 |
| Control | 61 | 0.14 | 28 |
| Complex syntax | 77 | 0.21 | 46 |
| Control | 51 | 0.15 | 31 |
| Complex logical structures | 39 | 0.20 | 44 |
| Control | 25 | 0.15 | 31 |
| Low syntactic redundancy | 51 | 0.18 | 38 |
| Control | 34 | 0.14 | 29 |
| Bridging inferences | 48 | 0.17 | 37 |
| Control | 42 | 0.16 | 34 |

*Note*. Fixation times are reported in milliseconds. To control for differences of word/phrase or question length between the two question versions, we divided all three eye-tracking parameters by the number of characters in the question. Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported here. Question fixation counts and question fixation times only refer to fixations on the question text, excluding fixations on answer options.

Finally, *question fixation times* were longer in the text feature condition compared to the control. Similar to the other two eye-tracking parameters, the text feature effects were significant for low-frequency words [$F(1, 41) = 17.66$, $p = .0001$, partial $\eta^2 = .30$], vague or imprecise relative terms [$F(1, 41) = 15.77$, $p = .0001$, partial $\eta^2 = .28$], vague or ambiguous noun phrases [$F(1, 41) = 8.49$, $p = .006$, partial $\eta^2 = .17$], complex syntax [$F(1, 41) = 13.21$, $p = .001$, partial $\eta^2 = .24$], complex logical structures [$F(1, 41) = 12.87$, $p = .001$, partial $\eta^2 = .24$], and low syntactic redundancy [$F(1, 41) = 4.94$, $p = .032$, partial $\eta^2 = .11$]. No significant effects were found for bridging inferences [$F(1, 41) = 0.00$, $p = .956$, partial $\eta^2 = .00$].

The design of our study made it unproductive to analyze text feature effects for every single item. We know from earlier findings (see Chapter 5) that the effects on the item level are of small to medium size (partial $\eta^2 < .13$; Cohen, 1988). A power

analysis (*F* test, $\alpha = .05$) indicated that a minimum sample size of $n = 257$, which is highly uneconomic in eye-tracking studies, would be required to detect any significant effects of medium size on the item level (G*Power 3, Faul, Erdfelder, Lang, & Buchner, 2007). Nevertheless, Appendix B (Chapter 6.6) reports the means for all three eye-tracking measures for every item and thus identifies those items which had the highest impact within each text feature.

### 6.3.2 Question types

After having analyzed the effects for each text feature we examined whether these effects were different for different question types. Each of the seven text features was operationalized with two attitudinal, one factual, and one behavioral question. For two questions, the distinction of question type was a little bit fuzzy. As a result of the text feature manipulation, the hypothetical questions Q17 and Q19 were attitudinal questions in the control condition but could have been conceived as either behavioral or attitudinal questions in the text feature condition. We treated Q17 as an attitudinal question and Q19 as a behavioral question. However, we also analyzed these two questions as if they were other question types but all of our conclusions remained unchanged.

For all three question types we observed similar patterns (Table 6.2). First, respondents had longer *word/phrase fixation times* when answering text feature questions compared to control questions. In analyses of variance with repeated measures on the individual questions per question type and reading rate as a covariate, the between-subjects effects were significant for attitudinal [$F(1, 41) = 24.30$, $p = .0001$, partial $\eta^2 = .37$], factual [$F(1, 41) = 10.83$, $p = .002$, partial $\eta^2 = .21$], and behavioral questions [$F(1, 41) = 17.57$, $p = .0001$, partial $\eta^2 = .30$]. Second, the *question fixation count* was significantly higher for the three question types when respondents answered text feature questions [attitudinal: $F(1, 41) = 9.14$, $p = .004$, partial $\eta^2 = .18$; factual: $F(1, 41) = 9.20$, $p = .004$, partial $\eta^2 = .18$; behavioral: $F(1, 41) = 21.98$, $p = .0001$, partial $\eta^2 = .35$]. And finally, *question fixation times* were significantly longer in the text feature condition for all three question types

Table 6.2. Mean word/phrase fixation time, question fixation count, and question fixation time for text feature versions and controls by question type

| | Word/phrase fixation time | Question fixation count | Question fixation time |
|---|---|---|---|
| Attitudinal (n=14) | | | |
| Text feature questions | 78 | 0.19 | 42 |
| Control | 44 | 0.15 | 32 |
| Factual (n=7) | | | |
| Text feature questions | 84 | 0.20 | 45 |
| Control | 37 | 0.15 | 30 |
| Behavioral (n=7) | | | |
| Text feature questions | 69 | 0.20 | 44 |
| Control | 38 | 0.14 | 28 |

*Note*. Fixation times are reported in milliseconds. To control for differences of word/phrase or question length between the two question versions, we divided all three eye-tracking parameters by the number of characters in the question. Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported here. Question fixation counts and question fixation times only refer to fixations on the question text, excluding fixations on answer options.

[attitudinal: $F(1, 41) = 9.54$, $p = .004$, partial $\eta^2 = .19$; factual: $F(1, 41) = 10.56$, $p = .002$, partial $\eta^2 = .21$; behavioral: $F(1, 41) = 26.92$, $p = .0001$, partial $\eta^2 = .40$].

### 6.3.3 Response times

Similar to the first study (Chapter 5), we used JavaScript to measure response times. Prior to analysis, these were log-transformed to reduce the skewness of the distribution (cf. Fazio, 1990; Yan & Tourangeau, 2008). Response times were analyzed on two levels: for all 28 questions in total and for the four questions per text feature. As was mentioned above, the small sample size of this study made it unproductive to examine response times on the item level. Nonetheless, Appendix C (Chapter 6.7) reports the mean response times between conditions for each question.

The overall effect for all 28 questions (total response time) was analyzed with a one-factor (group: text feature vs. control) analysis of covariance (ANCOVA) with average individual response time[8] as a covariate. The total mean response time was

[8] The average individual response time was computed from respondents' answer times on three introductory questions and hence, resembles the reading rate measure reported in Chapter 5.

Table 6.3. Analysis of response times per text feature

| Text feature | LFRW | VIRT | VANP | CSYN | CLOG | LSYR | BINF |
|---|---|---|---|---|---|---|---|
| Between groups | | | | | | | |
| $F(1, 41)$ | 18.37 | 4.83 | 4.05 | 6.59 | 14.95 | 2.58 | 0.16 |
| $p$ | <.001 | .034 | .51 | .014 | <.001 | .116 | .687 |
| partial $\eta^2$ | .31 | .11 | .09 | .14 | .27 | .06 | .00 |

*Note*. The seven analyses used a general linear model with the corresponding set of four questions each as repeated measures and average individual response time as a covariate.

375.5 seconds (*SD* = 103.8) in the text feature condition and 322.2 seconds (*SD* = 66.9) in the control condition. Respondents were significantly faster in responding to the more comprehensible questions, $F(1,41) = 7.63$, $p < .01$.

On the second level of analysis, the impact of each text feature was analyzed in separate general linear models with the corresponding set of four questions each as repeated measures and average individual response time as a covariate. The results in Table 6.3 show that four out of seven text features significantly accounted for longer response times: low-frequency words (LFRW), vague or imprecise relative terms (VIRT), complex syntax (CSYN), and complex logical structures (CLOG). A marginally significant effect was found for vague or ambiguous noun phrases (VANP) and no significant effect ($p > .05$) for low syntactic redundancy (LSYR) and bridging inferences (BINF). Hence, we were able to partly replicate the response times findings of the first study (Chapter 5) with a considerably smaller sample in the laboratory.

## 6.4 Discussion

Extending earlier research by Graesser et al. (2006) and Lenzner et al. (2010), this study examined whether survey question comprehension is impeded by seven psycholinguistic text features and whether these text features have different effects for different question types (attitudinal, factual, and behavioral questions). Using eye tracking methodology, we examined word/phrase fixation times, question fixation counts, and question fixation times while respondents answered two versions of similar questions (text feature version vs. control) in a Web survey. Moreover, we

partly replicated the response latency findings of the first study.

We found strong evidence that six of these text features reduce question comprehensibility and undermine the survey response process at the comprehension stage. Respondents had longer fixation times and needed more fixations in the text feature questions than in the control questions, indicating that processing of these questions required additional cognitive effort. Significant effects were found for low-frequency words, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, complex logical structures, and low syntactic redundancy. Only bridging inferences were not found to have a detrimental effect on question comprehensibility. In general, bridging inferences are drawn in order to establish coherence between implicit information from an introductory sentence and explicit information from the actual question. The purpose of the introductory sentences is usually to provide a context for the questions, however, understanding (or even reading) these sentences is not a prerequisite for answering the questions (i.e., introductory sentences do not necessarily determine the question focus). Hence, establishing coherence between introductory sentence and actual question is mostly optional rather than mandatory. Our results indicate that bridging inferences may only undermine question comprehension if the introductory sentence contains implicit information which is crucial for understanding and answering the question.

We also found strong support for our second hypothesis that those text features that negatively affect question comprehension do so independent of question type. Similar effects were found for attitudinal, factual, and behavioral questions, namely that respondents required longer fixation times and more fixations when these questions contained a text feature. Hence, the text feature effects can be generalized to all types of questions.

Finally, we were able to partly replicate the response time findings of the first study with a considerably smaller sample in the laboratory. The overall effect of the text features on the total response time was highly significant and four out of seven text features produced significantly longer response times (LFRW, VIRT, CSYN, and CLOG). It seems that the impact of these four text features on question comprehension is quite strong and observable under easily obtainable pretesting

conditions (with regard to sample size). In contrast, the comprehension difficulties imposed by vague or ambiguous noun phrases (VANP), low syntactic redundancy (LSYR) and bridging inferences (BINF) seem to be more subtle and making these difficulties observable via response times may require a considerably larger sample of respondents than the analysis of eye-tracking data.

With regard to the practical implications of using eye-tracking methodology for evaluating survey questions, it is important to note that the interpretation of fixation times and counts is by no means definite. Long fixation times and high numbers of fixations are not problematic per se, but may also indicate an increasing interest in the question or a more conscientious response style. For example, optimizing respondents may require considerable time to select the "optimal" response among the answer options offered. While retrieving relevant information, making a judgment, and formatting and editing the answer, these respondents would fixate longer on the answer options, resulting in a relatively large fixation time on the question as a whole. Hence, the interpretation of fixation times on the answer options and on the question as a whole is very complicated and rather speculative.

By contrast, it is much easier to interpret fixation times and counts on the question stem, excluding the answer options. Respondents fixate on the question stem while trying to understand what the question is about and usually they turn to the answer options as soon as they have retrieved the question's meaning. If the question is incomprehensible, respondents require more time to interpret it and require more fixations to re-read parts of the question to resolve uncertainties (see Rayner, 1998, for a general overview of eye-tracking measures and their interpretation). Thus, comprehension difficulties occurring in survey questions should become apparent in longer and higher numbers of fixations on the question stem. After all, there is no reason why respondents should fixate on this region after having retrieved its meaning (unless something remains unclear). In sum, eye-tracking methodology currently allows us to detect problems occurring at the comprehension stage of the response process only.

There are two limitations to this study. First, our experiment does not examine whether these text features reduce the quality of responses. While we know that

answering questions including the text features requires more time and cognitive energy, it is still unclear whether this additional cognitive effort leads to an increase in measurement error. The first study (Chapter 5) found some negative effects of the text features on response quality (e.g., that they produce more midpoint responses), however, further research is needed to systematically assess their influence on response quality. Second, our sample overrepresents higher educated individuals and therefore is by no means representative of the general population. However, assuming that our participants were comparatively good readers, the text feature effects may even be larger among poorer readers. Hence, we would argue that we can very likely generalize our findings to the broader population.

## 6.5 APPENDIX A: QUESTION WORDINGS

**Overview**

| | | | |
|---|---|---|---|
| QA - QC | Questions to compute reading rate and fixation rate (covariates) | | |
| Q1-Q4 | Low-frequency words (LFRW) | | |
| | Attitudinal: Q2, Q4 | Factual: Q1 | Behavioral: Q3 |
| Q5-Q8 | Vague or imprecise relative terms (VIRT) | | |
| | Attitudinal: Q7, Q8 | Factual: Q6 | Behavioral: Q5 |
| Q9-Q12 | Vague or ambiguous noun phrases (VANP) | | |
| | Attitudinal: Q9, Q10 | Factual: Q12 | Behavioral: Q11 |
| Q13-Q16 | Complex syntax (CSYN) | | |
| | Attitudinal: Q14, Q15 | Factual: Q16 | Behavioral: Q13 |
| Q17-Q20 | Complex logical structure (CLOG) | | |
| | Attitudinal: Q17, Q20 | Factual: Q18 | Behavioral: Q19 |
| Q21-Q24 | Low syntactic redundancy (LSYR) | | |
| | Attitudinal: Q22, Q24 | Factual: Q21 | Behavioral: Q23 |
| Q25-Q28 | Bridging inferences (BINF) | | |
| | Attitudinal: Q25, Q27 | Factual: Q26, | Behavioral: Q28 |

Grey = Areas of interest defined for ‘word/phrase fixation times’

**Questions to compute reading rate and fixation rate (covariates)**

(Q A) Wie stark interessieren Sie sich für Politik?

Answer options:
Sehr stark; Stark; Mittel; Wenig; Überhaupt nicht

(Q B) Es müsste verbindliche internationale Abkommen für den Umweltschutz geben, an die sich Deutschland und andere Länder halten müssen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q C) Wenn die Regierung die Wahl hätte, entweder die Steuern zu senken oder mehr für Sozialleistungen auszugeben, wofür sollte sie sich Ihrer Meinung nach entscheiden?

Answer options:
Die Steuern zu senken, selbst wenn dies bedeutet, dass weniger für Sozialleistungen ausgegeben wird; Mehr für Sozialleistungen auszugeben, selbst wenn dies höhere Steuern bedeutet.

**Low-frequency words (LFRW)**

(Q1) Text feature version:
Wie häufig kam es in den letzten vier Wochen vor, dass Sie **somatische** Beschwerden hatten?

Control:
Wie häufig kam es in den letzten vier Wochen vor, dass Sie **körperliche** Beschwerden hatten?

Answer options:
Immer; Oft; Manchmal; Fast nie; Nie

(Q2) Text feature version:
Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **AKW** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Control:
Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **Atomkraftwerk** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Answer options:
Sehr wahrscheinlich; Wahrscheinlich; Unwahrscheinlich; Sehr unwahrscheinlich

(Q3) Text feature version:
Man kann sich in seiner Freizeit auf unterschiedliche Weise beschäftigen. Bitte geben Sie an, wie häufig Sie Ihre Freizeit damit verbringen, **ersprießliche** Kontakte zu knüpfen.

