Cognitive burden of survey questions and response times: a psycholinguistic experiment

Lenzner, Timo; Kaczmirek, Lars; Lenzner, Alwine

Preprint / Preprint
Zeitschriftenartikel / journal article

Lenzner, Timo, Lars Kaczmirek, and Alwine Lenzner. 2010. "Cognitive burden of survey questions and response times: a psycholinguistic experiment". *Applied Cognitive Psychology* vol. 24, no. 7 1003-1020. doi:10.1002/acp.1602.

Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment

Timo Lenzner, Lars Kaczmirek, Alwine Lenzner

Submitted to Applied Cognitive Psychology, 18 Jun 2008. Accepted 25 Jun 2009. First published online 20 Aug 2009

Retrieve the published version from: http://dx.doi.org/10.1002/acp.1602

An important objective in survey question design is to write clear questions that respondents find easy to understand and to answer. This contribution identifies the factors that influence question clarity. Theoretical and empirical evidence from psycholinguistics suggests that specific text features (e.g., low-frequency words, left-embedded syntax) cause comprehension difficulties and impose a high cognitive burden on respondents. To examine the effect of seven different text features on question clarity, an online experiment was conducted in which well-formulated questions were compared to suboptimal counterparts. The cognitive burden of the questions was assessed with response times. Data quality was compared in terms of drop-out rates and survey satisficing behavior. The results show that at least six of the text features are relevant for the clarity of a question. We provide a detailed explanation of these text features and advise survey designers to avoid them when crafting questions.

1. Introduction[1]

Answering a survey requires respondents to invest a great deal of cognitive effort for little or no apparent reward (Krosnick, 1991). In order to guarantee that respondents are willing to invest this cognitive effort, it is important to construct the questions so that the required cognitive burden for answering the questions is kept at a minimum. Moreover, the cognitive burden 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;

[1] Acknowledgements: The Respondi AG provided the sample for this research. We therefore thank Otto Hellwig and Tom Wirth for their cooperation and support.

Foddy, 1993), to *satisfice* (i.e., to provide satisfying rather than optimal answers; Krosnick, 1991), to give incorrect answers (Schober & Conrad, 1997) or to refuse answering any further question of the survey (Ganassali, 2008). Consequently, an important objective in questionnaire design is to write clear questions and thus to minimize the cognitive effort required to process them.

Despite the importance of question clarity in survey design, only little is known about the factors that determine whether a question is easy or difficult to understand. For a long time, the only general rules about question wording offered to questionnaire designers were in the form of so-called "guidelines," "standards," or "principles" of asking survey questions. These are general and therefore 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 lack explicit definitions. Hence, it is up to the survey designer's subjective interpretation to decide what constitutes a complex question or an unfamiliar term. Thus, only highly experienced researchers are able to apply the guidelines correctly.

Only recently have survey researchers turned to the examination of specific text features in order to explain why some questions are difficult to comprehend or impose a high cognitive burden on respondents (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 relative terms, left-embedded syntax) cause comprehension difficulties and can thus have a strong impact on data quality.

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. This is done by extending a set of text features initially proposed by Graesser et al. (2006) which should be considered when crafting survey questions. Second, a Web experiment assessed the effects of these text features on the cognitive burden of survey questions and data quality. Cognitive burden was operationalized in terms of response time. Indicators for data quality were drop-out rates, "very short response times", amount of "no opinion" responses, acquiescence, and primacy effects.

2. Psycholinguistic Text Features

Evidence from various disciplines such as psycholinguistics, computational linguistics, cognitive psychology and artificial intelligence suggests that writers can reduce the cognitive burden on 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 important determinants of question clarity. The first five features are very similar to those incorporated into the Question Understanding Aid[2] (QUAID; Graesser et al., 2006).

