

Psycholinguistic Determinants of Question Difficulty: A Web Experiment

Timo Faaß¹, Lars Kaczmirek² and Alwine Lenzner³

¹GESIS-ZUMA, e-mail: timo.faass@gesis.org

²GESIS-ZUMA, e-mail: lars.kaczmirek@gesis.org

³University of Koblenz-Landau, e-mail: lenzner@uni-landau.de

Abstract

Question difficulty poses a serious threat to the reliability and validity of survey data. However, very little is known about the factors that determine question difficulty. Only recently have survey researchers started to look at specific text features in order to explain why some questions are easier to comprehend or have less cognitive burden than others. Theoretical and empirical evidence from psycholinguistics suggests that these features (e.g. low-frequency words, vague noun-phrases, left-embedded syntax) cause comprehension difficulties and can thus have a strong impact on survey response quality. In order to examine the effects of these text features on question difficulty, we conducted an online experiment in which two versions of the same questions were compared using response latencies. Response latencies are assumed to reflect the cognitive effort that is required to answer a survey question with poorly formulated questions resulting in longer response times. One group (n = 495) received well-formulated survey questions, the other group (n = 490) answered questions which were suboptimal with respect to several psycholinguistic text features. Consistent with our predictions, the results reveal a significant difference between both conditions for most questions. Moreover, the answer distributions of several questions differed significantly between conditions, a finding which is quite alarming given that the question alternatives are supposed to measure the same concept. Hence, questionnaire designers are advised to pay attention to these text features when crafting questions.

Keywords: Question wording effects, Online experiment, Response latency

1. Introduction

The difficulty of a survey question is known to be a serious source of response error (Bless, Bohner, Hild, & Schwarz, 1992; Knäuper, Belli, Hill, & Herzog, 1997; Velez & Ashworth, 2007). If questions are difficult to understand respondents are likely to arrive at different interpretations (Belson, 1981), to *satisfice* (i.e. to provide satisfying rather than optimal answers; Krosnick, 1991) or to give incorrect answers (Schober & Conrad, 1997). Moreover, they may become frustrated and may eventually refuse to answer any further question of the survey (Ganassali, 2008). Thus, an important objective in survey

question design is to write comprehensible questions that respondents find easy to answer accurately.

Despite the fact that most survey researchers certainly agree with this statement, comparatively little is known about the various factors that determine question difficulty. Whereas some factors such as question content (e.g. Belson, 1981), question context (e.g. Krosnick & Alwin, 1987; Tourangeau & Rasinski, 1988) and the response options (e.g. Clark & Schober, 1992; Schwarz, Hippler, Deutsch, & Strack, 1985) have been studied quite extensively, the ways in which the concrete *wording* of a question contributes to question difficulty has received less attention. For a long time, the only general rules about question wording offered to questionnaire designers were the so-called “guidelines,” “standards,” or “principles” of asking survey questions. These are rather vague suggestions emphasizing, for example, the need to avoid long or complex questions, unfamiliar terms, and questions that call for a lot of respondent effort (e.g. Belson, 1981; Bradburn, Sudman, & Wansink, 2004; Fink, 1995; Fowler, 1995). Even though the guidelines are useful in avoiding gross mistakes, their major drawback is that they are never explicitly defined. Hence, it is up to the survey designer’s subjective interpretation to decide what constitutes a complex question or an unfamiliar term.

Only recently have survey researchers turned to the examination of specific text features in order to explain why some questions are easier to comprehend or have less cognitive burden than others (Graesser, Cai, Louwerse, & Daniel, 2006; Lessler & Forsyth, 1996; Tourangeau, Rips, & Rasinski, 2000). Theoretical and empirical evidence from psycholinguistics suggests that these features (e.g. low-frequency words, vague noun-phrases, left-embedded syntax) can cause comprehension difficulties and can thus have a strong impact on survey response quality. Generally speaking, comprehending a question involves two processes which cannot be separated: decoding semantic meaning and inferring pragmatic meaning. Both have to work smoothly if a question is to be understood and answered correctly. The greater the effort of decoding and inferring required the more likely is it that respondents do not come up with an optimal answer. 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 purpose of this paper is twofold. First, we provide an overview of psycholinguistic text features that have been identified to be closely linked to comprehension difficulty. By reviewing findings from various disciplines concerned with the psychology of reading, we aim at establishing a more sophisticated basis for the formulation of survey questions. Second, we report a study that was conducted to examine the effects of these text features on question difficulty.

