

ABSTRACT

Title of Dissertation: GRICEAN EFFECTS IN SELF-ADMINISTERED SURVEYS

Ting Yan, Doctor of Philosophy, 2005

Dissertation directed by: Professor Roger Tourangeau
Joint Program in Survey Methodology

Despite the best efforts of questionnaire designers, survey respondents don't always interpret questions as the question writers intended. Researchers have used Grice's conversational maxims to explain some of these discrepancies. This dissertation extends this work by reviewing studies on the use of Grice's maxims by survey respondents and describing six new experiments that looked for direct evidence that respondents apply Grice's maxims.

The strongest evidence for respondents' use of the maxims came from an experiment that varied the numerical labels on a rating scale; the mean shift in responses to the right side of the rating scale induced by negative numerical labels was robust across items and fonts. Process measures indicated that respondents applied the maxim of relation in interpreting the questions. Other evidence supported use of the maxim of quantity — as predicted, correlations between two highly similar items were lower when

they were asked together. Reversing the wording of one of the items didn't prevent respondents from applying the maxim of quantity. Evidence was weaker for the application of Grice's maxim of manner; respondents still seemed to use definitions (as was apparent from the reduced variation in their answers), even though the definitions were designed to be uninformative. That direct questions without filters induced significantly more responses on the upper end of the scale — presumably because of the presuppositions direct questions carried — supported respondents' application of the maxim of quality. There was little support for respondents' use of the maxim of relation from an experiment on the physical layout of survey questions; the three different layouts didn't influence how respondents perceived the relation among items.

These results provided some evidence that both survey “satisficers” and survey “optimizers” may draw automatic inferences based on Gricean maxims, but that only “optimizers” will carry out the more controlled processes requiring extra effort. Practical implications for survey practice include the need for continued attention to secondary features of survey questions in addition to traditional questionnaire development issues. Additional experiments that incorporate other techniques such as eye tracking or cognitive interviews may help to uncover other subtle mechanisms affecting survey responses.

GRICEAN EFFECTS IN SELF-ADMINSTERED SURVEYS

By

Ting Yan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor Roger Tourangeau, Chair
Professor Frederick Conrad
Professor Stanley Presser
Professor John Robinson
Professor Norbert Schwarz

©Copyright by
Ting Yan
2005

Acknowledgements

First, I would like to express my most sincere gratitude towards my academic advisor, Professor Roger Tourangeau, for his continuous support throughout the Ph. D program. I deeply appreciate his constructive advice and input, his time, and ongoing patience. The regular discussions with him have definitely benefited me and have greatly motivated me to move forward during the Ph. D program. During the frustrating period of dissertation-writing, he was always there to provide both academic guidance and emotional support (I am very sympathetic to him for the pain he had going through the first drafts of the dissertation). Without Dr. T., this dissertation wouldn't have been possible.

I would like to acknowledge the help of my committee in various stages of the dissertation process. I am grateful for their comments on and assistance with ideas, and personal help with data collection, data analysis, and writing. I am extremely thankful for their cooperation and accommodating efforts.

Collection of data for the grid experiment in Chapter 2 was supported by grants from the National Science Foundation (SES-0106222) and National Institute of Health (R01 HD041386-01A1) to Roger Tourangeau, Mick Couper, Fred Conrad, and Reg Baker. I would like to extend my gratitude to the PIs, who allowed me to piggyback my experiment onto their web study.

Special thanks go to Reg Baker, who not only made possible my last data collection effort but also made it happen in time for me to finish the dissertation.

Lastly, I would like to thank my family for their constant reminders (“When will you be done?”), and Mr. Low Ke Bin for not asking when I will be done. It was their full support and complete understanding that made this happen.

Table of Contents

Acknowledgements	ii
Lists of Tables.....	v
Lists of Figures	vi
Chapter 1 Prior Work on Gricean Influences in Surveys.....	1
1.1 Introduction.....	1
1.2 Theoretical Formulation and Development of CP	3
1.3 Gricean Effects in Survey Research	7
1.3.1 Maxim of Quantity.....	9
1.3.2 Maxim of Quality.....	14
1.3.3 Maxim of Relation	17
1.3.4 Web Surveys	24
1.3.5 Conclusions.....	27
1.4 Outline of Dissertation.....	28
Chapter 2 The Maxim of Relation and Question Presentation in Web Surveys. 29	29
2.1 Introduction.....	29
2.2 Method	33
2.3 Results.....	36
2.4 Conclusions.....	40
Chapter 3 The Maxim of Manner and Providing Definitions	43
3.1 Introduction.....	43
3.2 Method	50
3.3 Results.....	53
3.3 Conclusions.....	60
Chapter 4 Three Maxims	62
4.1 The Maxim of Relation and Numerical Values of Rating Scales	62
4.1.1 Introduction.....	62
4.1.2 Method	67
4.1.3 Results.....	69
4.1.4 Conclusions.....	75
4.2 The Maxim of Quantity and Similar Items	76
4.2.1 Introduction.....	76
4.2.2 Method	79
4.2.3 Results.....	81
4.2.4 Conclusions.....	84
4.3 The Maxim of Quality and Presuppositions	85
4.3.1 Introduction.....	85
4.3.2 Method	90
4.3.3 Results.....	92

4.3.4	Conclusions.....	96
Chapter 5	Conclusions and Discussion	98
5.1	Summary of Results.....	98
5.2	Discussion.....	105
5.2.1	Automatic versus Controlled Processing.....	106
5.2.2	Gricean Effects or Satisficing.....	109
Reference	112

Lists of Tables

Table 1.1.	Grice’s Four Conversational Maxims	5
Table 1.2.	Horn’s Neo-Gricean Principles vs. Grice’s Original Formulation.....	7
Table 1.3.	Summaries of Studies on Maxim of Quantity.....	10
Table 1.4.	Correlations Between Responses to General and Specific Questions, by Study and Condition	11
Table 1.5.	Studies Examining the Maxim of Quality.....	16
Table 1.6.	Research Work on the Maxim of Relation.....	18
Table 2.1.	Number of Completed Cases Per Experimental Condition.....	34
Table 2.2.	Experiment 1: Introductions, Target Items and Follow-up Questions	35
Table 3.1.	Number of Participants Assigned to Experimental Conditions	52
Table 3.2.	Variances of Responses to Four Target Questions By Definition Condition	54
Table 3.3.	Percentage of “Yes” Responses to Borderline Instances by Definition Condition.....	54
Table 3.4.	Regression Coefficients from Multiple Regression Models by Target Item	56
Table 3.5.	ANOVA Results on Total Response Time	56
Table 4.1.	Experimental Conditions of Studies on Numerical Values of Rating Scales	65
Table 4.2.	Experiment 4: Number of Completes Per Experimental Condition.....	68
Table 4.3.	Questions Used in Experiment 4 on Maxim of Relation	69
Table 4.4.	Experiment 4: Mean Ratings By Experimental Condition	70
Table 4.5.	Experiment 4: Two-way ANOVA Results	70
Table 4.6.	Experiments 4 and 5: Number of Completes per Experimental Condition ..	80
Table 4.7.	Items for Experiment 5.....	80
Table 4.8.	Experiment 4: Correlation Coefficients by Experimental Condition.....	82
Table 4.9.	Experiment 5: Correlation Coefficients by Experimental Conditions	83
Table 4.10.	Experiment 6: Questions in the Spending Block	90
Table 4.11.	Experiment 6: Number of Completes by Experimental Condition.....	91
Table 4.12.	Importance of Four Issues by Question Format.....	93
Table 4.13.	Mean Importance Ratings (and Percent Selecting Options Above Midpoint) by Whether Issue Included In Prior Block.....	95
Table 4.14.	Correlations Between Importance Items (Inference Questions) and Concern Items.....	96
Table 5.1.	Summary of Results	99