Control:
Man kann sich in seiner Freizeit auf unterschiedliche Weise beschäftigen. Bitte geben Sie an, wie häufig Sie Ihre Freizeit damit verbringen, **nützliche** Kontakte zu knüpfen.

Answer options:
Sehr oft; Oft; Manchmal; Selten; Nie

(Q4) Text feature version:
Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **übervorteilen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Control:
Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **betrügen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Answer options:
fast immer versuchen, mich zu übervorteilen/betrügen; meistens versuchen, mich zu übervorteilen/betrügen; meistens versuchen, sich mir gegenüber fair zu verhalten; fast immer versuchen, sich mir gegenüber fair zu verhalten

**Vague or imprecise relative terms (VIRT)**

(Q5) Text feature version:
Ich verzichte **selten** beim Essen auf Fleisch.
Stimmt; Stimmt eher; Stimmt eher nicht; Stimmt nicht

Control:
Wie häufig verzichten Sie beim Essen auf Fleisch?
Immer; Manchmal; Selten; Nie

(Q6) Text feature version:
Haben Sie **kürzlich** einen oder mehrere Ärzte aufgesucht? Wenn ja, geben Sie bitte die Zahl der Arztbesuche an.

Control:
Haben Sie in den letzten **vier Wochen** Ärzte aufgesucht? Wenn ja, geben Sie bitte die Zahl der Arztbesuche an.

Answer options:
Kein Arztbesuch, Ein Arztbesuch, Zwei Arztbesuche, Drei oder mehr Arztbesuche

(Q7) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die Politik der Bundesregierung hat **wesentlich** zur momentanen wirtschaftlichen Lage Deutschlands beigetragen.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die Politik der Bundesregierung hat zu der momentanen wirtschaftlichen Lage in Deutschland beigetragen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q8) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Studenten aus einkommensschwachen Familien sollten **beträchtliche** finanzielle Unterstützung vom Staat erhalten.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die Studenten aus einkommensschwachen Familien sollten eine finanzielle Unterstützung vom Staat erhalten.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

**Vague or ambiguous noun phrases (VANP)**

(Q9) Text feature version:
Ganz allgemein gesprochen, würden Sie sagen, dass man Gesetze ohne Ausnahme befolgen muss, oder gibt es Ausnahmesituationen, in denen man seinem Gewissen folgen sollte, auch wenn dies bedeutet, **sie** zu übertreten?

Control:
Ganz allgemein gesprochen, würden Sie sagen, dass man Gesetze ohne Ausnahme befolgen muss, oder gibt es Ausnahmesituationen, in denen man seinem Gewissen folgen sollte, auch wenn dies bedeutet, **Gesetze** zu übertreten?

Answer options:
Man muss sie ohne Ausnahme befolgen; In Ausnahmesituationen seinem Gewissen folgen

(Q10) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Während der **Hochzeit** der Investmentbranche zwischen 2002 und 2007 hätte die Politik stärker in die Wirtschaft eingreifen müssen, um eine Finanzkrise zu verhindern.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Während der **Glanzzeit** der Investmentbranche zwischen 2002 und 2007 hätte die Politik stärker in die Wirtschaft eingreifen müssen, um eine Finanzkrise zu verhindern.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q11) Text feature version:
Wie häufig besuchen Sie in Ihrer Freizeit **kulturelle Veranstaltungen**?

Control:
Wie häufig besuchen Sie in Ihrer Freizeit **Theateraufführungen**?

Answer options:
Mehrmals in der Woche; Mehrmals im Monat; Mehrmals im Jahr oder seltener; Nie

(Q12) Text feature version:
Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die Sie bitten können, wichtige Entscheidungen, zu **Ihren** Gunsten zu beeinflussen?

Control:
Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die Sie bitten können, wichtige Entscheidungen, zu **ihren** Gunsten zu beeinflussen?

Answer options:
Ja, viele; Ja, einige; Ja, aber nur wenige; Nein, niemand

**Complex syntax (CSYN)**

(Q13) Text feature version:
Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, etwas gegen ein Gesetz, das der Bundestag berät und das Sie für ungerecht oder schädlich halten, zu **unternehmen versuchen**?

Control:
Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, **versuchen** etwas gegen ein Gesetz zu **unternehmen**, das der Bundestag berät und das Sie für ungerecht oder schädlich halten?

Answer options:
Sehr wahrscheinlich; Einigermaßen wahrscheinlich; Nicht sehr wahrscheinlich; Überhaupt nicht wahrscheinlich

(Q14) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen, die in Deutschland geboren sind, **werden** von Zuwanderern Arbeitsplätze **weggenommen**.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zuwanderer **nehmen** den Menschen, die in Deutschland geboren sind, die Arbeitsplätze **weg**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q15) Text feature version:
Ist es gerecht oder ungerecht, dass Menschen mit höherem Einkommen ihren Kindern eine bessere, **praxisnähere Hochschul**ausbildung zukommen lassen können als Menschen mit niedrigerem Einkommen?

Control:
Ist es gerecht oder ungerecht, dass Menschen mit einem höheren Einkommen ihren Kindern eine bessere **Ausbildung** zukommen lassen können als Menschen mit einem niedrigeren Einkommen?

Answer options:
Sehr gerecht; Eher gerecht; Weder gerecht noch ungerecht; Eher ungerecht; Sehr ungerecht

(Q16) Text feature version:
Wie häufig sind Sie oder ein Mitglied Ihrer Familie in den letzten fünf Jahren auf öffentliche Bedienstete, die als Gegenleistung für eine Dienstleistung andeuteten, eine Bestechung oder einen Gefallen zu wollen oder dies sogar forderten, **gestoßen**?

Control:
Wie häufig sind Sie oder ein Mitglied Ihrer Familie in den letzten fünf Jahren auf öffentliche Bedienstete **gestoßen**, die als Gegenleistung für eine Dienstleistung andeuteten, eine Bestechung oder einen Gefallen zu wollen oder dies sogar forderten?

Answer options:
Sehr oft; Relativ oft; Manchmal; Selten; Nie

**Complex logical structures (CLOG)**

(Q17) Text feature version:
**Stellen Sie sich bitte vor** Sie hätten eine erwachsene Tochter, die mit ihrem Partner ein Kind bekommen möchte, aber nicht heiraten will. Was meinen Sie, würden Sie ihr trotzdem dazu raten, zuerst zu heiraten?
Ja, auf jeden Fall; Eher ja; Weder noch; Eher nein; Nein, auf keinen Fall

Control:
Wie sehr stimmen Sie der folgenden **Aussage** zu: Menschen, die Kinder wollen, sollen vorher heiraten. Bitte antworten Sie auf einer Skala von „Stimme voll und ganz zu“ bis „Stimme überhaupt nicht zu“.
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q18) Text feature version:
Wie viele Erwachsene **und** Kinder leben **außer** Ihnen selbst in Ihrer Wohnung **oder** Ihrem Haushalt?

Control:
Wie ist das bei Ihnen, wie viele Personen leben insgesamt in Ihrer Wohnung **oder** Ihrem Haushalt?

Answer options:
(keine), 1,2,3,4,5,6 oder mehr

(Q19) Text feature version:
**Angenommen** Sie wären Bundeskanzler und stünden vor dem Problem, dass deutsche Interessen in einer Streitfrage mit denen anderer Länder nicht vereinbar sind. Würden Sie sich dafür einsetzen, dass die deutschen Interessen verfolgt werden, auch wenn dies zu Konflikten mit anderen Ländern führt?
Ja, auf jeden Fall; Ja; Nein; Nein, auf keinen Fall

Control:
Wie sehr stimmen Sie der folgenden **Aussage** zu oder nicht zu: Deutschland sollte seine eigenen Interessen verfolgen, selbst wenn dies zu Konflikten mit anderen Ländern führt. Bitte beantworten Sie diese Frage auf der folgenden Skala von „Stimme voll und ganz zu“ bis „Stimme überhaupt nicht zu“.
Stimme voll und ganz zu; Stimme zu; Stimme nicht zu; Stimme überhaupt nicht zu

(Q20) Text feature version:
Es gibt viele Möglichkeiten, mit denen einzelne **oder** Gruppen gegen Regierungsmaßnahmen **oder** Regierungsvorhaben protestieren können, wenn sie diese entschieden **oder** zumindest ein wenig ablehnen. Sollte in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein?
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren.

Control:
Es gibt verschiedene Möglichkeiten, mit denen einzelne **oder** Vereinigungen gegen eine Regierungsmaßnahme protestieren können, wenn sie diese Maßnahme entschieden ablehnen. Geben Sie bitte an, inwieweit in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein sollte:
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren.

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein

**Low syntactic redundancy (LSYR)**

(Q21) Text feature version:
In welchem Maß **werden** Sie durch Ihre Gesundheit daran **gehindert**, Ihre Freizeit so zu gestalten, wie Sie dies gerne tun würden?

Control:
In welchem Maß **hindert** Sie Ihre Gesundheit daran, Ihre Freizeit in der Weise zu gestalten, wie Sie dies gerne tun würden?

Answer options:
In sehr hohem Maß; In hohem Maß; Bis zu einem gewissen Maß; Überhaupt nicht

(Q22) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind für die **Sicherung** der Arbeitsplätze von Arbeitnehmern wichtig.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind wichtig um die Arbeitsplätze von Arbeitnehmern zu **sichern**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

(Q23) Text feature version:
Wie häufig verbringen Sie Ihre Freizeit mit dem **Erwerb** neuer Kenntnisse?

Control:
Wie häufig verbringen Sie Ihre Freizeit damit, neue Kenntnisse zu **erwerben**?

Answer options:
Sehr oft; Oft; Manchmal; Selten; Nie

(Q24) Text feature version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zu viel Geld **wird** vom Staat **ausgegeben**, um Zuwanderer zu unterstützen.

Control:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Der Staat **gibt** zu viel Geld **aus**, um die Zuwanderer zu unterstützen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu

**Bridging inferences (BINF)**

(Q25) Text feature version:
Auch Gerichte **können sich irren**. Was halten Sie dann für schlimmer...

Control:
Auch Gerichte **fällen falsche Urteile**. Was halten Sie dann für schlimmer...

Answer options:
eine unschuldige Person zu verurteilen; eine schuldige Person freizusprechen?

(Q26) Text feature version:
Von Arbeitnehmern wird heutzutage immer mehr **Mobilität** gefordert. Wie ist das bei Ihnen? Wie viele Nächte haben Sie letztes Jahr nicht zu Hause verbracht, weil Sie auf **Geschäftsreise** waren?

Control:
Von Arbeitnehmern wird heutzutage immer mehr **Reisebereitschaft** gefordert. Wie ist das bei Ihnen? Wie viele Nächte haben Sie letztes Jahr nicht zu Hause verbracht, weil Sie auf **Geschäftsreise** waren?

Answer options:
Ich war nicht über Nacht fort; 1-5 Nächte; 6-10 Nächte; 11-20 Nächte; 21-30 Nächte; Mehr als 30 Nächte

(Q27) Text feature version:
Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden. **Denken Sie einmal** an Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte:
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Control:
Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden, **wie zum Beispiel** Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte:
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein

(Q28) Text feature version:
Manche Menschen sind bereit, ihren Körper nach ihrem Tod der **Medizin zur Verfügung** zu stellen. Wären Sie bereit, nach Ihrem Tod ein **Organ zu spenden**?

Control:
Manche Menschen sind bereit, nach ihrem Tod **Organe zu spenden**. Wie ist das bei Ihnen – wären Sie dazu bereit, nach Ihrem Tod ein **Organ zu spenden**?

Answer options:
Ja, ganz bestimmt; Ja, wahrscheinlich; Nein, wahrscheinlich nicht; Nein, bestimmt nicht

## 6.6 APPENDIX B: GAZE DATA BY ITEM

Mean word/phrase fixation time, question fixation count and question fixation time for text feature versions and controls by item

| | Word/phrase fixation time | | Question fixation count | | Question fixation time | |
|---|---|---|---|---|---|---|
| Item | TF | Control | TF | Control | TF | Control |
| Low-frequency words | | | | | | |
| Q 01 low-frequency term | 276 | 23 | 0.32 | 0.13 | 77 | 52 |
| Q 02 acronym | 131 | 33 | 0.21 | 0.12 | 45 | 26 |
| Q 03 low-frequency term | 177 | 65 | 0.17 | 0.13 | 42 | 28 |
| Q 04 low-frequency term | 71 | 35 | 0.22 | 0.20 | 49 | 42 |
| Vague or imprecise relative terms | | | | | | |
| Q 05 vague frequency term | 99 | 43 | 0.35 | 0.17 | 78 | 38 |
| Q 06 vague temporal term | 44 | 27 | 0.15 | 0.12 | 33 | 23 |
| Q 07 vague intensity term | 58 | 52 | 0.19 | 0.17 | 40 | 35 |
| Q 08 vague quantification term | 40 | 19 | 0.17 | 0.12 | 36 | 23 |
| Vague or ambiguous noun phrases | | | | | | |
| Q 09 pronoun with multiple referents | 58 | 40 | 0.15 | 0.13 | 32 | 27 |
| Q 10 biased ambiguous noun | 224 | 88 | 0.21 | 0.15 | 46 | 31 |
| Q 11 abstract noun/hypernym | 38 | 25 | 0.18 | 0.11 | 38 | 22 |
| Q 12 ambiguous pronoun | 94 | 89 | 0.16 | 0.15 | 34 | 32 |
| Complex syntax | | | | | | |
| Q 13 left-embedded syntax | 43 | 36 | 0.21 | 0.17 | 46 | 35 |
| Q 14 ambiguous syntactic structure | 172 | 93 | 0.21 | 0.14 | 46 | 27 |
| Q 15 dense noun phrase | 51 | 38 | 0.17 | 0.14 | 35 | 28 |
| Q 16 left-embedded syntax | 43 | 35 | 0.24 | 0.17 | 56 | 35 |
| Complex logical structures | | | | | | |
| Q 17 hypothetical question | 28 | 19 | 0.17 | 0.11 | 46 | 32 |
| Q 18 numerous logical operators | 51 | 21 | 0.19 | 0.15 | 54 | 41 |
| Q 19 hypothetical question | 35 | 28 | 0.21 | 0.15 | 40 | 29 |
| Q 20 numerous logical operators | 40 | 31 | 0.24 | 0.20 | 36 | 22 |
| Low syntactic redundancy | | | | | | |
| Q21 passive | 38 | 29 | 0.17 | 0.15 | 36 | 29 |
| Q22 nominalization | 61 | 21 | 0.17 | 0.14 | 39 | 29 |
| Q23 nominalization | 51 | 37 | 0.17 | 0.12 | 38 | 25 |
| Q24 passive | 52 | 47 | 0.19 | 0.16 | 39 | 33 |
| Bridging inferences | | | | | | |
| Q25 bridging inference required | 79 | 62 | 0.18 | 0.17 | 39 | 39 |
| Q26 bridging inference required | 43 | 38 | 0.17 | 0.16 | 35 | 32 |
| Q27 bridging inference required | 31 | 33 | 0.22 | 0.21 | 46 | 43 |
| Q28 bridging inference required | 38 | 35 | 0.13 | 0.11 | 26 | 22 |

*Note*. Fixation times are reported in milliseconds. To control for differences of word/phrase or question length between the two question versions, we divided all three eye-tracking parameters by the number of characters in the question. Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported here. Question fixation counts and question fixation times only refer to fixations on the question text, excluding fixations on answer options.