2. Psycholinguistic Text Features

Evidence from various disciplines such as psycholinguistics, computational linguistics, cognitive psychology and artificial intelligence suggests that writers can enhance their comprehensibility to readers by paying attention to certain text features. The seven features that we selected in this study (low-frequency words, vague or imprecise relative terms, vague or ambiguous noun-phrases, complex syntax, working memory overload,

low syntactic redundancy, bridging inferences) do not necessarily exhaust the total set of relevant features. However, we believe that these are very important determinants of question difficulty. The first five features are similar to those incorporated into the Question Understanding Aid (QUAID; Graesser et al., 2006). QUAID is a computer tool that identifies problematic questions with respect to comprehension difficulty. The analytical detail of QUAID is unique in the questionnaire design literature and it provides an elaborate foundation for assessing and improving survey questions. It is currently available as a web facility (University of Memphis, n.d.).

Evidence from reading research, however, suggests that there are at least two more variables that affect comprehension difficulty to a similar degree, namely, low syntactic redundancy and bridging inferences. Incorporating these into QUAID might enhance the validity of this tool and cover additional aspects of the comprehensibility of survey questions.

2.1 Low-frequency Words

The frequency of a word (i.e., the number of times it occurs in large text corpora) has been one of the most investigated variables in reading research. It is a well-known finding that high-frequency words require less processing time and are thus easier to comprehend than rare words and words of medium frequency (e.g. Just & Carpenter, 1980; Mitchell & Green, 1978; Morton, 1969). This phenomenon is referred to as the *word frequency effect* and 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). Empirical evidence suggests that comprehension is impeded by low-frequency words, that is, people are slower at accessing these words and must work harder to comprehend sentences in which they occur. Consequently, low-frequency words such as technical terms, abbreviations, acronyms, and rare words should be avoided in survey questions. The following example, comparing a question with a low-frequency word to the same question using a high-frequency word illustrates this point:

- (1) Do you agree or disagree with the following statement?
The social *discrepancies* in Germany will certainly continue to exist.
- (2) Do you agree or disagree with the following statement?
The social *differences* in Germany will certainly continue to exist.

2.2 Vague or Imprecise Relative Terms

Vague or imprecise relative terms are predicates whose meanings are relative rather than absolute, as it is the case with quantitative adjectives or adverbs, for instance. They implicitly refer to an underlying continuum, however, the point on the continuum may be vague or imprecise. For example, adverbs such as *often* and *frequently* are imprecise relative terms. How often does an event need to occur in order to count as *often*? How frequent is *frequently*? And what is the difference between *often* and *frequently*? Clearly, this depends on the event that is being counted. 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, i.e., 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 stems, 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. Again, two question alternatives may illustrate this matter:

- (3) Do you agree or disagree with the following statement?
I *seldom* abstain from eating meat.
- (4) How often do you abstain from eating meat?
Always, Often, Sometimes, *Seldom*, Never.

When respondents are asked whether they seldom abstain from eating meat (as in (3)), without more information on how the adjective *seldom* is used in this context, no one - except vegetarians - can certainly say “yes” or “no”.

2.3 Vague or Ambiguous Noun-phrases

This term refers to noun-phrases, nouns, or pronouns which have an unclear or ambiguous referent. Abstract nouns, for example, often have unclear referents. This can be explained by their low hypernym value. A hypernym is a word that encompasses more specific words (hyponyms). For example, the hypernym *flower* encompasses the hyponyms *rose* and *tulip*. Every word can be assigned a hypernym value, which is low for abstract words and high for concrete words. In general, abstract words are more likely to be vague than concrete words and should be avoided in survey questions.

Ambiguous words have multiple senses associated with a single orthographic form (i.e., are polysemic), so that 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 and biased ambiguous words such as *bank*, which have one highly dominant meaning. Several studies (Duffy, Morris, & Rayner, 1988; Rayner, Pacht, & Duffy, 1994) 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*). This is explained by the fact that the context activates the non-dominant meaning while the 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 words, biased ambiguous words should be avoided in survey questions.