[2] The first text feature incorporated into QUAID is termed "unfamiliar technical terms" and differs from our first variable "low-frequency words". In addition to word frequency, the QUAID variable does also compute semantic familiarity using the familiarity rating in the Coltheart's (1981) MRC Psycholinguistic Database. We did not include familiarity because this concept is too vague in the psycholinguistic literature and is a rather subjective measure of word difficulty (e.g., Colombo, Pasini, & Balota, 2006; Rayner & Pollatsek, 2006). Because frequency and familiarity are highly correlated (Balota, Yap, & Cortese, 2006), familiarity can largely be subsumed under frequency, especially if up-to-date frequency lists are used in determining word difficulty.

QUAID is a manual and computer tool that identifies problematic questions with respect to comprehension difficulty (University of Memphis, n.d.). The analytical detail of QUAID is unique in the questionnaire design literature and it provides an elaborate foundation for assessing and improving survey questions.

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 (Horning, 1979) and bridging inferences (e.g., Vonk & Noordman, 1990). Incorporating these into QUAID might enhance the validity of this tool and cover additional aspects of the comprehensibility of survey questions. The following sections summarize the text features that determine question clarity to provide the theoretical background for the experiment.

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 well-known 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). Ample 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 (Q3)[3] illustrates this point, comparing a question with a low-frequency word to the same question with a high-frequency word:

(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. The following questions illustrate the associated problems of vagueness. 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 (cf. Graesser et al., 2006). 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

[3] The number in parentheses indicates that the example was used as a question in the Web experiment. In the examples in this text the suboptimal version of the question is always followed by the well-formulated question. The term "well-formulated" does not imply that the question is optimal but only that – with regard to wording – the second formulation is preferable. The text features presented in this paper and the recommendations for their use tap only one of many aspects that have to be taken into consideration during survey question design.

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 (Q5) may illustrate this matter:

(3) I *seldom* abstain from eating meat.

Answer categories: Agree, Disagree

(4) How often do you abstain from eating meat?

Answer categories: 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, are able to certainly "agree" or "disagree".

2.3 Vague or Ambiguous Noun-phrases

This term refers to noun-phrases, nouns, or pronouns which have an unclear or ambiguous referent. Firstly, abstract nouns 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.

Secondly, ambiguous noun-phrases 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[4] and biased ambiguous words such as *bank*, which have one highly dominant meaning.[5] 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 used in their non-dominant meaning should be avoided in survey questions.

Thirdly, 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, 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 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. Consider the following example (Q9):

[4] 1. straw of wheat, 2. straw to suck up a drink
[5] 1. financial institution, 2. river bank

(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 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 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, Kintsch & Keenan (1973) have shown that sentence (8) takes longer to read than sentence (9). This result is explained by the fact that (8) is propositionally more complex (eight propositions) than (9), which contains four propositions[6]. 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 (Q16):

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

(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, the following question (Q20):

(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?

Nevertheless, 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?” requires relatively little working memory to comprehend the question. However, it requires a quantitative mental calculation which imposes a high load on working memory to reach a response. 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 (Horning, 1979). 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 grammatically correct. They found that actives were faster identified as being correct 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 (Q21):

(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 everything explicit that they want to communicate in a text. Thus, a text always contains implicit information that the reader needs 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 times increase with the number of inferences readers need to generate (e.g., Haviland & Clark, 1974; Just & Carpenter, 1980). In questionnaires, inferences of this sort usually come in the form of bridging inferences.[7] These are drawn in order to establish coherence between the current

[7] Inferences are also drawn in other situations during the answering process. For example, respondents may try to establish coherence between the different questions of a survey and respond to a question on the basis of an answer to an earlier question. Here, however, we focus on the computation of implicit information that is required within the question. The concept of bridging inferences as a text feature may be conceived as a subtype of inferences.

information and previous information. In survey questions, bridging inferences are required when the actual question follows an introductory sentence, such as in:

(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. Experiment

3.1 Design and Hypotheses

We conducted an online experiment in order to test whether these seven text features reduce the clarity of questions and increase the cognitive burden on respondents and how this affects data quality. 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 defined above. This is the main factor in the experiment and operationalizes the clarity of survey questions. Dependent variables were response time as a measure of cognitive burden and drop-out rate and survey satisficing as measures of data quality.

Response Time

Response time 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, that is, it measures the cognitive burden of a question.