## 6.7 APPENDIX C: RESPONSE TIMES BY ITEM

Mean response times between conditions for each question: text feature questions (TF) vs. control questions (Control)

| | Means for raw data in seconds | | Means for log-transformed data | |
|---|---|---|---|---|
| Item | TF | Control | TF | Control |
| Low-frequency words | | | | |
| Q 01 low-frequency term | 13.53 | 6.93 | 4.04 | 3.81 |
| Q 02 acronym | 12.82 | 9.16 | 4.06 | 3.94 |
| Q 03 low-frequency term | 12.59 | 9.35 | 4.08 | 3.96 |
| Q 04 low-frequency term | 21.32 | 17.94 | 4.28 | 4.22 |
| Vague or imprecise relative terms | | | | |
| Q 05 vague frequency term | 8.23 | 5.57 | 3.88 | 3.73 |
| Q 06 vague temporal term | 9.00 | 8.38 | 3.91 | 3.90 |
| Q 07 vague intensity term | 12.16 | 11.94 | 4.04 | 4.03 |
| Q 08 vague quantification term | 10.78 | 9.45 | 4.01 | 3.95 |
| Vague or ambiguous noun phrases | | | | |
| Q 09 pronoun with multiple referents | 12.81 | 13.29 | 4.09 | 4.09 |
| Q 10 biased ambiguous noun | 16.54 | 11.79 | 4.19 | 4.03 |
| Q 11 abstract noun/hypernym | 8.37 | 6.24 | 3.86 | 3.77 |
| Q 12 ambiguous pronoun | 21.15 | 22.53 | 4.31 | 4.32 |
| Complex syntax | | | | |
| Q 13 left-embedded syntax | 15.18 | 12.76 | 4.16 | 4.08 |
| Q 14 ambiguous syntactic structure | 12.14 | 9.89 | 4.05 | 3.98 |
| Q 15 dense noun phrase | 12.62 | 10.20 | 4.08 | 4.00 |
| Q 16 left-embedded syntax | 20.79 | 15.18 | 4.24 | 4.15 |
| Complex logical structures | | | | |
| Q 17 hypothetical question | 14.19 | 9.28 | 4.13 | 3.95 |
| Q 18 numerous logical operators | 8.34 | 5.51 | 3.88 | 3.72 |
| Q 19 hypothetical question | 19.03 | 17.00 | 4.26 | 4.17 |
| Q 20 numerous logical operators | 22.56 | 19.85 | 4.33 | 4.29 |
| Low syntactic redundancy | | | | |
| Q21 passive | 8.32 | 8.36 | 3.90 | 3.89 |
| Q22 nominalization | 10.28 | 9.85 | 3.98 | 3.94 |
| Q23 nominalization | 6.36 | 6.38 | 3.78 | 3.77 |
| Q24 passive | 11.07 | 8.63 | 4.01 | 3.90 |
| Bridging inferences | | | | |
| Q25 bridging inference required | 12.61 | 12.79 | 4.06 | 4.08 |
| Q26 bridging inference required | 13.89 | 14.06 | 4.10 | 4.12 |
| Q27 bridging inference required | 20.64 | 21.79 | 4.29 | 4.31 |
| Q28 bridging inference required | 8.21 | 8.10 | 3.89 | 3.88 |

# 7 STUDY 3: EFFECTS OF SURVEY QUESTION COMPREHENSIBILITY ON RESPONSE QUALITY[9]

## 7.1 Research questions

Even though the earlier two studies (Chapter 5 and Chapter 6) provided strong empirical evidence that most of the psycholinguistic text features reduce question comprehensibility, only limited evidence was found that this also results in poorer response quality. The aim of this study was to take a closer look at the ways in which question comprehensibility, or more specific, the effort required to comprehend survey questions, affects response quality. In particular, the study examined whether the questions including these text features increase breakoff rates, increase the number of non-substantive and neutral responses, and lower reliability. Moreover, it examined whether the effects of question comprehensibility on response quality are moderated by respondent characteristics such as verbal intelligence and motivation.

## 7.2 Method

To examine the effects of these psycholinguistic text features on response quality, an experiment was conducted in which respondents were asked to complete two Web surveys during March and April 2010. In the first survey, they were randomly assigned to one of two questionnaire versions: a questionnaire including the seven problematic text features (text feature condition) or a questionnaire that did not include any questions with such features (control condition).

Dependent variables in this first survey were breakoff rates, number of non-substantive responses ("Don't know's" or skipped questions), and number of neutral responses (midpoint responses) as response quality indicators (cf. Galesic, 2006; Knäuper, Belli, Hill, & Herzog, 1997; Velez & Ashworth 2007). The rationale

[9] This chapter is based on:
Lenzner, T. (in press). Effects of survey question comprehensibility on response quality. *Field Methods*.
Parts of this chapter were presented at the 17th ISA World Congress of Sociology, July 11-17, 2010, Gothenburg, Sweden, and at the 13th General Online Conference (GOR11), March 14-16, 2011, Düsseldorf, Germany.

behind choosing these indicators was that dropping out of the survey or providing non-substantive or neutral responses are strategies by which respondents can simplify the survey endeavor if they are unable or unwilling to invest the cognitive effort required for answering incomprehensible questions (cf. Krosnick, 1991).

Assuming that low question comprehensibility reduces response quality, I expected to find more breakoffs, more non-substantive responses, and more neutral responses in the text feature condition than in the control condition (Hypothesis 1a). Moreover, according to satisficing theory (Krosnick, 1991), I hypothesized that these effects would be greater among respondents low in verbal intelligence (i.e., cognitive ability) and/or motivation (Hypothesis 1b).

Respondents who completed the first survey were re-invited to participate in a second Web survey two weeks after the initial invitation. This second survey asked exactly the same questions as the first one, making it possible to assess the reliability of the responses in both conditions. By comparing the answers given in the first Web survey with the answers given in the second Web survey, an index of over-time consistency could be calculated (Krosnick et al., 2002; Poe, Seeman, McLaughlin, Mehl, & Dietz, 1988). Higher over-time consistency is an indicator of higher reliability and thus superior response quality. Assuming that responses to the text feature questions are inaccurate, I hypothesized that the over-time consistency would be lower in the text feature condition than in the control condition (Hypothesis 2a) and that this effect would be more pronounced among respondents low in verbal intelligence and/or motivation (Hypothesis 2b).

### 7.2.1 Respondents

Respondents were recruited from the German nonprobability online panel Sozioland (Respondi AG). Members of this panel have signed up online to receive invitations for surveys on all kinds of topics covering society, media, health, and politics. For participation in this survey, panelists did not receive any incentives. Of the 7581 panel members who were invited, 1195 participated in the first Web survey. Some respondents were excluded from the dataset because they either finished the survey after breakoff ($n$ = 12), dropped out of the study before answering any experimental

question ($n$ = 152), reported having been interrupted or distracted during answering ($n$ = 133), clicked through the survey without answering ("lurkers", $n$ = 1), cheated on the WOCT vocabulary test (claiming to know the meaning of two or three fake words or skipping the test, $n$ = 8; see next section for a description of the test), or they did not complete the survey ($n$ = 64),[10] leaving 825 respondents in the analysis and resulting in a response rate of 10.9% (AAPOR RR1). Of these, 52.0% were female and 48.0% were male; 58.1% had received 12 or more years of schooling, 33.2% had received 10 years, and 8.7% had received 9 or less years of schooling. Respondents were between 16 and 77 years of age, with a mean age of 42 ($SD$ = 13.3). Following the random assignment, the two groups consisted of 407 respondents in the text feature condition and 418 respondents in the control condition.

These 825 respondents were re-invited to answer the second online questionnaire. In total, 515 (62.4%) respondents completed this second survey, allowing for the calculation of over-time consistency estimates for 248 respondents in the text feature condition and for 267 respondents in the control condition. Respondents in the second survey were between 16 and 77 years of age with a mean age of 44 ($SD$ = 12.5) and 51.3% were female. A total of 56.8% of the participants had received 12 or more years of schooling, 33.8% had received 10 years, and 9.4% had received 9 or less years of schooling. In both surveys, respondents in the two conditions did not differ with regard to gender, age, and educational attainment.

### 7.2.2 Instruments

The questionnaires in both surveys included 60 questions on various topics, covering the environment, health, leisure, role of government, national identity, and social inequality (ten questions for each topic). With the exception of one question designed by the author, the questions were taken from the International Social Survey Program (ISSP), the German General Social Survey (ALLBUS), and the German Socio-Economic Panel (GSOEP). To examine the effects of the seven psycholinguistic text

[10] Respondents who dropped out before completing the survey were only considered in the analysis of breakoff rates and excluded from the other analyses.

features on response quality, 28 (four questions per text feature) of the 60 questions were experimentally manipulated so that they contained a problematic text feature in one condition (text feature version) but not in the other (control version). The remaining 32 questions were used as filler items and were asked in the original wording. The experimental questions were constructed according to the following rewriting rules:

*1. Low-frequency words:* Replace a noun with its acronym (Q1). Replace a higher-frequency word with a low-frequency synonym (Q16, Q22, Q42).

*2. Vague or imprecise relative terms:* Raise a vague frequency term out of the response options into the question text (Q3, Q4). Replace a concrete temporal term by a vague temporal term (Q12). Add a vague quantification term to the question (Q44).

*3. Vague or ambiguous noun phrases:* Replace a concrete noun with an abstract noun (Q19, Q24). Replace an unambiguous pronoun with an ambiguous pronoun (Q38). Replace an unambiguous noun by an ambiguous noun (Q56).

*4. Complex syntax:* Make a noun phrase dense by modifying it with numerous adjectives (Q13). Create a left-embedded syntactic structure by moving a subordinate clause from the end of the sentence to the beginning (Q39, Q51). Create a syntactically ambiguous structure (a so-called garden-path, Q47).

*5. Complex logical structures:* Add a superfluous negative to the question (Q9). Add numerous logical operators such as *or* (Q21, Q32) Create a hypothetical question (Q53).

*6. Low syntactic redundancy:* Change an active sentence to a passive sentence (Q20, Q45). Nominalize the verb in the question (Q26, Q57).

*7. Bridging inferences:* Rewrite the question so that respondents need to draw a bridging inference between an introductory sentence and the actual question (Q33, Q34, Q41, Q59).

The language of the questionnaire was German. The exact wording of the experimental and filler questions is documented in Appendix A (Chapter 7.5).

To measure respondents' verbal intelligence, I administered an adapted version of the German *Vocabulary Test* (WST, Schmidt & Metzler, 1992). In the original version, the WST comprises 42 word sequences, each containing one real word (the target word) and five meaningless words. Participants are instructed to indicate which word in each sequence is the real word. For this study, the WST was modified so that it could be efficiently administered in a Web survey. The modified version (WSTmod) included 15 target words of variable word difficulty. For every word, respondents indicated on a two-point scale (yes/no) whether they knew the meaning of the word and could "explain it to someone else" (see Appendix B, Chapter 7.6). Verbal intelligence test scores were obtained by summing up the number of positive responses to the 15 words, and hence could range from 0 to 15 ($M$ = 11.43, $SD$ = 2.13).

To assess the convergent validity of the WSTmod, I correlated respondents' scores on the WSTmod with their scores on a second vocabulary test (WOCT, Ziegler & Kemper, 2010, see Appendix B). Respondents' scores on both measures were highly correlated ($r$ = .64, $p$ < .001). The verbal intelligence scores were moderately correlated with education ($r$ = .37, $p$ < .001), however, earlier research has shown that education is not a good proxy measure for cognitive ability among Web survey respondents (cf. Peytchev, 2009). Hence, I found it important to include this more direct measure of respondents' verbal intelligence in the questionnaire.

A potential problem of the WSTmod is that its yes/no answer format is prone to socially desirable responding. However, respondents' WSTmod scores were not correlated ($r$ = .05, p > .05) with respondents' scores on a social desirability index (composed of items adapted from Paulhus, 1991; Stöber, 1999, see Appendix C, Chapter 7.7). Hence, the WSTmod was an acceptable measure of respondents' verbal intelligence.

As indicators of respondents' motivation to answer survey questions, I measured their *need for cognition* (Cacioppo & Petty, 1982) and *need to evaluate* (Jarvis & Petty, 1996). While need for cognition (NFC) is a measure of how much people

enjoy thinking and performing effortful mental exercises, need to evaluate (NTE) is a measure of how opinionated people are and how willingly they engage in evaluation. People who are low in NFC and/or NTE are presumably more susceptible to satisfice in surveys than those high in these traits (cf. Krosnick, 1991; Toepoel, Vis, Das, & van Soest, 2009). Need for cognition and need to evaluate are usually measured with 36 and 16 items, respectively. For reasons of efficiency, however, I selected 5 items of the German NFC scale (Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994) and 6 items of the German NTE scale (Collani, 2009) on basis of their factor loadings, discrimination power, and face validity (see Appendix D, Chapter 7.8). The raw scores of both scales were combined to calculate an average index of respondent motivation (MOT, Cronbach's $\alpha = .75$).