A second form of ambiguous noun-phrases are ambiguous pronouns. Because of the fact that in written communication the writer is not present during reading there is basically no deictic use of pronouns or adverbs. Words such as *it*, *they*, *here*, *there*, and *this* “always refer anaphorically, that is, to something the writer has previously introduced explicitly or implicitly” (Morgan & Green, 1980). Hence, the task of connecting an anaphoric element to its antecedent in the text (antecedent search) is central to reading comprehension. When readers come across a pronoun such as *it*, they must identify an antecedent that matches it. If there is 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 (as in (5)), the pronoun is ambiguous and antecedent search might take longer.

- (5) 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*?
- (6) 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*?

2.4 Complex Syntax

According to current linguistic theories, syntax can become complex for two reasons: either the structures are ambiguous, lead to a wrong interpretation, and have to be corrected; 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 they see a word, people 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:

- (7) 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) found that the number of propositions influences the time required to read a passage. Consider the following two sentences:

- (8) Cleopatra's downfall lay in her foolish trust in the fickle political figures of the Roman world.
- (9) 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, sentence (8) took longer to read than sentence (9) (Kintsch & Keenan, 1973). This result is explained by the fact that (8) is propositionally more complex (eight propositions) than (9), which contains four propositions. 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 there are too 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. For example:

- (10) Do you agree or disagree with the following statement?
Even if the government does not agree with certain decisions, Germany as a member of international organizations *should* generally *follow* their decisions.
- (11) Do you agree or disagree with the following statement?
In general, Germany *should follow* the decisions of international organizations to which it belongs, even if the government does not agree with them.

2.5 Working Memory Overload

There is wide agreement on the fact that working memory capacity is limited (Baddeley, 1986; Ericsson & Kintsch, 1995; Just & Carpenter, 1992) and that people's working memory limitations affect the ease with which sentences are processed (Chomsky & Miller, 1963; Kimball, 1973; MacDonald & Christiansen, 2002). If a sentence requires readers to hold a lot of information in mind at the same time, working memory may be overloaded and break down. This has already been mentioned in the examples dealing with left-embedded structures and anaphora.

Another form of working memory overload occurs in sentences with numerous logical operators such as *or*. Disjunctions (expressions with *or*) quickly overload working memory because the reader needs to keep track of different options and possibilities. Sentences with two or more *or*'s are difficult to comprehend because people need to construct a mental table of the different options. Consider, for example, question (12):

- (12) 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 context, do you think it should be allowed or not allowed to organize public meetings to protest against the government?

However, working memory overload cannot be reduced to long sentences. For example, a question like *How many hours did you spent last year doing the housework?* is comparably short but requires a quantitative mental calculation which imposes a high load on working memory. Similarly, hypothetical questions might be short but difficult to process because they are not grounded in the real world, requiring the respondent to build

a mental representation of the situation and hold it in memory while processing the rest of the question.

2.6 Low Syntactic Redundancy

Syntactic redundancy refers to the predictability of the grammatical structure of a sentence. It is supposed that the higher the level of syntactic redundancy of a text, the quicker and easier one can process and comprehend it. Besides the operations mentioned in the section on *complex syntax*, syntactic redundancy is increased by changing passive sentences to active sentences and by denominalizing nominalizations.

In passive constructions the object of an action is turned into the subject of the sentence. Passives thus emphasize the action rather than the agent responsible for the action. This change of perspective makes it harder for the reader to predict the course of action and thus harder to comprehend. For example, Forster and Olbrei (1974) asked their participants to judge whether a sample of active and passive sentences were grammatical or ungrammatical. They found that actives were faster identified as grammatical than were passives.

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:

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

These days, it is the government's responsibility to *enforce a restriction* of top managers' salaries.

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

These days, it is the government's responsibility to *restrict* top managers' salaries.