Lists of Figures

Figure 2.1.	Plots of Cronbach’s Alphas by Experimental Condition.....	37
Figure 2.2.	Mean Perceived Relatedness Rating by Experimental Condition.....	39
Figure 3.1.	Mean Standardized Response Times for HNC and LNC Respondents.....	58
Figure 3.2.	Mean Ratings of Whether Survey Terms Are Used in Technical Sense by Definition Condition.....	60
Figure 4.1.	Example of a Faded Scale	68
Figure 4.2.	Mean Shifts Due to Negative Numbers by Their Font Across 4 Items.....	70
Figure 4.3.	Percent Correctly Recalling Numerical End Point by Scale Condition	72
Figure 4.5.	Percentage of Respondents Inferring “Presence of Failure” by Scale Condition.....	74

Chapter 1 Prior Work on Gricean Influences in Surveys

1.1 Introduction

Survey errors are often classified into coverage, sampling, nonresponse, and measurement errors (Groves, 1989). Measurement errors are further subdivided by source – the interviewer, the respondent, or the instrument (Groves, 1989). The traditional mathematical model of survey error, as set forth in the work by Hansen, Hurwitz, and Madow (1953), and Cochran (1977), starts with the variance of the sample mean of a variable measured without error and based on a simple random sample, and then adds additional variance components representing measurement error (Hansen, Hurwitz, & Bershad, 1961). Implicit in the Hansen-Hurwitz-Bershad measurement error model is the emphasis on the interview process (the “essential survey conditions”) and on the “processors” of survey data (including supervisors, interviewers, coders, etc.) rather than on individual respondents. The traditional model assigns respondents a passive role and largely ignores them. Another weakness of this model lies in its focus on the consequences (rather than causes) of measurement error on survey estimates. The model is not informative as to how measurement errors arise or how to prevent or reduce them.

The CASM (Cognitive Aspects of Survey Methodology) movement in the 80s and 90s fostered a shift in survey methods research to the cognitive paradigm. The CASM movement focused on the cognitive processes by which respondents arrive at and report an answer using concepts drawn mostly from cognitive psychology (see Hippler, Schwarz, & Sudman, 1987; Jabine, Straf, Tanur, & Tourangeau, 1984; Sirken, Herrmann, Schechter, Schwarz, Tanur, & Tourangeau, 1999; Sudman, Bradburn, & Schwarz, 1996). This new paradigm centered on the *causes* of measurement errors and motivated practical

measures to reduce these errors (Tourangeau, 2003). However, many of the problems in surveys seem to involve respondents' comprehension of the questions. As a result, the work inspired by CASM did not provide a complete account of measurement error (even from a cognitive perspective) because most of this research concerned memory and judgment (Tourangeau, 1999; O'Muirheartaigh, 1999).

Conversational analysis represents an alternative approach to understanding the survey response process (see Schaeffer, 1991). Suchman and Jordan (1990), in an analysis of two survey interviews, describe the very different view that respondents may have of the survey interview from that of the researchers. Criticizing the standardized interviewing style, they consider the interview process as involving the joint participation of the interviewer and the respondent in a speech event – a communication – where each has expectations as well as responsibilities (Suchman & Jordan, 1990). Following up on this critique of standardized interviewing, Conrad and Schober (2000) and Schober and Conrad (1997) provided empirical evidence that the standardized interviewing style can reduce response accuracy in certain circumstances. Conversational (or flexible) interviewing, by contrast, improves individual response accuracy, but at the expense of increasing interview time and, thus, increasing cost.

There is an additional discipline – linguistics – that could provide additional tools for understanding comprehension problems. As a matter of fact, linguistics remains one of the disciplines that have been neglected by survey researchers, a point made by Tourangeau (1999). Fillmore (1999) pointed out problems in surveys that could have been avoided or solved through the application of linguistic knowledge and called for greater attention to linguistics from survey researchers. Schwarz and his colleagues had

already begun to use linguistics, applying one area of linguistic theory – Grice’s Cooperative Principle (CP) – to the survey research setting. Schwarz and his colleagues have construed survey research as a conversation between survey researchers and survey respondents (see Schwarz, 1996, for a thorough review; see Tourangeau, Rips, & Rasinski, 2000 for related issues). Such a conversational model of research is also advocated in another field by Kihlstrom (1995), who argued in favor of viewing experiments from the subject’s point of view.

Schwarz and colleagues demonstrated that many response effects are the result of respondents’ (sometimes incorrect) belief that survey researchers are cooperative and abide by Grice’s conversational maxims. I will review the work of Schwarz and his colleagues as well as the work of other survey researchers on respondents’ application of Grice’s maxims in the survey research setting.

I begin by reviewing the theoretical formulation and development of Grice’s CP before discussing the (mostly) experimental studies that examine CP in the survey research setting.

1.2 Theoretical Formulation and Development of CP

Sentences versus utterances . H. P. Grice first proposed the Cooperative Principle (CP) and the associated maxims in the 1967 William James Lectures at Harvard University. The CP lays out a mechanism through which people go beyond simple sentence meanings to derive the speaker’s intentions. It bridges “what is said” with “what is implied” through implicatures – inferences the listener makes regarding the speakers’ intentions (Grice, 1975).

Traditionally, linguistics has regarded sentences as a collection of abstract linguistic symbols. Sentence meanings are understood as “the overall meaning composed from the meanings of all the constituents together with the meaning of the constructions in which they occur” (Levinson, 1995, p. 91). Sentence meanings fall into the realm of semantics (or, in Fillmore’s categorization (1999), lexical semantics plus compositional semantics), which deals with the common core of meaning and the systematic process by which overall meaning is built out of the meaning of the parts (Levinson, 1995; Sperber & Wilson, 1995). Utterances are treated as pairings of sentences and contexts; the meaning of an utterance refers to the “import” of a sentence in a particular context (Levinson, 1995). Different utterances of the same sentence may differ in their import because of the context in which each utterance occurs. Within linguistics, the study of utterance meanings belongs to pragmatics.

The distinction between sentences and utterances and between sentence meanings and utterance meanings is by now familiar to language researchers. We all have the experience of speaking one thing to mean something else. For instance, when one person says to another: “It is getting dark,” the speaker could mean that the listener should turn on the light, or that the listener should leave, or even that the listener should stay for dinner. Any of these utterance meanings could be right or all of them could be wrong. The deciding factor is not the linguistic structure (i.e., the sentence), which remains constant across contexts, but rather the context in which the utterance occurs. Consequently, the question remains how the listener should interpret the speaker’s utterance in a given context.

Grice's Cooperative Principle and the conversational maxims. Grice (1989)

offered one answer to these questions in his lectures:

Our talk exchanges do not normally consist of a succession of disconnected remarks. ... They are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction. ... We might then formulate a rough general principle which participants will be expected (*ceteris paribus*) to observe, namely: Make your conversational contribution such as is required, at the stage where it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged (p. 26).

Grice labeled this the Cooperative Principle (CP). Table 1.1 summarizes the associated subordinate maxims.

Table 1.1. Grice's Four Conversational Maxims

Maxim of Quantity	1) Make your contribution as informative as is required (for current purposes); 2) Do not make your contribution more informative than is required.
Maxim of Quality	Try to make your contribution one that is true; 1) Do not say what you believe to be false; 2) Do not say that for which you lack adequate evidence.
Maxim of Relation	Be relevant.
Maxim of Manner	Be perspicuous, and specifically: 1) Avoid obscurity; 2) Avoid ambiguity; 3) Be brief; 4) Be orderly.

Grice asserts that people generally follow these rules for efficient communication. However, he didn't intend his account of the CP to be a prescriptive one. On the contrary, he admits that people do not follow these guidelines to the letter. Rather "...in most ordinary kinds of talk these principles are oriented to, such that when talk does not proceed according to specifications, hearers assume that, contrary to appearances, the principles are nevertheless adhered to at some deeper level" (Levinson, 1983, p. 102). In other words, Grice believes that the CP and the associated maxims are normally observed by the participants in a talk exchange either at the level of what is said (the semantic

meaning of sentences) or failing that, at the level of what is implicated (the utterance meanings or implicatures).