### 7.2.3 Procedure

In total, the first survey consisted of 122 items with approximately 60% of the items presented on a separate screen.[11] All items were closed-end, requiring respondents to mark their answers by clicking on a radio button. First, respondents completed the WSTmod and the WOCT vocabulary tests, each consisting of 15 words of different word frequency. Then they answered four background questions on gender, age, education, and native language, followed by the NFC and NTE items, as well as three questions on political interest, international environmental laws, and social benefits. Subsequently, respondents were randomly assigned to either the text feature or the control condition. In both conditions, respondents answered a total of 60 questions in randomly-ordered blocks of thematically related questions. Of these 60 questions, 28 were experimentally manipulated so that they contained a text feature in the text feature condition but not in the control condition. Finally, respondents answered 11 social desirability items (adapted from Paulhus, 1991; Stöber, 1999) and three questions on Web survey administration and evaluation (problems with internet connection, interruption or distraction during answering, importance of surveys for

[11] Grids were used for administering the two vocabulary tests (15 items per screen), the need for cognition and need to evaluate scales (5 and 6 items per screen, respectively), and the social desirability items (5 and 6 items per screen, respectively). All of the experimental questions were presented on separate screens.

society). On average, respondents in the text feature and control condition completed the first Web survey in 19.8 minutes (*SD* = 9.3) and 18.8 minutes (*SD* = 7.4), respectively. To answer the second survey, which consisted of the 28 experimental and 32 filler items only, respondents required 12.8 minutes (*SD* = 9.8) in the text feature condition and 12.0 minutes (*SD* = 7.5) in the control condition on average.

## 7.3 Results

I first looked at differences in the response quality indicators across the two experimental conditions (Hypotheses 1a and 2a). Except for breakoffs, these analyses were followed by regression analyses and - if appropriate - simple slopes analyses to examine whether and to what extent effects of question comprehensibility were moderated by verbal intelligence and/or motivation (Hypotheses 1b and 2b). The descriptive statistics of the response quality indicators and the predictor variables, as well as the intercorrelations between all variables in both datasets, are shown in Table 7.1.

### 7.3.1 Breakoffs

A total of 64 respondents (7.2%) dropped out of the first survey before completing it. As expected, more respondents broke off in the text feature condition ($n$ = 38) than in the control condition ($n$ = 26). However, this difference was not statistically significant ($\chi^2 = 2.4$, df = 1, $p > .05$).

### 7.3.2 Non-substantive responses

The tendency to provide non-substantive responses was estimated by calculating the number of "Don't know's" (DKs) and missing answers across the 28 experimental questions.[12] On average, respondents in the text feature condition gave significantly

[12] I additionally analyzed the responses to the 32 filler questions (which were identical in the two questionnaire versions) and found no significant differences between both conditions with regard to the dependent variables (non-substantive responses, neutral responses, over-time consistency). These findings suggest that both groups were equivalent in their response behavior to unproblematic questions.

Table 7.1. Means, standard deviations, and intercorrelations for response quality indicators and predictor variables

| *Web Survey 1* | *M* | *SD* | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 1. Non-substantive responses | 1.57 | 2.15 | – | | | |
| 2. Neutral responses | 1.11 | 1.12 | .06 | – | | |
| 3. Comprehensibility (TF vs. Control) [+] | .49 | .50 | .08* | .11** | – | |
| 4. Verbal intelligence (WSTmod) | 11.43 | 2.13 | -.26*** | -.04 | .02 | – |
| 5. Motivation (MOT) | 4.78 | .83 | -.23*** | -.05 | .06 | .26*** |
| *Web Survey 2* | | | | | | |
| 1. Over-time consistency | 8.81 | 2.99 | – | | | |
| 2. Comprehensibility (TF vs. Control) | .48 | .50 | .09* | – | | |
| 3. Verbal intelligence (WSTmod) | 11.43 | 2.09 | -.14** | .01 | – | |
| 4. Motivation (MOT) | 4.73 | .82 | -.08 | -.03 | .25*** | |

*Note*. All coefficients are Pearson correlations, [+] 0 = Control questions, 1 = Text feature questions.
*$p < .05$, **$p < .01$, ***$p < .001$.

more non-substantive responses to the experimental questions (6.2% of the answers) than respondents in the control condition (4.9%), $\chi^2 = 16.1$, df = 1, $p < .001$.[13]

In a second step, I fitted two regression models. Since the dependent variable took the form of a count (number of non-substantive responses) and the data included a large number of zero counts (i.e., 327 out of 825 cases did not provide any non-substantive response), zero-inflated Poisson regression models were estimated

[13] All analyses were repeated excluding the four experimental questions that required bridging inferences, because these had not been found to induce comprehension difficulties in the eye-tracking study (Chapter 6). However, all of the conclusions remained unchanged (results available on request).

(cf. Federico & Schneider, 2007).[14] The models included question comprehensibility, verbal intelligence (WSTmod), motivation (MOT) and the two-way and three-way interactions of these variables. The question comprehensibility variable was dummy coded (0 = control condition, 1 = text feature condition) and the continuous predictor variables WSTmod and MOT were centered prior to analysis (cf. Whisman & McClelland, 2005). Robust standard errors were used in the analyses to adjust for heterogeneity in the models.

Table 7.2 summarizes the results of the regression models. In Model 1, only question comprehensibility, verbal intelligence and motivation were included to examine the main effects of these variables on non-substantive responses. Statistically significant effects were found for all three variables (comprehensibility: $b$ = .22, $p$ < .05, verbal intelligence: $b$ = -.08, $p$ < .01, motivation: $b$ = -.22, $p$ < .01), indicating that lower levels of comprehensibility, verbal intelligence, and motivation increased the number of non-substantive answers. Model 2 also included the two-way and three-way interactions of the three individual variables to examine whether the impact of question comprehensibility on providing non-substantive answers was moderated by respondents' verbal intelligence and/or motivation. In this model, the two-way interaction between comprehensibility and verbal intelligence was significant ($b$ = -.12, $p$ < .05), indicating that the effect of question comprehensibility on non-substantive responses depended upon the particular level of respondent's verbal intelligence. It is important to note that the coefficients of the individual predictors in moderator regression models do not estimate main effects (as in Model 1) but conditional effects that hold only when all other individual variables have a value of 0 (which represents the mean of the continuous variables that have been centered and the control condition of the categorical variable). Similarly, the two-way interactions are interpreted at a value of 0 (i.e., the mean) for the third variable.

[14] To confirm that this decision was appropriate, I conducted Vuong tests (Long, 1997) for all zero-inflated regression models that were performed in the analyses. These tests indicated that the zero-inflated models were more appropriate than the ordinary Poisson regression models (non-substantive responses, Model 1: z = 5.72, p < .0001, Model 2: z = 5.46, p < .0001; neutral responses, Model 1: z = 1.80, p < .05, Model 2: z = 2.69, p < .01). In each of the reported regressions, the inflation model contained the same set of predictor variables as the count model.

Table 7.2. Regression analyses summary for variables predicting non-substantive responses

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| Variable | *b* | SE *b* | *B* | SE *b* |
| Comprehensibility (TF vs. Control) | .22* | (.09) | .24* | (.10) |
| Verbal intelligence (WSTmod) | -.08** | (.03) | -.02 | (.04) |
| Motivation (MOT) | -.22** | (.08) | -.35*** | (.09) |
| Comprehensibility × verbal intelligence | | | -.12* | (.05) |
| Comprehensibility × motivation | | | .13 | (.15) |
| Verbal intelligence × motivation | | | .01 | (.04) |
| Comprehensibility × verbal intelligence × motivation | | | -.09 | (.06) |
| Constant | .63*** | (.07) | .60*** | (.07) |
| Log likelihood | -1449.25 | | -1427.53 | |
| Wald $\chi^2$ (degrees of freedom) | 36.37 | | 51.65 | |
| | (3)*** | | (7)*** | |
| N | 825 | | 825 | |

*Note*. Entries are zero-inflated Poisson regression coefficients and robust SEs. The functional form for the inflation models was the logistic; estimates for these models are not shown. The question comprehensibility variable was dummy coded (0 = control condition, 1 = text feature condition). Source: Web Survey 1.
*$p < .05$, **$p < .01$, ***$p < .001$.

Hence, these coefficients should not be interpreted as "main effects" (Whisman & McClelland, 2005).

To examine the interaction between question comprehensibility and verbal intelligence in more detail, I conducted simple slopes analyses (Aiken & West, 1991). These can be employed to determine whether the question comprehensibility effects are larger for respondents low in verbal intelligence (i.e. one standard deviation below the mean) than for respondents high in verbal intelligence (i.e., one standard deviation above the mean). The analyses revealed a significant relationship between question comprehensibility and the propensity to provide non-substantive

responses for respondents at low levels of verbal intelligence ($b$ = .49, $p$ < .001), but not for respondents at high levels of verbal intelligence ($b$ = .01, $p$ > .05). Hence, the effect of question comprehensibility on providing non-substantive responses was more pronounced among respondents with limited verbal skills.

In contrast to my expectations, comprehensibility did not interact with respondent motivation, suggesting that less comprehensible questions increased the number of non-substantive responses for highly and lowly motivated respondents alike. Also in contrast to my expectations, I found no significant three-way interaction, and hence neither of the two-way interactions was moderated by a third variable.

### 7.3.3 Neutral responses

The propensity to give neutral responses was estimated by calculating the number of "neither/nor" responses given to those eight experimental questions that offered a middle category. As hypothesized, respondents answering text feature questions provided more neutral responses (15.5% of the answers) than respondents answering control questions (12.4%), $\chi^2$ = 13.5, df = 1, $p$ < .001. Again, I fitted zero-inflated Poisson regression models to examine this effect in more detail (see Table 7.3). The regression models included the same set of variables as the regression models reported above. Again, Model 1 looked at the main effects of the three key independent variables and revealed a marginally significant effect of question comprehensibility ($b$ = .17, $p$ < .10) and a significant effect of motivation ($b$ = -.15, $p$ < .01) on the number of neutral responses. Model 2, which included the two-way and three-way interactions of the variables, showed a significant interaction between comprehensibility and motivation ($b$ = -.19, $p$ < .05), qualifying the main effects and suggesting that the effect of comprehensibility on neutral responses depended upon respondents' level of motivation. Simple slopes analyses revealed a significant simple slope for respondents with low levels of motivation ($b$ = .27, $p$ < .05), but not for highly motivated respondents ($b$ = -.04, $p$ > .05). Hence, low question comprehensibility only increased the number of neutral responses for respondents low in motivation.

Table 7.3. Regression analyses summary for variables predicting neutral responses

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| Variable | *b* | SE *b* | *b* | SE *b* |
| Comprehensibility (TF vs. Control) | .17$^{+}$ | (.10) | .11 | (.08) |
| Verbal intelligence (WSTmod) | -.02 | (.02) | .00 | (.03) |
| Motivation (MOT) | -.15** | (.05) | -.01 | (.07) |
| Comprehensibility × verbal intelligence | | | -.02 | (.04) |
| Comprehensibility × motivation | | | .19* | (.09) |
| Verbal intelligence × motivation | | | .04 | (.04) |
| Comprehensibility × verbal intelligence × motivation | | | -.09 | (.05) |
| Constant | .08 | (.09) | .14* | (.07) |
| Log likelihood | -1139.84 | | -1126.73 | |
| Wald $\chi^2$ (degrees of freedom) | 15.85 | | 18.30 | |
| | (3)** | | (7)* | |
| N | 825 | | 825 | |

*Note*. Entries are zero-inflated Poisson regression coefficients and robust SEs. The functional form for the inflation models was the logistic; estimates for these models are not shown. The question comprehensibility variable was dummy coded (0 = control condition, 1 = text feature condition). Source: Web Survey 1.
$^{+}p < .10$, $*p < .05$, $**p < .01$, $***p < .001$.

Surprisingly, the models revealed no effects of verbal intelligence, suggesting that this variable did not affect the likelihood of selecting neutral responses. Moreover, the three-way interaction was again not significant, and hence the significant two-way interaction of comprehensibility with motivation was not moderated by respondents' verbal intelligence.

### 7.3.4 Over-time consistency

To examine the consistency of respondents' answers to the same questions across the two Web surveys, I calculated the gross error rate (i.e., the simple response variance) for 26 of the 28 experimental questions (cf. Poe et al., 1988). Two questions were excluded from these analyses because they asked about behaviors during a specific time period (e.g., "during the last four weeks"), and thus were not comparable across the two surveys. To calculate the gross error rate, I computed a new variable for each question, coded 1 for respondents who gave the same answers and 0 for those who gave different answers in the two surveys (cf. Krosnick et al., 2002). The average gross error rate across all 26 questions was significantly higher in the text feature condition (35.0%) than in the control condition (32.9%), indicating that the text feature questions reduced the reliability of responses ($\chi^2 = 6.8$, df = 1, $p < .01$). Given that the dependent variable (i.e., number of inconsistent responses) did not contain any zero counts, I fitted Poisson regression models to look for any interaction effects (Table 7.4). Again, Model 1 only looked for main effects and revealed significant effects of question comprehensibility ($b = .06$, $p < .05$) and verbal intelligence ($b = -.02$, $p < .01$) on over-time consistency. The reliability of responses was not affected by respondents' level of motivation ($b = -.02$, $p > .05$). Moreover, Model 2 revealed no significant two- or three-way interaction predicting the consistency of responses, and hence the relation between question comprehensibility and over-time consistency was neither moderated by verbal intelligence nor by motivation.

## 7.4 Discussion

This study has found clear evidence that reduced survey question comprehensibility (operationalized by seven text features that undermine comprehension) reduces response quality: respondents receiving less comprehensible questions were more likely to drop out of the survey and they provided significantly more non-substantive (DKs and missings), more neutral (i.e., midpoint), and fewer reliable responses than respondents answering comprehensible questions. Moreover, some of these effects were conditional upon respondents' verbal skills (non-substantive responses), while

Table 7.4. Regression analyses summary for variables predicting over-time consistency of responses

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| Variable | *b* | SE *b* | *b* | SE *b* |
| Comprehensibility (TF vs. Control) | .06* | (.03) | .06$^{+}$ | (.03) |
| Verbal intelligence (WSTmod) | -.02** | (.01) | -.01 | (.01) |
| Motivation (MOT) | -.02 | (.02) | -.04 | (.03) |
| Comprehensibility × verbal intelligence | | | -.01 | (.01) |
| Comprehensibility × motivation | | | .04 | (.04) |
| Verbal intelligence × motivation | | | .01 | (.01) |
| Comprehensibility × verbal intelligence × motivation | | | .00 | (.02) |
| Constant | 2.14*** | (.02) | 2.14*** | (.02) |
| Log likelihood | -1279.73 | | -1278.02 | |
| $\chi^2$ (degrees of freedom) | 15.34 | | 18.78 | |
| | (3)** | | (7)** | |
| N | 515 | | 515 | |

*Note*. Entries are Poisson regression coefficients and robust SEs. The question comprehensibility variable was dummy coded (0 = control condition, 1 = text feature condition). Sources: Web Survey 1 and 2.
$^{+}p < .10$, $*p < .05$, $**p < .01$, $***p < .001$.

others were conditional upon respondents' motivation (neutral responses). Taken together, these findings indicate that survey data quality is reduced if questions are difficult to understand and exceed the processing effort that respondents are willing or able to invest.