2.7 Bridging Inferences

It is widely agreed that writers do not make explicit everything that they want to communicate in a text. Thus, a text does always contain implicit information that the reader needs to compute from the text. This computation of implicit information is called inferencing. Drawing inferences is generally assumed to be a time-consuming process (Vonk & Noordman, 1990). In questionnaires, inferences usually come in the form of bridging inferences. These are drawn in order to establish coherence between the current information and previous information. In survey questions, bridging inferences are required when the actual question follows an introductory sentence, such as in (15):

- (15) The government recently passed the *Patriot Act*. Do you think the authorities should have the right to detain people for as long as they want without putting them on trial?

In order to establish coherence between the introductory sentence and the question, respondents need to draw a bridging inference: *the Patriot Act* must somehow provide the authorities with the right mentioned in the question; otherwise, the two sentences would not be connected.

3. The Present Study

3.1 Design and Hypotheses

In order to test whether these text features have the predicted effects on question difficulty we conducted an online experiment. One group (n = 495) received well-formulated survey questions, the other group (n = 490) answered questions which were suboptimal with respect to the seven text features illustrated above. Three indicators of question difficulty were chosen as dependent variables: response latency, answer distribution and drop-out rate.

Response Latency

Response latency has received increasing attention in the survey research literature over the last decade (Yan & Tourangeau, 2008) and has been found to be a good indicator 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 answer. Consequently, we hypothesize that the poorly formulated questions will produce longer response times than their well-formulated counterparts.

Answer Distribution

The difficulty of a survey question threatens the reliability and validity of the answers respondents provide. Difficult questions may produce different question interpretations among respondents (Belson, 1981), respondent satisficing (Krosnick, 1991), or incorrect answers (Schober & Conrad, 1997). Therefore, we expect the answer distributions of several questions to differ considerably between conditions, despite the fact that the question alternatives are supposed to measure the same concept.

Drop-out Rate

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 heavier cognitive load reduce respondent motivation. Therefore, we hypothesize that the drop-out rate in the suboptimal condition will be larger than in the control condition.

3.2 Method

3.2.1 Participants

Participants were randomly drawn from the online access panel Sozioland (Respondi AG). 5000 people were invited and 1445 respondents started the survey (28.9%). 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 receiving any substantial questions ($n = 71$), technical problems prevented the collection of their response latency times ($n = 6$), or they did not complete the survey ($n = 136$). The upper and lower one percentile was defined as outlier (Ratcliff, 1993), excluding another 20 respondents and leaving 985 respondents in the analysis. The participants were 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 (suboptimal condition, $n = 490$) vs. 257 males and 238 females (control condition, $n = 495$). Educational achievement between the two randomized groups did not differ significantly.

3.2.2 Questions

With the exception of a few questions that were designed by the first author, 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 35 countries.

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. 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. An important requirement for the comparability of the questions through response latencies was that they were virtually equal in length: the question alternatives did not differ in more than two syllables from each other. The only exception to this rule were questions, in which the well-formulated version was longer than the poorly-formulated one, thus not affecting the response time in favor of our hypotheses. As a result of these manipulations, the question wording deviates to a certain degree from the original wording of the ISSP questions. Nevertheless, using ISSP topics allowed us to ask ecologically valid questions which are common in social science research.