Consequently, we hypothesize that the suboptimally formulated questions will produce longer response times than their well-formulated counterparts.

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 condition with well-formulated questions.

Survey Satisficing

The difficulty of a survey question threatens the quality of the answers respondents provide. According to satisficing theory (Krosnick, 1991), the likelihood that respondents provide low-quality data is a function of three factors: task 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 examine several indicators of satisficing across the two conditions (very short response times, "no opinion" responses, acquiescence and primacy effects) and expect to find more satisficing in the suboptimal condition.

3.2 Participants

Participants were randomly drawn from the online access panel Sozioland (Respondi AG). 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 being asked 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)[8]. For response times 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 (condition with well-formulated questions, n = 495). 65,1% of the participants had received twelve or more years of schooling, 20,1% received ten years and 14,8% received nine or less years of schooling. Educational achievement between the two randomized groups did not differ significantly.

3.3 Questions

With the exception of four questions that were designed by the first 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 35 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

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

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 German questions as well as a loose translation of the questions in English are attached as a supplement. The concrete rewriting rules for the suboptimal 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. Working memory overload: 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 well-formulated version was longer than the suboptimal one (Q7, Q10), thus not affecting the response time in favor of our hypotheses.

3.4 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 well-formulated questions or to the condition with suboptimal questions. First, respondents answered a series of background questions dealing with sex, age, and native language. Then they received 28 questions which were constructed with respect to the text features 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.

3.5 Results

3.5.1 Response Times

Response times were analyzed as an indicator 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 participants 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 (clarity of survey questions: well-formulated vs. suboptimal question formulations) analysis of covariance (ANCOVA) with reading rate[9] as a covariate. The total response time for a respondent during the

[9] 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

treatment was the sum over all 28 questions. The total mean response time was 370.3 seconds (*SD* = 150.2) in the suboptimal condition and 341.5 seconds (*SD* = 146.5) in the condition with well-formulated questions. 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 clarity 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 4 questions each as repeated measures and reading rate as a covariate. The results in table 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), working memory overload (MEMO), low syntactic redundancy (LSYR), 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 (α=.05) is p<=.007.

--- Table 1 around here ---

On the lowest level of analysis, that is the single questions in the survey, table 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

---

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.

- low-frequency terms
- vague quantification terms
- 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 condition with suboptimal question formulations with regard to all five text features in QUAID 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 what constitutes problematic questions.

--- Table 2 around here ---

3.5.2 Drop-out Rates

Drop-out rates were analyzed as a first indicator of data 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 suboptimal condition and 10.6% (n = 59) in the well-formulated condition. Drop-out rates did not differ between conditions, $\chi^2(1, N = 136) = 2.38$, $p = .12$.

3.5.3 Survey satisficing

Survey satisficing was analyzed as a second indicator of data quality. We examined four indicators of satisficing: very short response times, "no opinion" 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 cannot 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 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 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 of the respondents were in the condition with well-formulated questions and 19 respondents answered suboptimal questions. The direction of the effect is contrary to hypothesized satisficing behavior, showing more respondents with very short response times in the condition with well-formulated questions. This difference was not statistically significant, χ^2(df=1, N = 49) = 2.47, p = .12.

The propensity to give nonsubstantive answers was estimated by calculating item non-response rates[10] and by counting the number of neutral responses to the 12 question pairs offering a middle category. The item non-response rate was very low with only 125 (0. 45%) items being

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

left unanswered and there was no significant difference in item non-response between the two conditions, χ^2(df=1, N=985) = .404, p = .696. However, respondents gave more neutral responses to the 12 questions offering a middle category when answering suboptimal questions (1445 counts out of 5880 responses) than when answering well-formulated questions (1323 counts out of 5940 responses), χ^2(df=1, N=985) = .955, 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 suboptimal condition did not provide more answers in the acquiescent direction (3134 counts out of 5880 responses) than did respondents in the well-formulated condition (3212 counts out of 5940 responses), χ^2(df=1, N=985) = .528, 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. That way, another 11 questions were examined for primacy effects. Again, we found no primacy effect in the suboptimal condition (1305 counts out of 5390 responses) compared to the well-formulated condition (1390 counts out of 5445 responses), χ^2(df=1, N=985) = .737, p = .113.