With regard to satisficing theory, the study did not find any three-way interactions of question comprehensibility (i.e., task difficulty), verbal intelligence (i.e., cognitive ability), and motivation in the way that, for example, the question comprehensibility effects were strongest among respondents both low in verbal intelligence and motivation. Thus, survey satisficing may not generally be the results

of a three-way interaction of these variables. Instead, the significant two-way interactions suggest that respondents employ specific response strategies depending on their level of verbal intelligence on the one hand and on their level of motivation on the other hand. When confronted with less comprehensible questions, respondents with limited verbal skills (irrespective of their level of motivation) tended to provide non-substantive responses, whereas those with low motivation (irrespective of their verbal abilities) tended to provide neutral responses. It is conceivable that respondents with limited verbal skills prematurely decide that they do not have the necessary information to answer the questions if they already have problems to understand what these are about. Hence, these respondents may not even try to interpret the questions correctly but may satisfice instead by selecting a non-substantive response. On the other hand, respondents with low motivation to answer the questions may prematurely decide that they do not have or do not want to generate an opinion about the issue in question if understanding the question is burdensome. Hence, these respondents may satisfice by selecting a neutral response even though they may have been able to report an opinion. These issues call for future experimental studies that explore the underlying mechanisms that evoke these specific response strategies in more detail.

There are two limitations to this study. First, respondents were drawn from a nonprobability online panel which may restrict the generalizability of the results. At the same time, the low response rate (10.9%) together with the low breakoff rate (7.2%) suggest that only a small proportion of highly motivated respondents participated in the survey. These respondents may have been less influenced by the incomprehensible questions than less motivated respondents who may have exhibited even more satisficing behavior. Second, better educated respondents were overrepresented in the sample. However, assuming that more educated respondents are better and more competent readers, the question comprehensibility effects could have been even stronger if the sample had included a larger number of less educated respondents.

## 7.5 APPENDIX A: QUESTION WORDINGS

**Overview of the experimental questions**

Low-frequency words (LFRW): Q1, Q16, Q22, Q42

Vague or imprecise relative terms (VIRT): Q3, Q4, Q12, Q44

Vague or ambiguous noun phrases (VANP): Q19, Q24, Q38, Q56

Complex syntax (CSYN): Q13, Q39, Q47, Q51

Complex logical structures (CLOG): Q9, Q21, Q32, Q53

Low syntactic redundancy (LSYR): Q20, Q26, Q45, Q57

Bridging inferences (BINF): Q33, Q34, Q41, Q59

Grey = Experimental question

**Block 1: Environment**

(Q1) Text feature version:

LFRW1 Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **AKW** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Control:
Für wie wahrscheinlich halten Sie es, dass in den nächsten fünf Jahren ein Unfall in einem **Atomkraftwerk** zu langfristigen Umweltschäden in vielen Ländern führen wird?

Answer options:
Sehr wahrscheinlich; Wahrscheinlich; Unwahrscheinlich; Sehr unwahrscheinlich; Kann ich nicht sagen

(Q2) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Wir machen uns zu viele Sorgen über die Zukunft der Umwelt und zu wenig um Preise und Arbeitsplätze heutzutage.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q3) Text feature version:

VIRT1 Ich schränke **selten** der Umwelt zuliebe das Autofahren ein.
Stimmt; Stimmt eher; Stimmt eher nicht; Stimmt nicht; Kann ich nicht sagen

Control version:
Wie häufig schränken Sie der Umwelt zuliebe das Autofahren ein?
Immer; Manchmal; **Selten**; Nie; Kann ich nicht sagen

(Q 4) Text feature version:

VIRT2 Ich verzichte **selten** beim Essen auf Fleisch.
Stimmt; Stimmt eher; Stimmt eher nicht; Stimmt nicht; Kann ich nicht sagen

Control:
Wie häufig verzichten Sie beim Essen auf Fleisch?
Immer; Manchmal; **Selten**; Nie; Kann ich nicht sagen

(Q5) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die moderne Wissenschaft wird unsere Umweltprobleme bei nur geringer Veränderung unserer Lebensweise lösen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q6) Inwieweit fänden Sie es für sich persönlich akzeptabel, viel höhere Steuern zu bezahlen, um die Umwelt zu schützen?

Answer options:
Sehr akzeptabel; Eher akzeptabel; Weder akzeptabel noch inakzeptabel; Eher inakzeptabel; Sehr inakzeptabel; Kann ich nicht sagen

(Q7) Und inwieweit fänden Sie es für sich persönlich akzeptabel, Abstriche von Ihrem Lebensstandard zu machen, um die Umwelt zu schützen?

Answer options:
Sehr akzeptabel; Eher akzeptabel; Weder akzeptabel noch inakzeptabel; Eher inakzeptabel; Sehr inakzeptabel; Kann ich nicht sagen

(Q8) Wenn Sie zwischen den folgenden Aussagen entscheiden müssten, welche von beiden käme Ihrer Meinung am nächsten?

Answer options:
Die Regierung sollte es jedem selbst überlassen, wie er/sie die Umwelt schützt, auch wenn das dazu führt, dass nicht immer das Richtige für die Umwelt getan wird;
Die Regierung sollte Gesetze erlassen, um Leute zu zwingen, die Umwelt zu schützen, auch wenn dies in die Entscheidungsfreiheit des einzelnen eingreift; Kann ich nicht sagen

(Q9) Text feature version:
CLOG1 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Von ärmeren Ländern sollten **nicht** weniger Anstrengungen für den Umweltschutz erwartet werden als von reichen Ländern.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Von ärmeren Ländern sollten weniger Anstrengungen für den Umweltschutz erwartet werden als von reichen Ländern.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q10) Manche Länder tun mehr für den globalen Umweltschutz als andere. Ganz allgemein gesehen, tut Deutschland Ihrer Meinung nach…

Answer options:
mehr als genug?; in etwa genug?; zu wenig?; Kann ich nicht sagen

**Block 2: Health**

(Q11) Ist es gerecht oder ungerecht, dass sich Menschen mit höherem Einkommen eine bessere Gesundheitsversorgung leisten können als Menschen mit geringerem Einkommen?

Answer options:
Sehr gerecht, Eher gerecht, Weder gerecht noch ungerecht, Eher ungerecht, Sehr ungerecht, Kann ich nicht sagen

(Q12) Text feature version:
VIRT3 Haben Sie **kürzlich** einen oder mehrere Ärzte aufgesucht? Wenn ja, geben Sie bitte die Anzahl der Arztbesuche an.

Control:
Haben Sie in den letzten **vier Wochen** Ärzte aufgesucht? Wenn ja, geben Sie bitte die Anzahl der Arztbesuche an.

Answer options:
Kein Arztbesuch, Ein Arztbesuch, Zwei Arztbesuche, Drei oder mehr Arztbesuche

(Q13) Text feature version:

CSYN1 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ärzte verwenden häufig **für ihre Patienten schwer verständliche** Ausdrücke oder Formulierungen.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ärzte verwenden häufig Ausdrücke oder Formulierungen, die **für ihre Patienten schwer zu verstehen** sind.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q14) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ärzte können ihre Patienten immer seltener angemessen behandeln, weil es zu viele Bestimmungen und Vorschriften gibt.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q15) Und inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ärzte interessieren sich mehr dafür, Kosten zu begrenzen als dafür, was ihre Patienten brauchen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q16) Text feature version:

LFRW2 Wie häufig kam es in den letzten vier Wochen vor, dass Sie **somatische** Beschwerden hatten?

Control version:
Wie häufig kam es in den letzten vier Wochen vor, dass Sie **körperliche** Beschwerden hatten?

Answer options:
Sehr oft; Oft; Manchmal; Fast nie; Nie; Kann ich nicht sagen

(Q17) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen sollten Zugang zu allen nötigen Gesundheitsleistungen haben, auch wenn sie dafür nicht selbst bezahlen können.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q18) Inwieweit wären Sie bereit, höhere Steuern zu zahlen, um die Gesundheitsversorgung für alle Menschen in Deutschland zu verbessern?

Answer options:
Auf jeden Fall bereit, Eher bereit, Weder bereit noch nicht bereit, Eher nicht bereit, Auf keinen Fall bereit, Kann ich nicht sagen

(Q19) Text feature version:
VANP1 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ich esse nicht genug **pflanzliche Lebensmittel**.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ich esse nicht genug **Obst und Gemüse**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q20) Text feature version:
LSYR1 In welchem Maß **werden** Sie durch Ihre Gesundheit daran **gehindert**, Ihre Freizeit so zu gestalten, wie Sie dies gerne tun würden?

Control version:
In welchem Maß **hindert** Sie Ihre Gesundheit daran, Ihre Freizeit in der Weise zu gestalten, wie Sie dies gerne tun würden?

Answer options:
In sehr hohem Maß; In hohem Maß; Bis zu einem gewissen Maß; Überhaupt nicht; Kann ich nicht sagen

**Block 3: Leisure**

(Q21) Text feature version:
CLOG2 Wie viele Erwachsene **und** Kinder leben **außer** Ihnen selbst in Ihrer Wohnung **oder** Ihrem Haushalt?

Control version:
Wie ist das bei Ihnen, wie viele Personen leben insgesamt in Ihrer Wohnung **oder** Ihrem Haushalt?

Answer options:
(keine), 1,2,3,4,5 oder mehr

(Q22) Text feature version:
LFRW3 Man kann sich in seiner Freizeit auf unterschiedliche Weise beschäftigen. Bitte geben Sie an, wie häufig Sie Ihre Freizeit damit verbringen, **ersprießliche** Kontakte zu knüpfen.

Control version:
Man kann sich in seiner Freizeit auf unterschiedliche Weise beschäftigen. Bitte geben Sie an, wie häufig Sie Ihre Freizeit damit verbringen, **nützliche** Kontakte zu knüpfen.

Answer options:
Sehr oft; Oft; Manchmal; Selten; Nie; Kann ich nicht sagen

(Q23) Sind Sie in Ihrer Freizeit lieber mit anderen zusammen oder lieber allein?

Answer options:
Ich bin lieber…
meistens mit anderen zusammen; mehr mit anderen zusammen als allein; mehr allein als mit anderen zusammen; meistens allein; Kann ich nicht sagen

(Q24) Text feature version:
VANP2 Wie häufig besuchen Sie in Ihrer Freizeit **kulturelle Veranstaltungen**?

Control version:
Wie häufig besuchen Sie in Ihrer Freizeit **Theateraufführungen**?

Answer options:
Mehrmals in der Woche; Mehrmals im Monat; Mehrmals im Jahr oder seltener; Nie; Kann ich nicht sagen

(Q25) Wie häufig kommt es in Ihrer Freizeit vor, dass Sie an Ihre berufliche Arbeit denken?

Answer options:
Sehr oft, Oft, Manchmal, Selten, Nie; Trifft nicht zu; Kann ich nicht sagen

(Q26) Text feature version:
LSYR2 Und wie häufig verbringen Sie Ihre Freizeit mit dem **Erwerb** neuer Kenntnisse?

Control version:
Und wie häufig verbringen Sie Ihre Freizeit damit, neue Kenntnisse zu **erwerben**?

Answer options:
Sehr oft; Oft; Manchmal; Selten; Nie; Kann ich nicht sagen

(Q27) Es gibt unterschiedliche Meinungen zum Thema Sport. Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Sport zu treiben fördert die Charakterentwicklung von Kindern.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q28) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Sport bringt unterschiedliche Gruppen in Deutschland einander näher, etwa Gruppen verschiedener nationaler oder ethnischer Herkunft.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q29) Und inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Internationale Sportwettkämpfe erzeugen mehr Spannungen zwischen Ländern als positive Gefühle.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q30) Wie stolz sind Sie, wenn Deutschland bei internationalen Sportwettkämpfen gut abschneidet?

Answer options
Ich bin…
sehr stolz, etwas stolz, nicht sehr stolz, überhaupt nicht stolz, Kann ich nicht sagen

**Block 4: Role of government**

(Q31) Ganz allgemein gesprochen, würden Sie sagen, dass man Gesetze ohne Ausnahme befolgen muss, oder gibt es Ausnahmesituationen, in denen man seinem Gewissen folgen sollte, auch wenn dies bedeutet, Gesetze zu übertreten?

Answer options:
Gesetze ohne Ausnahme befolgen; oder In Ausnahmesituationen seinem Gewissen folgen, Kann ich nicht sagen

(Q32) Text feature version:
CLOG3 Es gibt viele Möglichkeiten, mit denen einzelne **oder** Gruppen gegen Regierungsmaßnahmen **oder** Regierungsvorhaben protestieren können, wenn sie diese entschieden **oder** zumindest ein wenig ablehnen. Sollte in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein?
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren.

Control version:
Es gibt verschiedene Möglichkeiten, mit denen einzelne **oder** Vereinigungen gegen eine Regierungsmaßnahme protestieren können, wenn sie diese Maßnahme entschieden ablehnen. Geben Sie bitte an, inwieweit in diesem Zusammenhang Ihrer Meinung nach die unten aufgeführte Protestaktion erlaubt sein sollte:
Öffentliche Versammlungen organisieren, um gegen die Regierung zu protestieren.

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein; Kann ich nicht sagen

(Q33) Text feature version:
BINF1 Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden. **Denken Sie einmal** an Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte:
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Control version:
Es gibt einige Menschen, deren Ansichten von den meisten anderen als extrem angesehen werden, **wie zum Beispiel** Menschen, die die Regierung durch eine Revolution stürzen wollen. Geben Sie bitte an, inwieweit diesen Menschen die folgende Tätigkeit erlaubt sein sollte:
Öffentliche Versammlungen abhalten, auf denen sie ihre Ansichten äußern.

Answer options:
Sollte auf jeden Fall erlaubt sein; Sollte schon erlaubt sein; Sollte eigentlich nicht erlaubt sein; Sollte auf keinen Fall erlaubt sein; Kann ich nicht sagen

(Q34) Text feature version:
BINF2 Auch Gerichte **können sich irren**. Was halten Sie dann für schlimmer...

Control version:
Auch Gerichte **fällen falsche Urteile**. Was halten Sie dann für schlimmer...

Answer options:
eine unschuldige Person zu verurteilen?; eine schuldige Person freizusprechen?; Kann ich nicht sagen

(Q35) Angenommen, staatliche Stellen haben den Verdacht, dass ein Terroranschlag droht. Was meinen Sie, sollten diese das Recht haben, Menschen einfach so auf der Straße anzuhalten und zu durchsuchen?