Theoretical development. Grice's maxims lay out a research direction more than a complete, well-specified framework. He acknowledges that his list of maxims is far from exhaustive and that there are all sorts of other maxims (aesthetic, social, or moral in character) that are normally observed by participants in talk exchanges (Grice 1989, p. 28). One line of development in pragmatics since Grice adds new maxims to supplement and extend Grice's original maxims. Leech and his Politeness Principle (1983) and Brown and Levinson's Politeness Principle (1987) are representatives of this approach.

This expansionism has been criticized as a seemingly endless addition of principles/maxims whenever situations call for it. Led by Sperber and Wilson (1986, 1995), reductionist critics replaces maxims with just a single Principle of Relevance (later renamed by them the Communicative Principle of Relevance in the 1995 edition): "the speaker tries to be as relevant as possible in the circumstances" (1986, p. 381). But as Davis (1998) pointed out, Sperber and Wilson's principle has the same difficulty as Grice's CP. Both are intended to be general enough to hold in all cases, but end up being too general to yield specific predictions. It is difficult to summarize Sperber and Wilson's theory because the formulation of the theory varies significantly from presentation to presentation (Davis, 1998, p. 99).

The Neo-Gricean school is the third line of development after Grice's initial work (Horn, 1984; Horn & Ward, 2004). It emphasizes the quantity maxim as the core of

Grice's system and collapses Grice's maxims into two principles—Q(uality) and R(elation). Table 1.2 summarizes Horn's Q and R principles (Horn, 1984).

Table 1.2. Horn's Neo-Gricean Principles vs. Grice's Original Formulation

	Horn's Principles	Encompassing Grice's maxims
The Q Principle (hearer-based)	Make your contribution sufficient (cf. Quantity ₁). Say as much as you can (given R).	Quantity ₁ : Make your contribution as informative as is required (for the given purpose of the exchange).
The R Principle (speaker-based)	Make your contribution necessary (cf. Relation, Quantity ₂ , and Manner). Say no more than you must (given Q).	Quantity ₂ : Do not make your contribution more informative than is required. Maxim of Manner Maxim of Relation

Note: Adapted from Horn (1984, p. 13).

Despite these various criticisms, Grice's CP and its maxims continue to be regarded as the main model of the tacit agreement between participants in natural conversational settings. Sperber and Wilson's relevance theory and Horn's Neo-Gricean principles are too abstract to permit the prediction of potential inferences drawn by participants; therefore, they are not directly testable. As a result, most survey researchers have focused on Grice's maxims (cf. Schwarz, 1996). Strack (1994) is the only researcher who has attempted to examine the politeness principle in survey research setting. In the next section, I will review experimental work that explores the relevance of Grice's maxims to the survey research setting and show examples of measurement error resulting from respondents' application of the CP in the context of survey research.

1.3 Gricean Effects in Survey Research

Survey interviews are similar to everyday conversations for two reasons. First of all, in both settings, at least two parties are involved – survey researchers (sometimes via interviewers) and survey respondents (see Kihlstrom, 1995; Schwarz, 1996). Secondly,

like natural conversational exchanges, survey interviews involve speech acts; in the case of surveys, a crucial speech act is the request for information from respondents as specified by the question's meaning. On the other hand, as noted quite often, survey interviews are also quite different from daily conversation (Suchman & Jordan, 1990). The most significant difference between the two settings is that survey interviews – especially standardized interviews – are highly constrained (Clark & Schober, 1992; Schwarz, 1996). Whereas speakers and addressees collaborate in ordinary conversations “to establish intended word meanings, intended interpretations of full utterances, implications of utterances, mutually recognized purposes, and many other such things” (Clark & Schober, 1992, p. 25), their opportunity to do so is severely limited in survey interviews, partially due to survey researchers' attempt to standardize the interview process (Schwarz, 1996, 1999).

Because of these differences between ordinary conversation and survey interviews, Schwarz speculates that respondents not only bring Grice's maxims to the survey interviewing situations but also are forced to rely more on these tacit assumptions than they would be in daily life to derive the intended meanings of the survey questions (Schwarz, 1996). Unlike participants in ordinary conversations, interviewees are not permitted to clarify their utterances. Respondents assume (correctly or incorrectly) that every contribution of the survey researcher is relevant to the aims of the ongoing conversation, that every contribution is informative, truthful, relevant, and clear (Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). Although the survey designers may have only the semantic meaning of a question in mind when asking the question, respondents still assume that the question obeys Grice's maxims (Schwarz, 1999). That is,

respondents may read between the lines, finding meaning in incidental features of the question and “rendering logically irrelevant information conversationally relevant” (Schwarz, 2000, p. 152). Accordingly, the question that respondents attempt to answer might be different from the one survey designers intend to ask. Such a mismatch or gap in understanding influences respondents’ responses to questions and contributes to measurement error in surveys. Numerous demonstrations of such measurement errors can be found in the survey literature. Each can be attributed to the application of Grice’s maxims by survey respondents.

1.3.1 Maxim of Quantity

The maxim of quantity requires speakers to make their contribution informative, but no more informative than is required. A number of studies illustrate the application of the maxims of quantity in a survey setting. Table 1.3 summarizes them.

There are two main situations in which respondents seem to apply the maxim of quantity to arrive at inferences about the intended meanings of survey questions. One situation involves part-whole questions. Respondents are asked one question that calls for an evaluation of some general domain or category (e.g., their overall happiness), and another question that calls for an evaluation of a highly salient member of that domain (e.g., marital happiness). Typically, when the specific question precedes the more general question, it alters answers to the general question (Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). One explanation for this finding goes as follows: when two questions are in the same conversation context and seem to ask for the same

Table 1.3. Summaries of Studies on Maxim of Quantity

Study	Questions	Results	Inference/Explanation
Strack, Schwarz, and Wanke (1991, Experiment 2)	Similar items – “How happy are you...?” and “How satisfied are you...?”	Lower correlations when the items are in the same conversational context than when they are in different conversational contexts	When in same conversational context, respondents infer difference in meanings between two highly similar items to avoid redundancy.
Kalton, Collins, and Brook (1978)	Specific-General – “Drivers in general,” and “Young drivers”	Higher percentage of people say driving standards are lower when the general question is asked first than when the general question is asked second	When a specific question is asked before general one, information used to answer the specific question is excluded from answering the subsequent general question because respondents infer general question calls for something new (avoiding redundancy).
Mason, Carlson, and Tourangeau (1994)	Specific-General – “Local economy,” and “State economy”	More optimism about state economy when it is asked first. Lower optimism when it follows specific question.	
McCabe and Brannon (2004)	Specific-General – “Romantic satisfaction,” and “Overall satisfaction”	Attenuated correlations between the two question when presented with a joint lead-in only for respondents who have a high need for cognition.	
Schwarz, Strack, and Mai (1991)	Specific-General – “Romantic satisfaction,” and “Overall satisfaction”	High correlations when general question first or with explicit inclusion wording; low correlations when specific question first, joint lead-in, or with explicit exclusion wording	
Strack, Martin, and Schwarz (1988)	Specific-General – “Happiness with dating”/ “Frequency of dating” and “Happiness with life in general”		
Tourangeau, Rasinski, and Bradburn (1991)	Specific-General – “Happiness with marriage,” and “Overall happiness”		

information, respondents reinterpret the general question as excluding the specific mentioned in the prior question, taking the general item to mean something like “aside from...” (Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). These speculations have been supported by a several similar studies (Kalton, Collins, & Brook, 1978; Mason, Carlson, & Tourangeau, 1994; McCabe & Brannon, 2004; Schwarz, Strack, & Mai, 1991; Strack, Martin, & Schwarz, 1988; Tourangeau, Rasinski, & Bradburn, 1991).