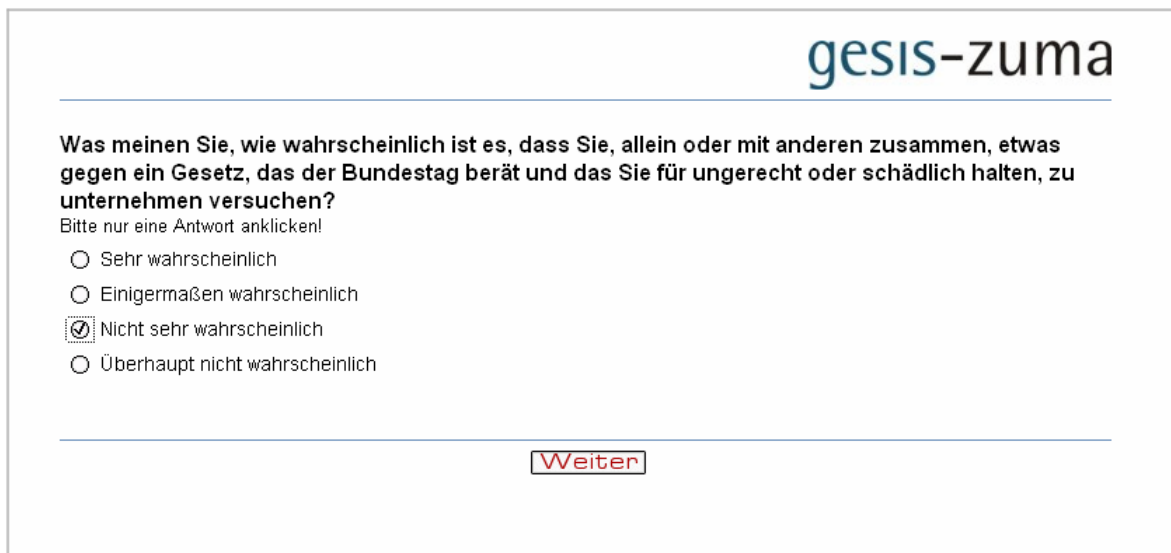
3.2.3 Procedure

The software used in this online study was EFS Survey (Globalpark, 2007), a software especially developed for conducting web-based surveys. We used JavaScript to measure response latencies. These were measured from the time a question was presented on the screen to the time an answer was selected using the computer mouse. The accuracy of the response latency times measurement was tested in a pretest and found to be very robust

and only minimally influenced by differing Web site loading times (Faaß, 2007; Kaczmirek, 2008).

Participants received a link to the online survey by e-mail. By clicking on this link, they were referred to the instruction page. Here they learned that the survey contained questions on a broad range of topics such as politics, society, and environment. The respondents were instructed to read each question in the given order and not to skip questions or return to a 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 at a time was displayed on the screen and participants had to use the computer mouse to mark their answers (see figure 1). The software automatically recorded the time lapse between the appearance of the question and the ticking of an answer option at millisecond accuracy. It also recorded the answer category selected by the respondent. Once an answer was given participants had to click on a next-button and the next question was presented. Participants were randomly assigned to one of the two conditions. First, respondents answered a series of background questions dealing with sex, age, and native language. Afterwards they received the 28 experimental questions in a random sequence in order to control for context effects of the question order. Finally, they answered additional background questions on education, work status, and speed of the Internet connection.



gesis-zuma

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?

Bitte nur eine Antwort anklicken!

- Sehr wahrscheinlich
- Einigermaßen wahrscheinlich
- Nicht sehr wahrscheinlich
- Überhaupt nicht wahrscheinlich

[Weiter](#)

Figure 1: Screenshot of an experimental question (Q16)

3.3 Results

3.3.1 Response Latencies

The distribution of response latencies is usually asymmetrical and skewed to the right (Ratcliff, 1993). The Kolmogorov Smirnov test for goodness of fit was used to check for the normality of distribution and confirmed that the response latencies were not normally distributed ($p < .0001$). Therefore, non-parametric tests were used in the analysis.

First, we compared the overall response latencies in the two conditions (i.e., the total time needed to answer the experimental questions). For the 28 questions as a whole, the mean response latency was 370.3 seconds ($SD = 150.2$) in the suboptimal condition and 341.5 seconds ($SD = 146.5$) in the control condition. A Mann Whitney U test on overall response latencies resulted in a significant difference between both groups ($z = -3.742$, $p < 0.0001$). Because of our strong hypotheses a one-sided testing approach was employed for the analysis of the individual questions. Mann Whitney U tests revealed that 21 questions differed significantly between conditions ($p < .05$); the latencies for 2 questions were marginally significant ($p < .1$); and the latencies for another 5 questions did not differ significantly between conditions ($p > .1$). With the exception of one question (Q 26), the response latencies were longer for the suboptimal questions. The results for the individual questions as well as the manipulated text feature for each question are described in table 1.