Answer options:
Auf jeden Fall, Eher ja, Eher nein, Auf keinen Fall; Kann ich nicht sagen

(Q36) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen wie ich haben keinen Einfluss darauf, was die Regierung macht.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q37) Und inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die Politiker, die wir in den Bundestag wählen, versuchen, ihre Versprechen aus dem Wahlkampf zu halten.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q38) Text feature version:
VANP3 Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die **Sie** bitten können, wichtige Entscheidungen zu **ihren** Gunsten zu beeinflussen?

Control version:
Manche Menschen haben aufgrund ihrer beruflichen oder gesellschaftlichen Stellung oder wegen ihrer Beziehungen Einfluss auf wichtige öffentliche Entscheidungen. Deshalb werden sie von anderen Menschen gebeten, zu deren Gunsten Einfluss zu nehmen. Wie ist das bei Ihnen? Gibt es Menschen, die sich an **Sie** wenden können, damit Sie wichtige Entscheidungen zu **deren** Gunsten beeinflussen?

Answer options:
Ja, viele; Ja, einige; Ja, aber nur wenige; Nein, niemand; Kann ich nicht sagen

(Q39) Text feature version:
CSYN2 Wie häufig sind Sie oder ein Mitglied Ihrer Familie in den letzten fünf Jahren auf öffentliche Bedienstete, die als Gegenleistung für eine Dienstleistung andeuteten, eine Bestechung oder einen Gefallen zu wollen oder dies sogar forderten, **gestoßen**?

Control version:
Wie häufig sind Sie oder ein Mitglied Ihrer Familie in den letzten fünf Jahren auf öffentliche Bedienstete **gestoßen**, die als Gegenleistung für eine Dienstleistung andeuteten, eine Bestechung oder einen Gefallen zu wollen oder dies sogar forderten?

Answer options:
Sehr oft; Oft; Manchmal; Selten; Nie; Kann ich nicht sagen

(Q40) Was meinen Sie, wie häufig behandeln Beamte Menschen wie Sie fair?

Answer options:
Fast immer, Oft, Manchmal, Selten, Fast nie, Kann ich nicht sagen

**Block 5: National identity**

(Q41) Text feature version:
BINF3 Manche Menschen sind bereit, nach ihrem Tod ihren Körper der **Medizin zur Verfügung** zu stellen. Wären Sie bereit, nach Ihrem Tod ein **Organ zu spenden**?

Control version:
Manche Menschen sind bereit, nach ihrem Tod **Organe zu spenden**. Wie ist das bei Ihnen? Wären Sie dazu bereit, nach Ihrem Tod ein **Organ zu spenden**?

Answer options:
Ja, ganz bestimmt; Ja, wahrscheinlich; Nein, wahrscheinlich nicht; Nein, bestimmt nicht; Kann ich nicht sagen

(Q42) Text feature version:
LFRW4 Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **übervorteilen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Control version:
Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu **betrügen** oder aber versuchen, sich Ihnen gegenüber fair zu verhalten?
Andere Leute würden...

Answer options:
fast immer versuchen, mich zu übervorteilen/betrügen; meistens versuchen, mich zu übervorteilen/betrügen; meistens versuchen, sich mir gegenüber fair zu verhalten; fast immer versuchen, sich mir gegenüber fair zu verhalten; Kann ich nicht sagen

(Q43) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Es ist die Aufgabe des Staates, die Einkommensunterschiede zwischen Arm und Reich abzubauen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q44) Text feature version:
VIRT4 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Studenten aus einkommensschwachen Familien sollten **beträchtliche** finanzielle Unterstützung vom Staat erhalten.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Die Studenten aus einkommensschwachen Familien sollten eine finanzielle Unterstützung vom Staat erhalten.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q45) Text feature version:
LSYR3 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Von Menschen in reichen Ländern sollte eine zusätzliche Steuer **entrichtet werden**, um Menschen in armen Ländern zu helfen.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen in reichen Ländern sollten eine zusätzliche Steuer **entrichten**, um den Menschen in armen Ländern zu helfen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q46) Manche Leute meinen, dass es für ein Land besser ist, wenn Gruppen verschiedener nationaler Herkunft oder Hautfarbe ihre eigenen Sitten und Gebräuche beibehalten. Andere finden es besser, wenn solche Gruppen sich anpassen und in der Gesamtgesellschaft aufgehen. Welche Meinung kommt Ihrer eigenen Ansicht näher?

Answer options:
Es ist besser für die Gesellschaft, wenn solche Gruppen ihre unterschiedlichen Sitten und Gebräuche beibehalten; Es ist besser, wenn solche Gruppen sich anpassen und in der Gesamtgesellschaft völlig aufgehen; Kann ich nicht sagen

(Q47) Text feature version:
CSYN3 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Menschen, die in Deutschland geboren sind, **werden** von Zuwanderern Arbeitsplätze **weggenommen**.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zuwanderer **nehmen** den Menschen, die in Deutschland geboren sind, die Arbeitsplätze **weg**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q48) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zuwanderer sind im Allgemeinen gut für die deutsche Wirtschaft.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q49) Und inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Zuwanderer bereichern Deutschland durch neue Ideen und Kulturen.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q50) Meinen Sie, dass die Zahl der Zuwanderer nach Deutschland heutzutage…

Answer option:
Deutlich erhöht werden sollte, leicht erhöht werden sollte, so bleiben sollte, wie sie ist; leicht verringert werden sollte, deutlich verringert werden sollte?; Kann ich nicht sagen

**Block 6: Politics and social inequality**

(Q51) Text feature version:
CSYN4 Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, etwas gegen ein Gesetz, das der Bundestag berät und das Sie für ungerecht oder schädlich halten, zu **unternehmen versuchen**?

Control version:
Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen
zusammen, **versuchen** etwas gegen ein Gesetz zu **unternehmen**, das der Bundestag berät und das Sie für ungerecht oder schädlich halten?

Answer options:
Sehr wahrscheinlich; Einigermaßen wahrscheinlich; Nicht sehr wahrscheinlich;
Überhaupt nicht wahrscheinlich; Kann ich nicht sagen

(Q52) Welche dieser zwei Aussagen kommt ihrer Ansicht am nächsten?

Answer options:
In internationalen Organisationen sollten Entscheidungen den Vertretern der nationalen Regierungen überlassen werden; In internationalen Organisationen sollten Bürgervereinigungen direkt am Entscheidungsprozess beteiligt sein; Kann ich nicht sagen

(Q53) Text feature version:
CLOG4 **Angenommen** Sie wären Bundeskanzler/in und stünden vor dem Problem, dass deutsche Interessen in einer Streitfrage mit denen anderer Länder nicht vereinbar sind. Würden Sie sich dafür einsetzen, dass die deutschen Interessen verfolgt werden, auch wenn dies zu Konflikten mit anderen Ländern führt?
Ja, auf jeden Fall; Ja; Nein; Nein, auf keinen Fall; Kann ich nicht sagen

Control version:
Wie sehr stimmen Sie der folgenden **Aussage** zu oder nicht zu: Deutschland sollte seine eigenen Interessen verfolgen, selbst wenn dies zu Konflikten mit anderen Ländern führt. Bitte beantworten Sie diese Frage auf der folgenden Skala von „Stimme voll und ganz zu“ bis „Stimme überhaupt nicht zu“.
Stimme voll und ganz zu; Stimme zu; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q54) Was meinen Sie, wie erfolgreich ist zurzeit der Staat, wenn es darum geht mit Bedrohungen der inneren und äußeren Sicherheit Deutschlands umzugehen?

Answer options:
Sehr erfolgreich, Ziemlich erfolgreich, Weder noch, Ziemlich erfolglos, Äußerst erfolglos, Kann ich nicht sagen

(Q55) Und wie erfolgreich ist zurzeit der Staat, wenn es darum geht die Arbeitslosigkeit zu bekämpfen?

Answer options:
Sehr erfolgreich, Ziemlich erfolgreich, Weder noch, Ziemlich erfolglos, Äußerst erfolglos, Kann ich nicht sagen

(Q56) Text feature version:
VANP4 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Während der **Hochzeit** der Investmentbranche zwischen 2002 und 2007 hätte die Politik stärker in die Wirtschaft eingreifen müssen, um eine Finanzkrise zu verhindern.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Während der **Glanzzeit** der Investmentbranche zwischen 2002 und 2007 hätte die Politik stärker in die Wirtschaft eingreifen müssen, um eine Finanzkrise zu verhindern.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q57) Text feature version:
LSYR4 Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind für die **Sicherung** der Arbeitsplätze von Arbeitnehmern wichtig.

Control version:
Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Gewerkschaften sind wichtig um die Arbeitsplätze von Arbeitnehmern zu **sichern**.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q58) Und inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Ohne Gewerkschaften wären die Arbeitsbedingungen für Arbeitnehmer viel schlechter.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

(Q59) Text feature version.
BINF4 Zurzeit wird in Deutschland viel über die **alternde Gesellschaft und deren Folgen** diskutiert. Unten finden Sie drei mögliche Maßnahmen, um die Probleme der gesetzlichen Rentenversicherung zu lösen. Wenn Sie sich für eine davon entscheiden müssten, welche würden Sie wählen?
Um die Probleme der gesetzlichen Rentenversicherung zu lösen, …

Control version:
Zurzeit wird in Deutschland viel über **Rente, Rentenfinanzierung und Rentenalter** diskutiert. Unten finden Sie drei mögliche Maßnahmen, um die Probleme der gesetzlichen Rentenversicherung zu lösen. Wenn Sie sich für eine davon entscheiden müssten, welche würden Sie wählen?
Um die Probleme der gesetzlichen Rentenversicherung zu lösen, …

Answer options:
Sollte das Rentenalter erhöht werden, sollten die Rentenbeiträge erhöht werden; sollten die gesetzlichen Renten gekürzt werden; Kann ich nicht sagen

(Q60) Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?
Erwachsene Kinder haben die Pflicht, sich um ihre betagten Eltern zu kümmern.

Answer options:
Stimme voll und ganz zu; Stimme zu; Weder noch; Stimme nicht zu; Stimme überhaupt nicht zu; Kann ich nicht sagen

# 7.6 APPENDIX B: VOCABULARY TESTS

## 1. WOCT

Bevor wir mit der Umfrage beginnen, möchten wir Sie bitten, zwei Listen mit jeweils 15 Wörtern durchzulesen. Es handelt sich hierbei um Wörter, die in Befragungen vorkommen können. Um unsere Umfragen zu verbessern, möchten wir herausfinden, wie bekannt diese Wörter sind.

Bitte geben Sie bei jedem Wort an, ob Sie es kennen, das heißt jemand anderem seine Bedeutung erklären könnten oder nicht.

Ich kenne die Bedeutung des Wortes…

| | Ja | Nein |
|---|---|---|
| 1. Mahnung | ☐ | ☐ |
| 2. Eichmaß | ☐ | ☐ |
| 3. Holozän | ☐ | ☐ |
| 4. Offerte | ☐ | ☐ |
| 5. Habitat | ☐ | ☐ |
| 6. Altenuse (fake word) | ☐ | ☐ |
| 7. Erosion | ☐ | ☐ |
| 8. Platine | ☐ | ☐ |
| 9. Halali | ☐ | ☐ |
| 10. Fazit | ☐ | ☐ |
| 11. Kürschner | ☐ | ☐ |
| 12. Triasmus (fake word) | ☐ | ☐ |
| 13. Sparta | ☐ | ☐ |
| 14. Hybris | ☐ | ☐ |
| 15. Enklivie (fake word) | ☐ | ☐ |

**2. WSTmod**

Und hier die zweite Liste mit 15 Wörtern.

Bitte geben Sie bei jedem Wort an, ob Sie es kennen, das heißt jemand anderem seine Bedeutung erklären könnten oder nicht.

Ich kenne die Bedeutung des Wortes…

| | Ja | Nein |
|---|---|---|
| 1. Ironie | ☐ | ☐ |
| 2. Koalition | ☐ | ☐ |
| 3. salopp | ☐ | ☐ |
| 4. Kaskade | ☐ | ☐ |
| 5. Detail | ☐ | ☐ |
| 6. Fiasko | ☐ | ☐ |
| 7. Eruption | ☐ | ☐ |
| 8. Diskrepanz | ☐ | ☐ |
| 9. Votum | ☐ | ☐ |
| 10. Kausalität | ☐ | ☐ |
| 11. kontaminieren | ☐ | ☐ |
| 12. Sukzession | ☐ | ☐ |
| 13. evident | ☐ | ☐ |
| 14. Flageolett | ☐ | ☐ |
| 15. Kassiterit | ☐ | ☐ |

## 7.7 APPENDIX C: SOCIAL DESIRABILITY ITEMS

Im Folgenden finden Sie mehrere Aussagen, mit denen Sie sich selbst beschreiben können. Diese Aussagen können mehr oder weniger auf Sie zutreffen. Bitte geben Sie bei jeder dieser Aussagen an, wie sehr die Aussage auf Sie persönlich zutrifft.

| | Trifft völlig zu | Trifft ziemlich zu | Trifft etwas zu | Trifft wenig zu | Trifft nicht zu |
|---|---|---|---|---|---|
| In einem Gespräch höre ich meinem Gegenüber immer aufmerksam zu. | □ | □ | □ | □ | □ |
| Ich zögere nie, jemandem in einer Notsituation beizustehen. | □ | □ | □ | □ | □ |
| Ich habe schon mal zu viel Wechselgeld zurückbekommen ohne es der Verkäuferin/dem Verkäufer zu sagen. | □ | □ | □ | □ | □ |
| Es ist schon mal vorgekommen, dass ich jemanden ausgenutzt habe. | □ | □ | □ | □ | □ |
| Ich fluche niemals. | □ | □ | □ | □ | □ |
| Wenn ich etwas versprochen habe, halte ich es ohne Wenn und Aber. | □ | □ | □ | □ | □ |
| Es ist schon mal vorgekommen, dass ich schlecht über jemanden geredet habe. | □ | □ | □ | □ | □ |
| Ich habe schon mal geliehene Sachen nicht zurückgegeben. | □ | □ | □ | □ | □ |
| Ich habe noch nie absichtlich etwas gesagt, um die Gefühle anderer zu verletzen. | □ | □ | □ | □ | □ |
| Manchmal helfe ich jemandem nur, wenn ich eine Gegenleistung erwarten kann. | □ | □ | □ | □ | □ |
| Ich bin immer ehrlich zu anderen. | □ | □ | □ | □ | □ |