Schwarz et al. (1991), Strack et al. (1988), and Tourangeau et al. (1991) asked respondents about their happiness with marriage/romance/dating (the specific question) and about general happiness (the general question). All three studies show that the correlations between the general and specific items are lower when the general question comes second (indicating that respondents exclude their marriages/dating when they evaluate their overall happiness under this order), when the question explicitly calls for their exclusion, or when the two questions are perceived to be in same conversational context. The findings are displayed in Table 1.4 (adapted from Table 7.2 in Tourangeau, Rips, & Rasinski, 2000, p. 205).

Table 1.4. Correlations Between Responses to General and Specific Questions, by Study and Condition

Condition	Strack et al. (1988)		Tourangeau et al. (1991)	Schwarz et al. (1991)	
	Experiment 1	Experiment 2		One Specific Question	Three Specific Questions
General-specific	.16 (60)	-.12 (60)	.54 (60)	--	.32 (50)
Specific-general No introduction	.55 (60)	.66 (60)	--	.67 (50)	.46 (50)
Specific-general Joint lead-in	.26 (60)	.15 (60)	.28 (53)	.18 (56)	.48 (56)
Specific-general Exclusion wording	--	--	.27 (54)	.20 (50)	.11 (50)
Specific-general Inclusion wording	--	--	.52 (59)	.61 (50)	.53 (50)

Note: Parenthetical entries are sample sizes.

A third study, by Kalton, Collins, and Brook (1978), looked into another topic—driving standards for drivers in general and for young drivers (see Table 1.3). Not surprisingly, they found a lower percentage of people saying driving standards were getting lower for drivers in general when the general question came after a question about young drivers. Again, when the general question followed the specific question, the general driving standards seemed to be reinterpreted as meaning ‘drivers aside from young drivers.’

In addition to being the basis for respondents’ inference that the general question is meant to exclude material and/or information used in answering an earlier, more specific question, Grice’s maxim of quantity may also cause respondents to disregard accessible information simply because they feel that the task or question calls for something new and different (Tourangeau, Rips, & Rasinski, 2000). Mason, Carlson, and Tourangeau (1994) provide an example. They asked respondents to report their expectations about the prospects for both their state and local economy and to explain why they felt that way. When the local economy question was asked first, the reasons pertaining to the local economy (e.g., high unemployment) were not given again for the state economy, apparently because respondents felt their answers to the second question should be based on something new (Mason et al., 1994). As a result, the distribution of reasons cited for the general item was altered, when that item followed the more specific question.

McCabe and Brannon (2004) reported a partial replication of Schwarz, Strack, and Mai (1991). However, this study differs from the previous studies on part-whole questions in that it linked respondents’ application of the maxim of quantity with their

need for cognition. McCabe and Brannon found that respondents with a high need for cognition displayed an attenuated correlation between satisfaction ratings when the two items had a joint lead-in, but not respondents with a low need for cognition. This finding seemed to support the notion that the conversational norm to avoid redundancy is not automatically applied in the survey context and that the observed reduction in correlations is not the result of respondents' 'satisficing' (i.e., their tendency to reduce cognitive burden by taking various shortcuts). Rather, those with a high need for cognition – who are thought to pay more attention to questions and to process survey questions more carefully – are affected by the manipulation of the order of the two questions.

The other situation that seems to reflect the application of the maxim of quantity involves two highly similar items in the same questionnaire. For instance, Strack, Schwarz, and Wanke (1991, Experiment 2) asked respondents to rate both their happiness ("How happy are you...?") and their satisfaction with life ("How satisfied are you...?") in a single questionnaire. The two items are quite similar and appear to invite redundant answers from respondents. However, respondents appeared to infer that the two similar items must mean something different when they were presented in the same part of the questionnaire; according to Strack and his coauthors, respondents exaggerated the small difference in meaning between happiness and satisfaction so that the questions no longer violated the Cooperative Principle. The findings confirmed this account – correlations between the two ratings were much smaller when the two items were administered in a single questionnaire than in different ones (Strack, Schwarz, & Wanke, 1991).

1.3.2 Maxim of Quality

Respondents sometimes offer opinions about highly obscure or even fictitious issues. Indeed, Converse (1970) argued they often provide random answers to survey questions based on a mental flip of a coin. Grice's CP and the maxim of quality offer an alternative explanation for why respondents provide answers to survey questions.

Respondents answer questions on fictitious or highly obscure issues because they assume that the survey designers are being cooperative. The sheer fact that a researcher asks a question about some issue presupposes that the issue exists (or else asking a question about it would be uncooperative). As a result, the respondents search the context for cues and attempt to make sense of the obscure or fictitious issue.

An experiment from Strack, Schwarz, and Wanke (1991, Experiment 1) demonstrates this process. German respondents were asked about their support for an undefined 'educational contribution.' Based on the presupposition of the question – that something called an 'educational contribution' exists – respondents tried to assign a reasonable meaning to the term. When the question followed another question asking for their estimate of the amount of tuition students have to pay at U.S. universities, they inferred that 'educational contribution' was to be taken from them. But when the same question was preceded by a question asking them to estimate the amount of money the Swedish government pays students, they inferred that the 'educational contribution' was to be given to them. They then answered the question according to their interpretation of the educational contribution; they showed more support for it when the contribution was to be given to them than to be taken from them (Strack, Schwarz, & Wanke, 1991). In this case, an inference based on the maxim of relation resolves an apparent violation of

the maxim of quality. Another study by Tourangeau and Rasinski (1988), which will be discussed more in detail later, exhibited a similar use of the maxim of relation.

In the same way, questions such as “How many times did you go shopping in the past month?” or “How concerned are you with pollution?” presuppose that one went shopping or one should be concerned about pollution, which may not be an appropriate assumption. Nonetheless, assuming that these presuppositions are true, respondents may reinterpret the question. For example, they may reinterpret shopping broadly as encompassing, say, trips to the drug store or grocery. However, a prior “filter” item (e.g., “Did you go shopping in the past month?” or “Are you concerned with pollution?”) cancels such presuppositions, explicitly legitimizing the “did not go shopping” and “not concerned” responses.

Knauper (1988) reports a study that seems to illustrate these interpretive processes. She found that respondents reported more crimes (and less severe ones) when they were asked “In the past 10 years, how many times did you witness a crime?” directly, but reported fewer and more severe crimes when a filter “Did you witness a crime in the past 10 years?” was used first. Similar findings were reported by Sterngold, Warland, and Herrmann (1994), who argued that direct questions about respondents’ level of concern encouraged respondents to overstate their concerns. The use of filter questions greatly reduced respondents’ reported concern and produced fewer responses in the upper part of the scale. Table 1.5 summarizes the studies demonstrating inferences based on the maxim of quality in the context of survey questionnaires.

Table 1.5. Studies Examining the Maxim of Quality

Study	Target Questions	Result	Inference/Explanation
Strack, Schwarz, and Wanke (1991, Experiment 1)	Fictitious issue – “educational contribution”	More support for the target item when it followed an item about government financial support for students in Sweden than one about tuition in US.	Respondents assigned meaning to the fictitious issue using contextual information (the preceding question).
Tourangeau and Rasinski (1988)	Obscure term – “Monetary Control Bill” (MCB)	More support for MCB when the item followed block of inflation questions than when inflation questions are scattered among questions on unrelated issues.	Respondents inferred topic of Monetary Control Bill must be inflation when the item followed a block of inflation questions.
Knauper (1998)	With filter vs. without filter – “How many times did you experience X?”	Significantly fewer reports when first asked filtered question than without filter question.	Question ‘How many times did you experience X?’ presupposes one has experienced X; so respondents inferred X included less serious, more frequent incidents. The filtered question, on the other hand, canceled the presupposition and respondents reasoned survey designers meant more serious, rarer incidents.
Sterngold, Warland, and Herrmann (1994)	With filter vs. without filter – “How concerned are you about Y?”	A higher percentage of ‘not concerned’ responses and fewer responses at the upper end of the response scale with filtered question than when items without filters asked	Question “How concerned are you with...” implies one should be concerned about it, encouraging respondent to overstate their concerns. A filter reduces such an implication by legitimizing ‘not concerned’ responses.