3.3.2 Answer Distribution

To analyze between-group differences in answer distribution, again Mann Whitney U tests were conducted because of the ordinal data. In some cases, the question alternatives were not comparable because of differences in the response options. For example, one question asked the 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 Example (3)). In total, modifications like these occurred in five questions, leaving 23 questions for the analysis of the answer distribution. The Mann Whitney U tests revealed significant differences between the two groups for 8 questions ($p < .05$). For example, participants systematically gave different responses to the question if they believe that the social differences/discrepancies in Germany will continue to exist, depending on whether the question included the low-frequency word *discrepancies* or its higher-frequency synonym *differences* ($z = -2.332$, $p < .05$). The results of the Mann Whitney U tests are illustrated in table 2.

3.3.3 Drop-out Rates

As was mentioned above, 136 participants (11,9%) dropped at some point from the survey. The drop-out rates were 13.2% ($n = 77$) for the suboptimal condition and 10.6% ($n = 59$) for the control condition. The results of a chi-square test for drop-out revealed no significant differences across conditions ($p = .123$).

Table 1: Results of the Mann Whitney *U* tests on response latencies

Item	<i>z-value</i>	<i>p (1-sided)</i>
Low-frequency words		
Q 01 low-frequency term	-2.512	.006**
Q 02 acronym	-4.447	.000***
Q 03 low-frequency term	-1.674	.048*
Q 04 low-frequency term	-3.564	.000***
Vague or imprecise relative terms		
Q 05 vague quantification term	-8.523	.000***
Q 06 imprecise relative term	-1.359	.087 ⁺
Q 07 vague temporal term	-.480	.316
Q 08 vague intensity term	-1.769	.039*
Vague or ambiguous noun-phrases		
Q 09 pronoun with multiple referents	-1.602	.055 ⁺
Q 10 abstract noun/hypernym	-.038	.485
Q 11 abstract noun/hypernym	-.540	.239
Q 12 ambiguous pronoun	-2.319	.010*
Complex syntax		
Q 13 left-embedded syntactic structure	-8.198	.000***
Q 14 ambiguous syntactic structure	-6.130	.000***
Q 15 dense noun-phrase	-4.283	.000***
Q 16 left-embedded syntactic structure	-.506	.308
Working memory overload		
Q 17 hypothetical question	-3.115	.001***
Q 18 quantitative mental calculation	-14.755	.000***
Q 19 hypothetical question	-11.685	.000***
Q 20 numerous logical operators	-6.420	.000***
Low syntactic redundancy		
Q21 nominalization	-2.524	.006**
Q22 nominalization	-3.393	.000***
Q23 passive	-2.442	.008**
Q24 passive	-2.764	.003**
Bridging inferences		
Q25 bridging inference required	-6.469	.000***
Q26 bridging inference required	-2.004	.023*
Q27 bridging inference required	-3.094	.001**
Q28 bridging inference required	-.553	.290

⁺ p < 0.1 * p < 0.05, ** p < 0.01, *** p < 0.001

Table 2: Results of the Mann Whitney *U* tests on answer distribution

Item	<i>z-value</i>	<i>p (2-sided)</i>
Low-frequency words		
Q 01 low-frequency term	-.525	.600
Q 02 acronym	-.940	.347
Q 03 low-frequency term	-2.332	.020*
Q 04 low-frequency term	-2.696	.007**
Vague or imprecise relative terms		
Q 05 vague quantification term	n.a.	n.a
Q 06 imprecise relative term	n.a.	n.a
Q 07 vague temporal term	-6.601	.000***
Q 08 vague intensity term	-7.852	.000***
Vague or ambiguous noun-phrases		
Q 09 pronoun with multiple referents	-.789	.430
Q 10 abstract noun/hypernym	-.311	.756
Q 11 abstract noun/hypernym	-.466	.641
Q 12 ambiguous pronoun	-.752	.452
Complex syntax		
Q 13 left-embedded syntactic structure	-2.062	.039*
Q 14 ambiguous syntactic structure	-.031	.975
Q 15 dense noun-phrase	-.059	.953
Q 16 left-embedded syntactic structure	-6.208	.000***
Working memory overload		
Q 17 hypothetical question	n.a.	n.a.
Q 18 quantitative mental calculation	n.a.	n.a.
Q 19 hypothetical question	n.a.	n.a.
Q 20 numerous logical operators	-13.293	.000***
Low syntactic redundancy		
Q21 nominalization	-1.171	.241
Q22 nominalization	-2.246	.025*
Q23 passive	-.223	.824
Q24 passive	-.543	.587
Bridging inferences		
Q25 bridging inference required	-1.402	.161
Q26 bridging inference required	-.050	.960
Q27 bridging inference required	-.257	.797
Q28 bridging inference required	-.150	.881

n.a.= not applicable, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. Discussion

This study examined how seven psycholinguistic text features affect the difficulty of survey questions. Using response latencies as a measure of the cognitive effort required to answer a survey question, we compared two versions of the same questions in a Web experiment. Additional dependent variables were answer distribution and drop-out rate.