# 7.8 APPENDIX D: MEASURES OF MOTIVATION

## 1. Need for cognition

Inwieweit treffen die folgenden Aussagen auf Sie zu oder nicht zu?

| | trifft voll und ganz zu | | | | | | trifft überhaupt nicht zu |
|---|---|---|---|---|---|---|---|
| Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss. | □ | □ | □ | □ | □ | □ | □ |
| Ich würde lieber etwas tun, das wenig Denken erfordert, als etwas, das mit Sicherheit meine Denkfähigkeit herausfordert. | □ | □ | □ | □ | □ | □ | □ |
| Ich würde komplizierte Probleme einfachen Problemen vorziehen. | □ | □ | □ | □ | □ | □ | □ |
| Ich finde wenig Befriedigung darin, angestrengt und stundenlang nachzudenken. | □ | □ | □ | □ | □ | □ | □ |
| Ich versuche, Situationen vorauszuahnen und zu vermeiden, in denen die Wahrscheinlichkeit groß ist, dass ich intensiv über etwas nachdenken muss. | □ | □ | □ | □ | □ | □ | □ |

**2. Need to evaluate**

Inwieweit treffen die folgenden Aussagen auf Sie zu oder nicht zu?

| | trifft voll und ganz zu | | | | | | trifft über-haupt nicht zu |
|---|---|---|---|---|---|---|---|
| Ich bilde mir zu allem eine Meinung. | □ | □ | □ | □ | □ | □ | □ |
| Es ist mir wichtig, einen festen Standpunkt zu vertreten. | □ | □ | □ | □ | □ | □ | □ |
| Oft ziehe ich es vor, bei schwierigen Fragen neutral zu bleiben. | □ | □ | □ | □ | □ | □ | □ |
| Wenn mich etwas nicht betrifft, urteile ich gewöhnlich nicht darüber, ob es gut oder schlecht ist. | □ | □ | □ | □ | □ | □ | □ |
| Ich bilde mir nur dann eine feste Meinung, wenn es sein muss. | □ | □ | □ | □ | □ | □ | □ |
| Ich habe lieber einen festen Standpunkt als gar keine Meinung. | □ | □ | □ | □ | □ | □ | □ |

# 8 CONCLUSION

It is universally acknowledged that the wording of a survey question can have a strong influence on the answers that respondents provide. For example, many studies have shown that vague and ambiguous terms are often interpreted idiosyncratically by respondents, and thus can increase measurement error (Bradburn & Miles, 1979; Fowler, 1992; Schaeffer, 1991; Smith, 1987; Sturgis & Smith, 2010). In addition to ambiguity, the cognitive effort required to comprehend survey questions may affect data quality in a similar way. This aspect of survey question design has received comparatively little attention to date and has rarely been examined experimentally (see Krosnick, 1991, for a theoretical discussion of this issue).

At the outset of this thesis, I suggested that applying a psycholinguistic perspective to survey question design may shed some light on the relationship between the cognitive effort required to comprehend survey questions and the quality of respondents' answers. This perspective conceives survey questions as specific linguistic objects that are more or less difficult to process depending on the complexity of their components and structures. Respondents are conceived as efficient processing devices who consistently (though mostly unconsciously) monitor the efforts and rewards associated with the processing of survey questions and who simplify or terminate the response process as soon as the rewards no longer warrant the efforts. From this perspective, an important goal during survey question design is to write questions that are easy for respondents to comprehend. Attempts should be made to simplify the question comprehension process by avoiding specific text features that undermine the comprehensibility of texts.

To examine the virtue of this psycholinguistic perspective for survey question design, two overall research questions were addressed:

1. Which factors determine the comprehensibility of survey questions?

2. How does the cognitive effort required to comprehend survey questions affect response quality and measurement error?

These overall research questions were analyzed in three consecutive studies. The main findings of these studies are summarized in the following section. Afterwards I discuss the implications of these findings for survey question design and close with some suggestions for future research.

## 8.1 Summary of the results

### 8.1.1. Effects of the text features on question comprehensibility

Chapter 3 described seven psycholinguistic text features that have been found to reduce the comprehensibility of texts: low-frequency words (LFRW), vague or imprecise relative terms (VIRT), vague or ambiguous noun phrases (VANP), complex syntax (CSYN), complex logical structures (CLOG), low syntactic redundancy (LSYR), and bridging inferences (BINF). Study 1 (Chapter 5) and study 2 (Chapter 6) examined whether these text features also affect the comprehensibility of survey questions. In both studies, each of the seven text features was operationalized by a set of four questions.

Study 1 revealed that six of the seven text features reduce question comprehensibility as indicated by significantly longer response times. Only vague or ambiguous noun phrases (VANP) were not found to affect the cognitive burden of the questions. For the most part, these findings were supported by study 2, in which eye-tracking parameters were used as measures of cognitive effort. Again, six of the seven text features were identified to reduce question comprehensibility. In this study, vague or ambiguous noun phrases significantly reduced question comprehensibility while bridging inferences (BINF) had no impact on cognitive effort. Both theoretical and methodological considerations may explain these partially disagreeing results.

From a theoretical point of view, it is conceivable that respondents often interpret vague or ambiguous noun phrases (VANP) idiosyncratically, a procedure which does not cause any difficulties. For example, when answering the question *In your free time, how often do you attend cultural events?,* some respondents may automatically think of their visits to the opera, to the theater, and to classical

concerts, while others may automatically think of their visits to the movies and to pop concerts. In both cases, interpreting the vague term *cultural events* and answering this question would be easy for these respondents to do. However, other respondents may perceive the vagueness of the term *cultural events*, they may wonder whether they are supposed to consider some events (e.g., breakdance performances, rock festivals) or not, and hence may find it quite difficult to answer the question. In this case, the text feature VANP would indeed affect the comprehensibility of survey questions. The analyses per item in study 1 and study 2 suggest that the effects of this text feature on question comprehensibility depend to some degree on the specific instance of the feature. Some forms of VANP (e.g., pronouns with multiple referents, biased ambiguous nouns) seem to be more problematic than others (e.g., abstract nouns, ambiguous pronouns) in that they make it more difficult for respondents to ignore the vagueness/ambiguity of these terms and to interpret them idiosyncratically. These different impact levels of the various forms of VANP may also be the reason why the effects of this text feature on question comprehensibility are weaker or more subtle than the effects of, for example, LFRW or CLOG (see effect sizes in study 2).

From a methodological point of view, the experimental design of study 1 may not have been adequate for identifying the seemingly subtle effects of VANP on question comprehensibility. First, the field experiment made it impossible to control for or minimize all potential confounding factors. Second, response times are imperfect indicators of question comprehensibility because they do not allow us to distinguish between the time required to comprehend a question and the time it takes to arrive at an answer (including information retrieval, judgment, and response selection). In contrast, when we examined this text feature under carefully controlled conditions in the laboratory using eye-tracking methodology (study 2), we found that vague or ambiguous noun phrases clearly reduced the comprehensibility of survey questions.

With regard to bridging inferences (BINF), it has already been discussed in Chapter 6 that their effects on comprehensibility may depend on whether drawing these inferences is essential for answering the questions or not. Theoretically,

bridging inferences are drawn in order to establish coherence between implicit information from an introductory sentence and explicit information from the actual question. The purpose of these introductory sentences is usually to provide a context for the question. Practically, however, understanding (or even reading) these sentences is often not a prerequisite for answering the questions (i.e., introductory sentences do not necessarily determine the question focus). In these cases, respondents may refrain from establishing coherence between the introductory sentence and the question. Hence, bridging inferences may only reduce question comprehensibility if the introductory sentence contains implicit information which is crucial for understanding and answering the question.

The introductory sentences of the BINF questions used in the present studies did not contain essential information for answering the questions. Thus, from a methodological point of view, they may not have been appropriate for eliciting the detrimental effects of bridging inferences on question comprehensibility. It is also possible that the comprehension difficulties imposed by bridging inferences are so subtle that making them observable requires a larger sample of respondents than the one used in study 2. With regard to the significant effects of BINF found in study 1, it is likely that these are methodological artifacts, particularly in light of the eye-tracking results of study 2. All in all, further research is needed to pinpoint the specific effects of bridging inferences on question comprehensibility.

Summarizing, study 1 and study 2 found strong evidence that survey question comprehensibility is reduced by the following six text features: low-frequency words (LFRW), vague or imprecise relative terms (VIRT), vague or ambiguous noun phrases (VANP), complex syntax (CSYN), complex logical structures (CLOG), and low syntactic redundancy (LSYR). Moreover, study 2 identified these effects for attitudinal, factual, and behavioral questions, and thus irrespective of question type. The analyses per item show specifically which instances of the text features had the strongest impact on question comprehensibility and allow for formulating specific guidelines for survey question design.

### 8.1.2. Effects of question comprehensibility on response quality

The second overall research question was examined in study 1 (Chapter 5) and study 3 (Chapter 7). Question comprehensibility was operationalized by the seven psycholinguistic text features and the effects of these features on response quality were examined with the following indicators: drop-out rates, very short response times, acquiescence, primacy effects, neutral (midpoint) responses, non-substantive responses, and over-time consistency. In addition to examining the main effects of the text features on response quality indicators, study 3 included two moderator variables and examined whether there are interaction effects of question comprehensibility with verbal intelligence and motivation.

In both studies question comprehensibility had no significant effects on drop-out rates. Even though more respondents dropped out of the surveys when they received text feature questions, the decision to quit answering the survey was not explicitly related to the cognitive effort imposed by the questions. Insofar as drop-out is mediated by respondent motivation, it is likely that both samples in study 1 and study 3 consisted of highly motivated respondents who would try to complete the surveys irrespective of the cognitive effort required to do so. Indicators of these high levels of motivation are, for example, the low initial response rates in both survey (28.9% and 10.9%, respectively), suggesting that only a small proportion of highly motivated respondents participated in the surveys. Moreover, respondents did not receive any incentives and completed the survey for no apparent reward. Finally, they were drawn from online access panels and were presumably experienced in answering (poor) questionnaires.

Study 1 also examined whether question comprehensibility increases the number of respondents who rush through a survey (very short response times), who agree with assertions (acquiescence), and who select one of the first answer options presented (primacy effects). The text features had no effects on these three response quality indicators. With regard to very short response times, we assumed that respondents may start to rush through a survey if they find it burdensome to answer the questions thoroughly. However, in order to detect this kind of behavior, respondents either need to switch to this response strategy quite early in

the questionnaire or the questionnaire must be considerably longer than the one we used in study 1 (28 questions). Otherwise, the additional time required to answer the less comprehensible questions in the beginning of a survey may counterbalance the shorter response times provided later in the survey. Moreover, there is reason to believe that the sample consisted of highly motivated respondents and it seems unlikely that these respondents would have rushed through the survey, particularly with regard to the fact that they received no incentive for completing it. Instead, it seems that the respondents tried to cope with the cognitive demands of the text feature questions, which in turn required longer response times.

The non-significant findings regarding acquiescence and primacy effects may also be due to the characteristics of the sample used in this study. According to Krosnick (1991), question difficulty (i.e., comprehensibility) may not necessarily affect response quality if respondents are highly motivated or high in cognitive ability. As was mentioned above, it is very likely that the sample in study 1 consisted of highly motivated respondents. With regard to cognitive ability, 66.9% of the respondents received 12 or more years of schooling, suggesting that higher educated individuals were overrepresented in the sample. A final explanation could be that primacy effects are unlikely to occur in responses to the types of questions (or rather the types of answer formats) used in this study. All of our questions asked respondents to answer either on 3- to 5-point rating scales or on short lists of categorical response options. However, previous research on primacy effects found that these effects are more pronounced for questions involving longer lists of response categories (Galesic et al., 2008).

With regard to neutral (midpoint) responses, study 1 and study 3 showed that text feature questions significantly increase the number of neutral responses. In addition, study 3 showed that this effect depends upon respondents' level of motivation for answering survey questions: low question comprehensibility only increases the number of neutral responses for respondents low in motivation but not for highly motivated respondents. These findings indicate that when confronted with incomprehensible questions respondents with low motivation to

answer survey questions prematurely decide that they do not have or do not want to generate an opinion about the issue in question. Instead, these respondents seem to adjust their response strategy and to satisfice by selecting a neutral response even though they may have been able to report an opinion.

Study 3 also found that question comprehensibility affects the number of non-substantive responses: respondents receiving the text feature questions provided significantly more non-substantive responses than respondents answering control questions. This effect was moderated by their level of verbal intelligence and was more pronounced among respondents with limited verbal skills. When responding to questions that are difficult to comprehend, respondents with limited verbal skills seem to decide prematurely that they do not have the necessary information to answer the questions if they already have problems to understand what they are about. Hence, they may not bother trying to interpret the questions correctly but may satisfice by providing a non-substantive response.

Finally, study 3 revealed that the text feature questions significantly reduce the reliability of responses as indicated by lower over-time consistency of the answers across two surveys asking identical questions. The over-time consistency of responses is a relatively direct measure of data quality which suggests that the responses to the text feature questions are to some extent inaccurate and increase the measurement error in the survey data.

Taken together, these findings indicate that response quality is reduced if questions are difficult to comprehend and exceed the processing effort that respondents are willing or able to invest during survey responding. All the more, survey designers should avoid the problematic text features discussed above when writing questions. At the same time, it is important to note that study 3 did not examine the specific effects of the individual text features on response quality but only their aggregated effect. Analogical to the different impact levels of the text features on question comprehensibility, the individual text features may also affect response quality to variable degrees.

## 8.2 Implications

The findings summarized above have some practical implications for survey question evaluation and design. With regard to survey question evaluation, one could argue that the text feature *low syntactic redundancy* should be included into QUAID (University of Memphis, n.d.) given that it was found to reduce question comprehensibility. An extension of QUAID's five components would increase the validity of this tool in identifying questions that are difficult for respondents to comprehend.

With regard to survey question design, the findings imply that, whenever possible, survey designers should try to minimize the cognitive effort required to comprehend a question by avoiding the following six problematic text features: low-frequency words, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, complex logical structures, and low syntactic redundancy. The analyses on the item-level reported above suggest a number of specific recommendations on how to enhance the comprehensibility of survey questions. The following recommendations may supplement the existing guidelines of asking questions, lend further precision to these rules, and help practitioners to systematically check and improve the comprehensibility of their questions:

1. Avoid the use of **low-frequency words** that are relatively uncommon in written and spoken language. Consult linguistic thesauruses that include up-to-date word frequency lists and look for higher frequency synonyms to replace the low-frequency words.
2. Write out all words in the question and avoid **acronyms**.
3. Avoid the use of **vague quantification terms** and **vague frequency terms** in the question stems, because these refer to imprecise points on continuums. Instead, ask respondents to report an absolute metric when you are interested in the quantity or frequency of something.