1.3.3 Maxim of Relation

This maxim is probably the most powerful and general of Grice's maxims, which justifies efforts (like Sperber and Wilson's) to reduce the other maxims to this one. In the survey research context, the maxim of relation implies that everything about the questions is relevant to the question-answer process; they make use of every possible piece of information (such as nearby questions, the numerical values assigned to response scales, the range, and even the physical layout of the scales) to interpret the question and make the required judgment. Relevant research is summarized in Table 1.6.

For instance, when a question is ambiguous or involves an obscure or unfamiliar issue, respondents try to infer its meaning from the questions around it. Such inferences are based on the maxim of relation. A study by Tourangeau and Rasinski (1988) demonstrates the operation of the maxim of relation. When the target question about the obscure "Monetary Control Bill" (MCB) was placed in a block of questions on inflation, respondents seemed to infer that the Monetary Control Bill must concern inflation (see also Schuman & Presser, 1981, pp. 154-160). By contrast, when questions shifted from one topic to the next in the questionnaire, respondents no longer assumed the previous questions were relevant and were more likely to give a "No Opinion" response to the item about the MCB (Tourangeau & Rasinski, 1988). A similar finding was observed in the Strack, Schwarz, and Wanke's study (1991), where respondents tried to infer the meaning of "educational contribution" by assuming the questions preceding the target item were related to it.

Table 1.6. Research Work on the Maxim of Relation

Study	Target Questions	Results	Inference/Explanations
Strack, Schwarz and Wanke (1991, Experiment 1)	Fictitious issue – “Educational contribution”	More support for the target item when it followed an item about government financial support for students in Sweden than one about tuition in the US.	With the aid of relation maxim, respondents assigned meaning to the fictitious issue using contextual information (preceding question).
Tourangeau and Rasinski (1988)	Obscure term – “Monetary Control Bill” (MCB)	More support for the MCB when the item followed a block of inflation questions than when the inflation questions were scattered among questions on unrelated issues.	With the aid of relation maxim, respondents inferred topic of Monetary Control Bill must be inflation when the item followed block of inflation questions.
Haddock and Carrick (1999, Experiment 1)	Rating scale – “Ratings of Blair on four attributes” -5 to 5 vs. 0 to 10	Blair was rated more favorably (mean=6.9) in the -5 to 5 condition as compared to 0 to 10 condition (6.4). Blair was judged to be more effective when respondents were asked to rate him on four attributes on the -5 to 5 scale than on the 0 to 10 scale.	Numerical labels served as information that affected respondents’ ratings of Blair on four attributes. Furthermore, the trait ratings on the different numerical scales influenced a subsequent judgment about Blair’s predicted effectiveness as Prime Minister.
O’Muirheartaigh, Gaskell, and Wright (1995, Experiment 1)	Rating scale – “Opinion on adverts on TV” -5 to 5 vs. 0 to 10	More respondents selected value equal or less than the midpoint given the 0 to 10 scale than when given the -5 to 5 scale.	“Much less entertaining than the programmes” implies presence of negative evaluations when combined with -5, but absence of positive evaluations when combined with 0. Confirmed Schwarz et al (1991)’s finding.
Schwarz, Grayson, and Knäuper (1998, Experiment 1)	Frequency scales – 0 to 10 vs. 1 to 11 (‘rarely’ to ‘often’)	Higher mean frequency ratings along the 0 to 10 scale (mean=3.8) than the 1 to 11 scale (2.9).	Respondents interpreted ‘rarely’ to mean ‘never’ when combined with 0, but to mean ‘a low frequency’ when combined with 1.
Schwarz and Hippler (1995)	Rating scale – “Opinion toward politicians” -5 to 5 vs. 0 to 10	More positive rating of politicians along -5 to 5 scale (mean=5.6) than along 0 to 10 scale (4.9)	Respondents interpreted verbal label ‘don’t think very highly’ to indicate absence of positive thoughts when combined with 0, but the presence of negative thoughts when combined with -5.
Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991, Experiment1)	Rating scale – “Success in life” -5 to 5 vs. 0 to 10	Fewer people endorsed a value equal or less than the midpoint on the -5 to 5 scale than on scale from 0 to 10.	-5 implies ‘presence of failure’ for end label ‘Not at all successful,’ and 0 implies ‘absence of success’ for same end label.

Study	Target Questions	Results	Inference/Explanations
Schwarz, Grayson, and Knäuper (1998, Experiment 2)	Different shapes of a scale: a set of stacked boxes; onion format; pyramid format	Students rated their academic performance less favorably with pyramid scale than with stacked boxes and with onion format.	The wider bottom in the pyramid scale implies researchers believe distribution includes more people at the bottom.
Gaskell, O’Muircheartaigh and Wright (1994)	Frequency scale – “Annoyance,” “Feeling unsafe,” and “Being in pain” Low frequency vs. high frequency	Fewer reports with low frequency scale; more reports with high frequency scale	Low frequency scale implies lower rate of occurrence in population
Schwarz, Strack, Muller, and Chassein (1988)	Frequency scale – “Irritation” Low frequency vs. high frequency	Students either provided a more serious irritation or rated the standard example as more annoying given low frequency scale.	Low frequency scale implies that “irritation” mean “serious irritations,” and high frequency scale implies it means “minor irritations”
Schwarz and Bienias (1990)	Frequency scale – “Weekly TV consumption and alcohol consumption” Low frequency vs. high frequency	Lower weekly TV watching and alcohol consumption; more pronounced effect with proxy reports than with self-reports	Respondents use the range of the frequency scales as a frame-of-reference to estimate their own or others’ behavior, or to make comparative judgment. Respondents make the inference that the scale reflects the distribution of the behavior, checking a response alternative is equivalent to locating one’s own position in the distribution.
Schwarz, Bless, Bohner, Harlacher, and Kellenbenz (1991, Experimental 2)	Frequency scale – “Seriousness of illness” (Vignettes) Low frequency vs. high frequency	Suffering from a given symptom rated as more severe and more likely to require consultation when presented on a low frequency scale.	
Schwarz, Hippler, Deutsch, and Strack (1985)	Frequency scale – “TV consumption;” “Satisfaction with leisure time activities” Low frequency vs. high frequency	Less TV watching reported with low frequency scale; also lower satisfaction with leisure time activities; and higher ratings of importance of TV	
Schwarz and Scheuring (1988)	Frequency scale – “Masturbation;” “Satisfaction with intimate relationship;” “Interest in extramarital affairs” Low frequency vs. high frequency	Lower frequency of masturbation/intercourse; lower ratings of satisfaction with intimate relationship and higher ratings of interest in extramarital affairs	

Study	Target Questions	Results	Inference/Explanations
Ji, Schwarz, and Nisbett (2001)	Frequency scale – “Mundane behaviors for self and for others” Low frequency vs. high frequency (cross-cultural)	For western respondents, replicated Schwarz et al. (1985); but not for eastern respondents	Respondents use the range of the frequency scales ‘frame-of-reference’ to estimate their own and others’ behavior, and to make comparative judgment. Respondents make the inference that the scale reflects the distribution of the behavior, checking a response alternative is equivalent to locating one’s own position in the distribution.
Norenzayan and Schwarz (1999)	Researchers’ affiliation on causal attributions	Provided more situational attributions when researcher identified as a social scientist rather than a personality psychologist	Respondents infer researchers are either social scientist or personality psychologist from the affiliation or the letterhead, and attempt to make their answers relevant to the goals of researcher.
Wanke (2002)	Reference group – “students” vs. “population at large” Reference behavior – “Leisure activities” vs. “cultural activities”	Lower frequency report of going to movies when reference group is students; Lower frequency report of going to movies when survey introduced as assessing leisure activities	Respondents make use of the reference persons and reference behaviors given in the survey as a base for comparison
Winkielman, Knauper, and Schwarz (1998)	Reference period – “Anger” during “last year” vs. “last week”	More frequent and less severe episodes of anger reported when question pertained to one week than one year.	Shorter inference period implies survey researcher interested in more frequent and less severe episodes of anger