The present findings largely confirm our hypotheses. First, the response latencies for most questions differed significantly between conditions: respondents answering the suboptimal questions had longer response latencies. The most considerable determinants of question difficulty were low-frequency words, working memory overload and low syntactic redundancy. Although the response time data largely support our predictions, a few questions did not produce the hypothesized effects. For example, abstract nouns with a low hypernym value did not result in longer response latencies than their more concrete counterparts (Q10, Q11). Abstract nouns certainly lead to variable interpretations among respondents, however, response latencies seem to be no sensitive measure for revealing these differences. Obviously, as a measure of cognitive processing depth, they are not sensitive for detecting differing question interpretations. In another question, it is very likely that the suboptimal question was so difficult to answer correctly that participants made a guess instead of an informed response. This seems to have happened in question Q07, for which the answer distribution differs significantly across conditions. People systematically gave different answers to the question if they already had a private Internet connection at the time of the New Economy crash depending on whether the question included the year of the economic crash or not ($z = -6.601$, $p < .0001$). The term *New Economy* seems to be so vague that without any temporal reference, respondents started to guess and chose the more probable “no” answer option.

The second hypothesis, that several question alternatives produce different answer distributions, was supported. Despite the fact that the question alternatives are supposed to measure the same concept, the answer distribution for 8 questions differed significantly between conditions. This means that in 35% of the questions, a slight change in question wording led to systematically biased responses. Given that the participants were randomly assigned to the two conditions, the differences between the answer distributions are likely to represent response error.

Contrary to our predictions, the manipulation of the questions did not affect the drop-out rate of the survey. The decision to quit answering the survey was therefore not related to the difficulty of the questions. It is possible that respondent characteristics such as motivation or cognitive ability as well as other features of the questionnaire (e.g. length) have a stronger influence on survey drop-out than question difficulty.

5. Conclusion

In accordance with our hypotheses we have demonstrated that the wording of a question and specifically, seven psycholinguistic text features are important determinants of question difficulty. Question difficulty, in turn, is a serious source of response error. Therefore, questionnaire designers are advised to pay attention to these text features when crafting survey questions.

Our findings have theoretical as well as practical implications. From a theoretical point of view, we found strong empirical evidence for effects of psycholinguistic text features on survey question difficulty. Given that all seven text features had the predicted effects on comprehension difficulty, we would argue for an extension of QUAID's five components and an inclusion of *low-syntactic redundancy* and *bridging inferences* into this tool. On the practical side, a specification of text features and their effect on question difficulty can help practitioners to write comprehensible questions. Manuals describing these text features in detail are likely to replace or at least to supplement the out-dated "guidelines" or "standards" of asking survey questions. Moreover, collecting response latencies appears to be a suitable approach for measuring the cognitive effort required to answer a survey question.