4. Discussion and Conclusion

This study examined how seven psycholinguistic text features affect the cognitive burden and data quality of survey questions. Five text features are considered in the tool QUAID and two additional text features are proposed. Using response times as a measure of the cognitive effort required to answer a survey question, 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 data quality.

The present findings show a strong support for the relevance of 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 suboptimal questions had 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 working memory overload. In addition, survey designers should also optimize survey questions with regard to low frequency words, vague or imprecise relative terms, low syntactic redundancy, and 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, the effect size in some questions did not reach a significant level. For two questions (Q7, Q10) this might be due to the fact that the well-formulated questions contained three and four syllables more than the suboptimal ones. Question length may thus have suppressed the impact of question clarity on the response times. However, the relevance of vague or ambiguous noun-phrases was not confirmed. The words associated with this text feature are usually interpreted idiosyncratically by respondents and do thus not necessarily require more processing effort. In this respect, this text feature could be different from the other six text features in this study.

Data 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 suboptimal 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 clarity. Insofar as drop-out is mediated by respondent motivation, it is likely that our sample consisted of highly motivated respondents which 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, 2004). Moreover, respondents did not receive any incentives and thus agreed to answer the questions for no apparent reward.

Several indicators of data quality were assessed which concern survey satisficing behavior among respondents. Examining four indicators of satisficing (very short response times, "no opinion" responses, acquiescence, primacy effects), the suboptimal questions only resulted in more neutral responses (i.e., selecting the middle category). 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 high in motivation and cognitive ability, so that the

cognitive burden induced by the survey questions did not affect data quality. Instead, respondents were willing and able to cope with the higher demands of suboptimal questions while still providing longer response times.

There are two 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, 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 data quality.

Our findings have theoretical and practical implications. From a theoretical point of view, we found strong empirical evidence for effects of psycholinguistic text features on the cognitive burden of survey questions. Given that six 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 applied side, the specification of text features and their relation to question clarity can help practitioners to systematically check and improve the quality and comprehensibility of their questions. Manuals describing these text features in detail may supplement the existing "guidelines" or "standards" of asking survey questions and lend further precision to these rules.

References

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., 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* (3), 390-399.

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

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.). San Francisco, CA: Jossey-Bass.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 269-321). New York: Wiley.

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

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.

Couper, M. P., Tourangeau, R., & Conrad, F. G. (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). New York: Wiley.

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

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

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 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*. Thousand Oaks: Sage.

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

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

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.

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, March). *Data Quality of Paradata: A Comparison of Three Response Time Measures in a Randomized Online Experiment*. Poster session presented at the annual meeting of the General Online Research (GOR), Hamburg, Germany.

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. (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 N. Schwarz & S. Sudman (Eds.), *Answering questions* (pp. 259-291). San Francisco, CA: Jossey-Bass.

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.

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

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 R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 113-140). Hillsdale, NJ: Erlbaum.

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.

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.

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

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

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 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.

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.

Table 1. Analysis of response times per text feature

Text feature	LFRW	VIRT	VANP	CSYN	MEMO	LSYR	BINF
Between groups							
F(df1=1)	22.97	17.05	.184	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 4 questions each as repeated measures and reading rate as a covariate.

Table 2. Mean response times between conditions for each question: suboptimal questions (poor) vs. well-formulated questions (good)

Item	Means for raw data in sec.		Means for log-transformed data		F-value (df1=1)	df2	*p*
	Poor	Good	Poor	Good			
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 or 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 or ambiguous noun-phrases							
Q 09 pronoun with 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 syntactic structure	15.69	13.03	4.14	4.04	55.852	977	<.001*
Q 14 ambiguous syntactic structure	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 syntactic structure	14.87	15.05	4.11	4.10	.421	979	.517
Working memory overload							
Q 17 hypothetical question	20.94	19.81	4.25	4.20	9.672	979	.002
Q 18 quantitative 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 numerous 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 α=.05 (.05/28=.00179).