Respondents may also rely on the maxim of relation in drawing inferences from the numerical values assigned to a rating scale. Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991) reported the first demonstration of this type of inference, showing that respondents came to different interpretations of the verbal end points of a scale when the scale labels ran from 0 to 10 than when it ran from -5 to 5. Presuming that every piece of information was relevant, respondents inferred that the same end label (“Not at all successful”) meant the “mere absence of noteworthy success” when 0 was assigned to that scale point, but “the presence of failure” when -5 was assigned to the scale point (Schwarz, Knauper, et al., 1991). As a result, respondents were less likely to select values less than or equal to the midpoint with the -5 to 5 scale labels than with the 0 to 10 labels. This finding is replicated on questions about different topics in four other studies (Haddock & Carrick, 1999; O’Muirheartaigh, Gaskell, & Wright, 1995; Schwarz, Grayson, & Knäuper, 1998; Schwarz & Hippler, 1995). The experiment reported by O’Muirheartaigh et al. was part of a large-scale face-to-face survey. The experiment reported by Schwarz, Grayson, and Knäuper involved a unipolar frequency scale and contrasted 0 to 10 and 1 to 11 scales.

The visual appearance of a rating scale can influence respondents’ interpretation of the scale as well. Smith (1995), in a non-experimental study, noted that respondents seemed to draw inferences from the shape of the rating scale when asked to place themselves in the social hierarchy. When the scale was presented in a shape of pyramid with the bottom wider than the middle and the top, respondents inferred that the scale conveyed the distribution by class that the researchers had in mind and that more people should be included at the bottom than in the middle (Smith, 1995). This same result was

observed in an experiment by Schwarz, Grayson, and Knauper (1998, Experiment 2) among US college students, who were asked to evaluate their academic performance on either a stacked-box scale, a pyramid scale, or an onion-like scale (with the widest categories in the middle). The different physical displays of the scales affected the response distributions in the predicted way (Schwarz, Grayson, & Knauper, 1998).

Frequency scales provide other examples in which respondents apply the maximum of relation to derive pragmatic implicatures about the questions (Schwarz & Hippler, 1991). Respondents assume that the scale the researchers present in the questionnaire is meaningful and reflects the researchers' knowledge about the population distribution of the behavior in question (Schwarz, 1996; Schwarz, Hippler, Deutsch, & Strack, 1985). Values in the middle range of the scale are assumed to reflect the "average" or "typical" behavior, whereas the extremes of the scale are assumed to correspond to the extremes of the distribution (Schwarz, 1996).

Respondents may interpret ambiguous questions based on this assumption. For instance, when asked to indicate how frequently they were "really irritated" recently, respondents inferred that 'irritation' meant minor annoyances when they were given a high frequency response scale, but severe annoyances when presented a low frequency scale (Schwarz, Strack, Muller, & Chassein 1988). Another study by Gaskell, O'Muircheartaigh, and Wright (1994) made the same point – respondents rely on the frequency range of the response alternatives to comprehend the meaning of vague events rather than relying solely on the wording of the question.

The range of response alternatives may have a strong impact on the estimated frequency of one's own and others' behaviors as well. As Schwarz, Hippler, Deutsch,

and Strack argue (1985), everyday behaviors are not always represented in memory as distinct episodes, but rather as a blended, generic representation. Accordingly, respondents often answer questions about the overall numbers of a given behavior (e.g., visits to an ATM) by estimating the frequency rather than recalling and counting each episode (Sudman, Bradburn, & Schwarz, 1996). Applying the relation maxim in this context, respondents consider the range of the response options relevant to the question asked and use it as a frame of reference for their estimate (Schwarz, 1996; Schwarz, Hippler, et al., 1985). The study by Schwarz and colleagues, for example, observed higher frequency estimates with scales that presented higher rather than lower frequency response alternatives (Schwarz and Bienias, 1990; Schwarz, Hippler, et al., 1985; Schwarz & Scheuring, 1988; see also Tourangeau & Smith 1996). The same pattern was also observed in a cross-cultural setting (Ji, Schwarz, & Nisbett, 2001). Both American and Chinese respondents made use of the range of response alternatives to estimate the frequency of unobservable behaviors (such as telling a lie). Chinese respondents relied less on the response alternatives to estimate observable behaviors (Ji, Schwarz, & Nisbett, 2001). This finding is important in that it demonstrated the pervasiveness of the CP and maxims across cultures.

In addition to influencing respondents' behavioral reports, the range of the response alternatives can also affect subsequent comparative judgments (Schwarz, Hippler, et al., 1985; Schwarz & Scheuring, 1988; Schwarz, Bless, et al., 1991). Because the scale values are assumed to reflect the distribution of the behavior in question, the selection of a response is equivalent to locating oneself in the distribution. Respondents subsequently use these inferences about their relative positions within the population in

making later judgments (Schwarz, 1996). For instance, respondents in the Schwarz, Hippler, et al. (1985) study reported that television played a more important role in their leisure time when they had chosen an answer from the low rather than the high frequency scale, since this scale makes it look as if they were above average in their TV viewing. This effect of the scale range on comparative judgments has been replicated across other behaviors and judgments.

Other features of the questions, such as the length of the period covered, the reference group, or reference behaviors given in the question text – even the letterhead on which the questionnaire is printed with the researchers’ affiliations – can trigger the maxim of relation, affecting the inferences respondents draw about the researchers’ intended meanings (Norenzayan & Schwarz, 1999; Wanke, 2002; Winkielman, Knauper, & Schwarz, 1998). For instance, a question that covers a long reference period may imply that the events must be memorable. The impact of the relation maxim also is apparent in some psychological experiments concerning base-rate information utilization and attribution errors (Krosnick, Li, & Lehman, 1990; Schwarz, Strack, Hilton, & Naderer, 1991).

1.3.4 Web Surveys

The studies reviewed so far were done either in self-administered or face-to-face surveys. With developments in computer technology, newer data collection methods such as web surveys are gaining popularity among survey researchers. Compared to the traditional modes for conducting surveys, web surveys can impart information through a richer range of media; picture images, hyperlinks, and video clips can be easily

embedded in a web survey, even though it is not yet clear whether these rich media are more likely to help or to confuse respondents.

Couper, Tourangeau, and Kenyon (2004) make a distinction between task and style elements in web surveys (2004, p. 256-257). Task elements are those that are essential to the task of completing the survey and include the question wording, the response options, instructions, navigational cues, and other essential material. Typically, the task elements are verbal, though they can also be visual. By contrast, the style elements refer to those features of the instrument that are incidental or completely unrelated to the task of answering the question (at least from the viewpoint of the survey designers). They include the elements that create the overall appearance of the web site or survey instrument, the typeface used to represent the survey questions, the background color, and so on. Style elements are typically but not necessarily visual.

Some elements intended only as style elements can affect survey responses to web questions. For example, Couper, Traugott, and Lamias (2001) reported a web survey of University of Michigan students, in which the width of a numeric entry box was inadvertently varied. A random subset of respondents received a longer box than was necessary for the task. They found that respondents were guided by the size of the entry box and provided more information than was required (Couper, Traugott, & Lamias, 2001). Respondents gave longer answers such as “between 4 and 5” to the longer entry box instead of providing a numeric value (e.g., 5). The length of the box had a significant effect on the content of answers provided (Couper, Traugott, & Lamias, 2001).