4. Avoid the use of **biased ambiguous nouns** in their non-dominant meaning. Consult linguistic thesauruses to check whether a word has more than one meaning of which one is more dominant than the other.
5. Avoid the use of **abstract nouns**. Consult linguistic thesauruses to determine the hypernym value of a word and replace an abstract noun by a more specific hyponym of the word.
6. Avoid **left-embedded syntactic structures** (i.e., questions beginning with many subordinate clauses embedded in the main clause). Instead, first present the main clause (e.g., assertion or question) and subsequently add clauses and phrases that qualify the first clause.
7. Avoid **ambiguous syntactic structures** and ensure that all words in the question can be assigned to distinct linguistic categories (e.g., noun phrase, verb phrase, prepositional phrase, etc.).
8. Avoid the use of **dense noun phrases** (i.e., nouns which are supplemented by many adjectives and adverbs). Instead, attach subordinate clauses to the main clause of the question to narrow down the precise meaning of the noun.
9. Avoid asking **hypothetical questions** that are not grounded in the real world. Re-formulate these questions so that they relate to the concrete circumstances and experiences of the respondents.
10. Avoid asking questions that contain **numerous logical operators** such as *or*. Instead, consider to split the question into two or more questions containing less logical operators.
11. Avoid **nominalizations** (i.e., verbs that have been transformed into nouns) and replace these by active verbs.
12. Avoid asking questions in the **passive voice** and change passive to active constructions.

## 8.3 Suggestions for future research

The main goal of this thesis was to gain deeper insights into the relationship between the cognitive effort required to comprehend survey questions and the quality of respondents' answers. The findings presented above indicate that question comprehensibility has a strong impact on response quality and that survey designers should try to minimize the cognitive effort required to comprehend their questions in order to obtain high-quality answers. At the same time, there are some limitations to the studies presented above which suggest directions for future research.

First, none of the three empirical studies was conducted on a probability sample of respondents, and hence the findings cannot simply be generalized to the general (German) population. As was already mentioned in the discussions of the individual studies, it is very likely that similar (if not more definite) findings were obtained in studies using probability samples. For example, study 2 (Chapter 6) was conducted with relatively young and highly-educated respondents which were likely to be very competent readers. The question comprehensibility effects found in this study may have been even larger if the study had been conducted on a more diverse sample including less proficient readers as well. It is also important to note that the focus of the current studies was on randomization (i.e., on the analysis of differences between experimental treatments) rather than generalization to a population (cf. Couper, Tourangeau, Conrad, & Crawford, 2004; Kish, 1987). Nevertheless, it may be worth conducting this research on a probability sample of the German population to examine if the effects also hold for the general population and whether question comprehensibility affects the responses of different subgroups to different degrees.

Second, the response quality measures used in study 1 and study 3 are indicators of measurement error but not of response bias. Measurement errors are "deviations of the answers of respondents from their true values on the measure" (Groves, 1991, p. 2). For example, to the extent that a respondent provides a "don't know" response in a situation in which a more comprehensible question would elicit a substantive answer, the answer provided departs from the true value

for that respondent. If many respondents answered “don’ know” in responses to incomprehensible questions, then the estimates of means would be inaccurate. Certainly, measurement errors like these jeopardize the validity of survey results. However, if these errors are unsystematic, they do not necessarily lead to wrong conclusions drawn on basis of the survey data. In other words, they do not necessarily lead to response bias, which is a systematic response effect on the direction of the answers. Future research studies may examine the effect of question comprehensibility on response bias. Bias would occur, for example, if the answers of respondents with limited verbal skills would have differed from those of other respondents had they not responded “don’t know.” In this case, the resulting means would either overestimate or underestimate the true value in the population. It may be fruitful for future research studies to compare the response distributions resulting from different question versions and to examine whether different conclusions would be drawn on basis of these data. Furthermore, it would be interesting to apply a within-subject design and to examine whether the same respondents give different answers to the two question versions, and if so, whether these differences are systematic.

Finally, future research may extend the present findings by examining the impact of additional text features on survey question comprehensibility and response quality. The seven text features examined in this thesis do not necessarily exhaust the total set of text features which affect the cognitive effort required to comprehend a question. In light of the present research findings, it may be worthwhile to strive for a more complete understanding of the various factors that influence question comprehensibility. Minimizing respondent burden and making it easy for respondents to process and answer survey questions is certainly an effective strategy for obtaining accurate responses, and hence valid data.

# REFERENCES

Adams, P. F., Hendershot G. E., & Marano M. A. (1999). Current estimates from the National Health Interview Survey 1996. National Center for Health Statistics. *Vital Health Statistics, 10* (200).

Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Akiyama, M. M., Brewer, W. F., & Shoben, E. J. (1979). The yes-no question answering system and statement verification. *Journal of Verbal Memory and Verbal Behavior, 18,* 365-380.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D., & Hitch, G. T. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283-316.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler, & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (Vol. 2, pp. 285–375). Amsterdam: Elsevier.

Bassili, J. N. (1996). The how and the why of response latency measurement in telephone surveys. In N. Schwarz, & S. Sudman (Eds.), *Answering questions* (pp. 319–346). San Francisco, CA: Jossey-Bass.

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

Belson, W. A. (1968). Respondent understanding of survey questions. *Polls, 3,* 1-13.

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot: Gower.

Bishop, G., Tuchfarber, A., & Oldendick, R. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly, 50*, 240-250.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. R., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. [Need for cognition: A scale measuring engagement and happiness in cognitive tasks.]. *Zeitschrift für Sozialpsychologie, 25*, 147-154.

Bloomer, A., Griffiths, P., & Merrison, A. J. (2005). *Introducing language in use. A Coursebook.* London: Routledge.

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly, 43*, 92-101.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions* (2nd ed.). San Francisco, CA: Jossey-Bass.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131.

Cannel, C., Miller, P., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology 1981* (pp. 389-437). San Francisco: Jossey-Bass.

Cantril, H. (1944). *Gauging public opinion*. Princeton, NJ: Princeton University Press.

Carroll, D. W. (2004). *Psychology of language*. Belmont, CA: Thomson-Wadsworth.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472-517.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 15-48). New York: Russell Sage Foundation.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Coleman, E. B. (1964). The comprehensibility of several grammatical transformations. *Journal of Applied Psychology, 48,* 186–190.

Collani, G. (2009). Eine Deutsche Skala zum Bedürfnis nach Bewertung (Need to Evaluate). [A german scale for the need to evaluate]. In A. Glöckner-Rist (Ed.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 13.00*. Bonn: GESIS.

Colombo, L., Pasini, M., & Balota, D. A. (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition, 34,* 1312–1324.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Beverly Hills, CA: Sage.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004). What they see is what we get. *Social Science Computer Review, 22,* 111–127.

Draisma, S., & Dijkstra, W. (2004). Response latency and (para)linguistic expressions as indicators of response error. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 131–147). Hoboken, NJ: Wiley.

Duffelmeyer, F. A. (1979). The effect of rewriting prose material on reading comprehension. *Reading World, 19,* 1–16.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27,* 429–446.

Ericsson, K. A., & Kintsch, W. A. (1995). Long-term working memory. *Psychological Review, 102,* 211–245.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behaviour Research Methods, 39,* 175-191.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick, & M. S. Clark (Eds.), *Research methods in personality and social psychology (review of personality and*

*social psychology)* (Vol. 11, pp. 74–97). Newbury Park, CA: Sage Publications.

Federico, C. M., & Schneider, M. C.( 2007). Political expertise and the use of ideology: Moderating effects of evaluative motivation. *Public Opinion Quarterly, 71,* 221-252.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348–368.

Fillmore, C. J. (1999). A linguistic look at survey research. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 183-198). New York: Wiley.

Fink, A. (1995). *How to ask survey questions*. Thousand Oaks, CA: Sage.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: theory and practice in social research.* Cambridge: Cambridge University Press.

Forster, K. I. (1970). Visual perception on rapidly presented word sequences of varying complexity. *Perception & Psychophysics, 8,* 197–202.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627–635.

Forster, K. I., & Olbrei, I. (1974). Semantic heuristics and syntactic analysis. *Cognition, 2,* 319–347.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item and position upon decision times. *Journal of Verbal Learning and Verbal Behavior, 8,* 457–462.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly, 56,* 218-231.

Fowler, F. J. (1995). *Improving survey questions.* Thousand Oaks: Sage.

Fowler, F. J. (2001). Why it is easy to write bad questions. *ZUMA-Nachrichten, 48,* 49-66.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178–210.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22*, 313-328.

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

Galesic, M., & Yan, T. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester, & L. Kaczmirek (Eds.), Social and Behavioral Research and the Internet (pp. 349-370). New York, NY: Routledge Academic.

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods, 2*, 21–32.

Garrod, S., Freudenthal, S., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language, 33*, 39–68.

Globalpark. (2007). EFS survey [computer software, pc]. Hürth: Author.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior, 19*, 135–151.

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.

Graesser, A. C., McMahen, C. L., & Johnson, B. K. (1994). Question asking and answering. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 517-538). San Diego, CA: Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Groves, R. M. (1991). Measurement error across disciplines. In P. P. Biemer, R.M. Groves, L. E. Lyberg, N. M. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp.1-25). New York: John Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: Wiley.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior, 13,* 512–521.

Horning, A. S. (1979). On defining redundancy in language: Case notes. *Journal of Reading, 22,* 312–322.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40,* 431–439.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70,* 172-194.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122–149.

Kaczmirek, L., & Faaß, T. (2008). Data quality of paradata: A comparison of three response time measures in a randomized online experiment. *Poster presented at the annual meeting of the General Online Research conference (GOR),* March 10-12, Hamburg, Germany.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition, 2,* 15–47.

Kintsch, W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5,* 257–279.

Kish, L. (1987). *Statistical design for research*. New York: John Wiley.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics, 13,* 181–199.

Kohn, M. L. (1969). *Class and conformity: A study in values*. Homewood, IL: Dorsey Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213-236.

Krosnick, J. A. & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of 'no opinion' response options on data quality. *Public Opinion Quarterly, 66,* 371-403.

Lenzner, T. (in press). Effects of survey question comprehensibility on response quality. *Field Methods*.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23,* 361-373.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24,* 1003-1020.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz, & S. Sudman (Eds.), *Answering questions* (pp. 259–291). San Francisco, CA: Jossey-Bass.

Levelt, W. J. M. (1992). Psycholinguistics: An overview. In W. Bright (Ed.), *International Encyclopedia of Linguistics* (Second Edition/Vol. 3, pp. 290-294). Oxford University Press.

Loftus, E. F., Smith, K. D., Klinger, M. R., & Fiedler, J. (1992). Memory and mismemory for health events. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 102-137). New York: Russell Sage Foundation.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

McCutchen, D., Dibble, E., & Blount, M. M. (1994). Phonemic effects in reading comprehension and text memory. *Applied Cognitive Psychology, 8,* 597–611.

Morgan, J. L., & Green, G. M. (1980). Pragmatics and reading comprehension. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 113–140). Hillsdale, NJ: Erlbaum.

Mosier, C. I. (1941). A psychometric study of meaning. *Journal of Social Psychology, 13,* 123–140.

Myers, J. L., Cook, A. E., Kambe, G., Mason, R. A., & O'Brien, E. J. (2000). Semantic and episodic effects on bridging inferences. *Discourse Processes, 29,* 179-199.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–60). San Diego, CA: Academic Press.

Payne, S. L. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly, 73,* 74-97.

Poe, G. S., Seeman, I., McLaughlin, J., Mehl, E., & Dietz, M. (1988). Don't know boxes in factual questions in a mail questionnaire. *Public Opinion Quarterly, 52,* 212-222.

Porst, R. (2008). *Fragebogen. Ein Arbeitsbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

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

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191–201.

Rayner, K., Pacht, J. M., & Duffy, S. A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of Memory and Language, 33,* 527–544.

Rayner, K., & Pollatsek, A. (2006). Eye-movement control in reading. In M. J. Traxler, & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (Vol. 2, pp. 613–658). Amsterdam: Elsevier.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: Wiley.

Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly, 55*, 395-423.

Schmidt, K.-H., & Metzler, P. (1992). *Wortschatztest (WST) [Vocabulary test (WST)]*. Weinheim: Beltz Test GmbH.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61,* 576-602.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys.* New York: Academic Press.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Mahwah, NJ: Erlbaum.

Smith. A. F., & Jobe, J. B. (1994). Validity of reports of long-term dietary memories: Data and a model. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 121-140). Berlin: Springer-Verlag.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter. *Public Opinion Quarterly, 51*, 75-83.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Spyridakis, J. H., & Isakson, C. S. (1998). Nominalizations vs. denominalizations: Do they influence what readers recall? *Journal of Technical Writing and Communication, 28,* 163–188.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität [The social desirability scale-17 (SDS 17): Development and first findings on reliability and validity]. *Diagnostica, 45,* 173-77.

Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H. J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123-148). New York: Springer-Verlag.

Sturgis, P., & Smith, P. (2010). Assessing the validity of generalized trust questions: What kind of trust are we measuring? *International Journal of Public Opinion Research, 22*, 74-92.

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

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Toepoel, V., Vis, C., Das, M., & van Soest, A. (2009). Design of Web questionnaires. An information-processing perspective for the effect of response categories. *Sociological Methods & Research, 37*, 371-392.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R. (2004). Experimental design considerations for testing and evaluating questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 209-224). Hoboken, NJ: Wiley.

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

University of Memphis. (n.d.). Question understanding aid. Retrieved 31 May 2011 from http://mnemosyne.csl.psyc.memphis.edu/QUAID/quaidindex.html

Velez, P., & Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods, 1*, 69–74.

Vonk, W., & Noordman, L. G. M. (1990). On the control of inferences in text understanding. In D. A. Balota, G. B. F. d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 447–464). Hillsdale, NJ: Lawrence Erlbaum.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17,* 143–154.

Whisman, M. A., & McClelland, G. H. (2005). Designing, testing, and interpreting interactions and moderator effects in family research. *Journal of Family Psychology, 19,* 111-120.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16,* 312-339.

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.

Ziegler, M., & Kemper, C. J.. (2010). Wortschatz- und Overclaiming-Test (WOCT) [Vocabulary and overclaiming test (WOCT)]. Manuscript in progress.