References

- Baddeley A. D. (1986) *Working Memory*, Oxford University Press, Oxford
- Bassili J. N. (1996) The how and the why of response latency measurement in telephone surveys, in: *Answering Questions*, Schwarz, N. & Sudman, S. (Eds.), Jossey-Bass, 319-346
- 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. (1981) *The Design and Understanding of Survey Questions*, Gower, Aldershot
- Bless H., Bohner G., Hild T., Schwarz N. (1992) Asking difficult questions: task complexity increases the impact of response alternatives, *European Journal of Social Psychology*, 22, 309-312
- Bradburn N., Sudman S., Wansink, B. (2004) *Asking Questions* (2nd ed.), Jossey Bass, San Francisco
- Chomsky N., Miller G. A. (1963) Introduction to the formal analysis of natural languages, in: *Handbook of Mathematical Psychology (Vol. 2)*, Luce, R. D., Bush, R.R. & Galanter, E. (Eds.), Wiley, 269-321
- Clark H. H., Schober, M. F. (1992) Asking questions and influencing answers, in: *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, Tanur, J. M. (Ed.), Russel Sage Foundation, 15-48
- Coleman E. B. (1964) The comprehensibility of several grammatical transformations, *Journal of Applied Psychology*, 48, 186-190
- Draisma S., Dijkstra W. (2004) Response latency and (para)linguistic expressions as indicators of response error, in: *Methods for Testing and Evaluating Survey Questionnaires*, Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J.T., Martin, E., Martin, J. & Singer, E. (Eds.), Wiley, 131-147
- 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
- Faaß T. (2007) The readability of survey questions: an interdisciplinary approach to improving question wording. Master of Arts Thesis, University of Mannheim.
- Fillmore C. J. (1999) A linguistic look at survey research, in: *Cognition and Survey Research*, Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M. & Tourangeau, R. (Eds.), Wiley, 183-198
- Fink A. (1995) *How to Ask Survey Questions*, Sage, Thousand Oaks
- Forster K. I. (1970) Visual perception on rapidly presented word sequences of varying complexity, *Perception and 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. (1995) *Improving Survey Questions*, Sage, Thousand Oaks
- 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
- Ganassali S. (2008) The Influence of the Design of Web Survey Questionnaires on the Quality of Responses, *Survey Research Methods*, 2, 21-32
- Graesser A. C., Cai Z., Louwse 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., Hoffman N. L., Clark L. F. (1980) Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior*, 19, 135-151
- Globalpark. (2007) *Surveycenter 5.1 [computer software, pc]*, Author, Hürth
- 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
- Kaczmarek L. (2008) Human-survey interaction: usability and nonresponse in online surveys. Doctoral Dissertation, University of Mannheim.
- 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
- Kimball J. (1973) Seven principles of surface structure parsing in natural language, *Cognition*, 2, 15-47
- 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
- Krosnick J. A., Alwin D. F. (1987) An evaluation of a cognitive theory of response-order effects in survey measurement, *Public Opinion Quarterly*, 51, 201-219
- Krosnick J. A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology*, 5, 213-236
- Lessler J. T., Forsyth B. H. (1996) A coding system for appraising questionnaires, in: *Answering Questions*, Schwarz, N. & Sudman, S. (Eds.), Jossey-Bass, 259-291
- MacDonald M. C., Christiansen M. H. (2002) Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996), *Psychological Review*, 109, 35-54
- Mitchell D. C., Green D. W. (1978) The effects of content on immediate processing in reading, *Quarterly Journal of Experimental Psychology*, 30, 29-63
- Morgan J. L., Green G. M. (1980), Pragmatics and reading comprehension. in: *Theoretical Issues in Reading Comprehension*, Spiro, R. J., Bruce, B. C. & Brewer, W. F. (Eds.), Erlbaum, 113-140
- Morton J. (1969) The interaction of information in word recognition, *Psychological Review*, 76, 165-178
- Ratcliff R. (1993) Methods for dealing with reaction time outliers, *Psychological Bulletin*, 114, 510-532

- 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
- Schober M. F., Conrad F. G. (1997) Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602
- Schwarz N., Hippler H., Deutsch B., Strack F. (1985) Response scales: Effects of category range on reported behavior and comparative judgements, *Public Opinion Quarterly*, 49, 388-395
- 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
- Tourangeau R., Rasinski K. A. (1988) Cognitive processes underlying context effects in attitude measurement, *Psychological Bulletin*, 103, 299-314
- Tourangeau R., Rips L. J., Rasinski K. (2000) *The Psychology of Survey Response*. Cambridge University Press, Cambridge
- University of Memphis. (n.d.). Question Understanding Aid. Retrieved 16 May, 2008 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: *Comprehension Processes in Reading*, Balota, D. A., d'Arcais, G. B. F. & Rayner, K. (Eds.), Erlbaum, 447-464
- Whaley C. P. (1978) Word-nonword classification time, *Journal of Verbal Learning and Verbal Behavior*, 17, 143-154
- 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