In another web survey, Couper, Tourangeau, and Kenyon (2004) examined the effect of presenting pictures on responses to behavioral frequency questions. They

embedded in the survey 1) a picture of a salient but low frequency instance of the behavior in question (e.g., a picture of a department store in a question about shopping), 2) a picture of a less salient but higher frequency instance (a picture of a grocery store) , 3) both pictures, or 4) neither picture. Relying on the maxim of relation, respondents used the picture to interpret the type of behaviors covered in the questions. Accordingly, the picture showing the high frequency instance of the categories prompted higher reporting on the average than the picture showing the low frequency instance (Couper, Tourangeau, & Kenyon, 2004).

The two studies by Couper and his colleagues seem to demonstrate that web surveys are not immune to the Gricean effects observed in the more traditional modes of administration. Respondents appear to apply Grice's maxims whether they are conducting a daily conversation or doing a survey, and whether the survey is online or offline.

Summary. To sum up, many incidental features of the questionnaire (even survey research setting itself) can affect respondents in unanticipated ways; seemingly incidental features can produce implicatures that affect respondents' interpretation of the questions, and, subsequently, their responses. The overlap in meaning between items, the items' sequencing, the numerical range of the response alternatives, and the range and labeling of ratings scales can all produce implicatures about the survey designers' intent in asking the questions, and these in turn shape the way respondents think about and answer the questions (Tourangeau, Rips, & Rasinski, 2000).

1.3.5 Conclusions

Schwarz's insight that Grice's CP and his maxims apply in survey interviews is ground-breaking in that it points to an additional source of measurement error in surveys. Survey researchers generally intend for their questions to be taken literally. Features other than the question text are thought to be peripheral and used out of convenience. They are, in the terms of Couper and his colleagues, style elements (Couper, Tourangeau, & Kenyon, 2004). Survey respondents, however, being cooperative communicators themselves, try their best to understand and to answer the questions. They often seem to read between lines and to find meaning in every feature of the question. As a result, the question respondents attempt to answer might turn out to be different from the one the researchers intended to ask. This mismatch or gap in understanding that influences respondents' responses to questions and leads to measurement error.

Schwarz (1998, 2000) presents three alternatives for surveys to take these Gricean effects into account. We can 1) reliably create many biases by flouting conversational norms; 2) reliably attenuate such biases by undermining the assumption that the survey designers are cooperative communicators; or 3) avoid such biases by being cooperative communicators in the first place.

Obviously the last two alternatives are the only way to reduce or eliminate response biases. However, as Tourangeau, Rips, and Rasinski (2000) point out, there are many properties associated with each survey item and respondents cannot attend to (and misinterpret) all of them. So the question is how we can predict which elements of a question will produce a Gricean effect. At this point, we don't know enough about the mechanisms underlying these effects and don't have a good basis for prediction. We

need systematic studies that provide direct evidence of the operation of the Cooperative Principle and the associated maxims in the survey research setting.

1.4 Outline of Dissertation

This dissertation extends research on Gricean effects in surveys, with a focus on obtaining direct evidence of respondents' inferences based on Grice's maxims. The remainder of the dissertation consists of four chapters. Chapter 2 looks into the maxim of relation and the effects of physical arrangement of survey questions on web pages. Providing self-evident definitions to everyday terms in a survey questionnaire would seem to violate the maxim of manner; Chapter 3 describes a study examining the consequences of including such definitions. Chapter 4 describes four experiments embedded in one web study; these experiments examine the maxims of relation, quantity, and quality, and their effects on measurement errors arising from web surveys. The final chapter of the dissertation summarizes the findings, discusses their implications and the limitations of the research.

Chapter 2 The Maxim of Relation and Question Presentation in Web Surveys

2.1 Introduction

Prior studies have established that many incidental features of survey questions can affect respondents' answers and create response errors (Schwarz, 1996, 1999). The present study explores a formal feature of web survey questions – the physical arrangement of the questions on the web page – and its effects on respondents' inferences about the meaning of the questions.

Survey questions can be presented one question per screen (the interactive format for web questionnaires; see Couper, 2000) or questions can be shown on a single HTML page (the scrollable format). A key difference between interactive and scrollable web surveys is that in the latter respondents can browse the entire survey before answering a single question. Scrollable web surveys are very similar to mail surveys, while dynamic web surveys are comparable to interviewer-administered surveys in that they do not allow respondents to read ahead. The different formats may have different implications for measurement error. For instance, Dillman and Bowker (2001) argued that interactive web surveys result in a lack of context for the questions and could enhance order effects since respondents are not able to see the entire survey.

If multiple questions are presented on the same screen, they can be presented in a grid (or matrix) or as a series of individual items (Graf, 2002). With the grid format, many questions, which usually share the same answer categories, are positioned in a table. The question text is located in the far left-hand column of the table, and the answer categories fill the columns to the right-hand edge of the screen.

From the survey researchers' point of view, the physical arrangement of the questions is a style element; the decisions on which one to use are more likely to be driven by considerations of space, cost, and convenience. However, this style element is also a formal feature that may be viewed as a task element by the respondents, leading to interpretive errors.

For instance, one of the interpretive heuristics used by web respondents (see, Tourangeau, Couper, & Conrad, 2004) – the “Near Means Related” heuristic – may lead them to expect items that are physically near each other on the screen to be closely related conceptually. As a result, respondents will see stronger similarities among items that are displayed on a single screen than among those displayed on separate screens, boosting the correlations among them. Tourangeau and his colleagues presented eight related items either on eight separate screens (“interactive” presentation), on two screens with four questions on each screen in a grid (i.e., two grids on two separate screens), or in a single grid on one screen (Tourangeau, Couper, & Conrad, 2004). Respondents were randomly assigned to one of the three presentation conditions. The study found higher intercorrelations among the items when they were in a grid on one screen than when they were spread across two screens or presented on eight separate screens (Tourangeau, Couper, & Conrad, 2004). In addition, Tourangeau and his colleagues found evidence that respondents inferred more similarity among the items than was warranted when the items were in a single grid; respondents were more prone to select the same answers for the eight items and were less sensitive to the fact that two of them were actually reverse-worded when the eight items were in a single grid (Tourangeau, Couper, & Conrad, 2004).

A second web study offered further support that the grouping of related items on a single screen is likely to lead respondents to view the items as more strongly related, increasing the correlation among them (Couper, Traugott, & Lamias, 2001). In their second experiment, Couper and his colleagues varied the physical presentation of a group of questions by displaying them either in a grid on one screen (the ‘multi-item screen’ condition in the paper) or on separate screens (the ‘single-item screen’ condition). As expected, the inter-item correlations were consistently higher in the multi-item screen condition than in the single-item screen condition. However, the overall effect was not large, and none of the differences between each pair of correlations reach statistical significance (Couper, Traugott, & Lamias, 2001).

Reips (2002) also examined the impact of the grouping of items by screens on respondents’ answers to earlier items in a web survey. He displayed two items asking respondents’ potential expenditures on donations and on the Internet connection either on one or two screens. Respondents tended to allocate more of their income to the first item than to the second item when the two items were presented on separate pages; however, when the two questions were displayed on the same page, respondents allocated more of their income to the second item than to the first (Reips, 2002).

Even though it is hard to say what pragmatic implicatures were generated in Reips’ study, all three studies suggest that the physical arrangement of survey questions in a web survey provides cues to respondents, affecting their comprehension of the questions. Respondents take the style element of arrangement on screen as a task element.

I varied the physical arrangement of a set of six questions in a web survey. The survey presented the six items in one of the three ways: 1) one question per screen; 2) all six on the same screen; or 3) in a grid on a single screen. Unlike the studies by Tourangeau et al. (2004) and Couper et al. (2004), the six items in this study may not seem related at first sight, even though they shared the same theme (risky behaviors). Nonetheless, if respondents regarded the physical arrangement as meaningful, they would infer the strongest relation among the items when they were presented in a grid regardless of the content of the questions. This is especially true when the items were embedded in a multiple-page web survey. To respondents, it is natural to infer that the survey designers put the items together in a grid or on a single screen because they are related – why else would they be grouped together? Accordingly, I predicted that respondents would infer greater relatedness among the items when they were displayed in a single grid and that their answers would show higher intercorrelations than when the items were presented in other formats.

This experiment also varied the introduction to the items. One introduction – the facilitating introduction – indicated that the questions were taken from the same source (thus, encouraging respondents to apply Grice's maxim of relation); a second introduction – the inhibitory introduction – indicated that the items came from different sources so as to discourage the application of the relation maxim. The third introduction was intended to be neutral and didn't reveal anything about the relatedness (or unrelatedness) of the questions. The purpose of the introduction variable was to see whether the verbal introduction might strengthen or override the inference from the physical arrangement of the survey questions.

2.2 Method

Overview. The data came from a web survey conducted by MSInteractive. Survey Sampling Inc. (SSI) selected the sample for the web survey, using two different frames – the Survey Spot frame and the America Online Opinion Place. The Survey Spot frame includes more than a million web users who have signed up to receive survey invitations. SSI selected 29,772 email addresses for the survey and sent out email messages inviting the recipients to take part in “a study of attitudes and lifestyle.” The email invitations included the URL of the web site with our questionnaire and a unique ID number (which prevented respondents from completing the survey more than once). A total of 1,361 respondents completed the survey for a response rate of 5% (AAPOR [2000] RR1).

The America Online Opinion Place provides access to approximately 25 million AOL account holders, which is estimated to include about 50 million individuals worldwide. Opinion Place uses a technique in which survey invitations are posted on banners throughout the AOL service and related sites. Users willing to click through are screened against the respondent requirements for active surveys and then passed through to a survey for which they qualify. Respondents who complete a survey accrue miles in the American Airlines AAdvantage Program. This sampling technique makes it impossible to compute a response rate. The study was fielded from November 5, 2004, through November 14, 2004, and included questions on a range of topics such as health, diet, and travel. Since my goal is not generalization to a population but rather analysis of

differences between experimental treatments, all the analyses were done on unweighted data from both sample sources. A total of 2,587 respondents completed the survey.¹

Experimental manipulation. The experiment employed a 3 (physical arrangement of the survey questions) by 3 (introductions) factorial design. Crossing the physical arrangement and introduction variables produced the nine experimental conditions in Table 2.1 (see Table 2.2 for exact wording of the three introductions). Respondents were randomly assigned to one of the nine conditions. Table 2.1 also shows the number of completed cases in each cell of the design.²

Table 2.1. Number of Completed Cases Per Experimental Condition

	One Question per Screen	All On One Screen	In a Grid	Total
Facilitating Introduction	304	269	248	821
Neutral Introduction	293	280	273	846
Inhibiting Introduction	298	316	306	920
Total	895	865	827	2,587

Target questions. A set of six questions asked respondents about various risky behaviors (see Table 2.2. for exact wording). The questions ranged from whether the respondents wore seatbelts to whether they would use the sunscreen if they stayed in the sun for more than an hour to whether they took fruits everyday. On first sight, the six questions might not be very related, even though the questions were all about risky behaviors.

Follow-up questions. Three follow-up questions were intended to assess respondents' inferences about the relatedness of the six target questions. One question

¹ I analyzed the data by sample and found no noteworthy differences in the results.

² For approximately half of the respondents, this experiment was placed in the middle of the survey. For the other half, it came in the front part of the survey. However, the location didn't change the results and I ignore this variable in the analyses below. Another manipulation has to do with the response categories of the target items. Half of the respondents had same response categories for all six target items, while the other half got different response categories. Again, the analyses below ignore this variable, which had no noticeable effects.

asked respondents to rate the relatedness of the target questions on a seven-point scale.

The other two questions asked respondents what the six target questions were about (see Table 2.2 for exact wording).

Table 2.2. Experiment 1: Introductions, Target Items and Follow-up Questions

Introduction	
Facilitating Introduction: The following questions are taken from a standardized psychological battery that measures a person's tendency to take risky behaviors.	
Neutral Introduction: Here are a few more questions for you...	
Inhibitory Introduction: The following questions were selected from our database of questions from different online surveys about lifestyles.	
Target Items	
Same Response Categories	Different Response Categories
Do you always use seatbelts when you drive or ride in a car, van, sports utility vehicle (SUV), or pick-up?	How often do you use seatbelts when you drive or ride in a car, van, sports utility vehicle (SUV), or pick-up?
Between not having medical insurance and buying medical insurance with your own money, would you opt for not having a medical insurance?	Between not having medical insurance and buying medical insurance with your own money, which one would you opt for?
When you ride a bicycle, do you always wear a helmet?	When you ride a bicycle, do you always wear a helmet?
Do you think it is safe to keep a firearm in or around your house?	Do you think it is safe to keep a firearm in or around your house?
Do you always use sunscreen of SPF 15 or greater when you go out in the sun for more than 1 hour?	How often do you use sunscreen of SPF 15 or greater when you go out in the sun for more than 1 hour?
Not counting juice, do you eat fruit every day?	Not counting juice, do you eat fruit every day?
Follow-up Questions	
Without looking back, what was your impression of the items on the last six screens/last screen? What do you think these six items were about? (Open-ended)	
Would you say that the 6 items are about....	
1 Recreational activities	
2 Risky behaviors	
3 Life styles	
4 American culture	
5 Medical care	
6 Some other topics	
Sometimes surveys ask questions about closely related topics, but other times they shift from one unrelated topic to the next. Again, based on your impression of the last six items on the last six screens/last screen, how much were the items related?	
1 = Very closely related	
7 = Not at all related	

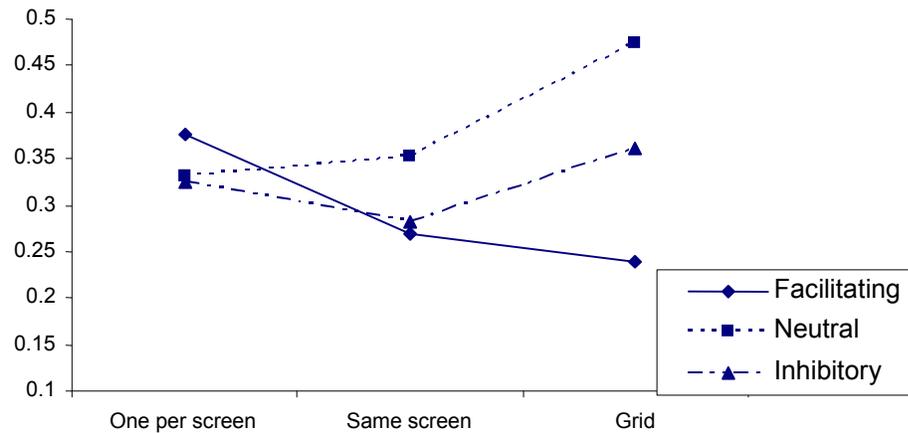
2.3 Results

My main hypotheses involved the correlations among responses by experimental condition and the inferences respondents derived about the relation among the items. However, I checked the survey answers by conditions first.

Survey responses. To examine whether responses varied by experimental condition, I computed a risk score for every respondent based on their responses to the six target questions. A two-way ANOVA on the risk scores showed that the risk scores didn't vary by experimental condition. Neither the main effects nor the interaction of the two experimental factors had significant impact on risk scores.

Intercorrelations. I predicted that the highest intercorrelations would be observed in the grid condition and when the respondents got the facilitating introduction. The lowest intercorrelations should occur when the items were spread across six screens and when the respondents got the inhibitory introduction. I computed the Cronbach's alpha among the six items across nine experimental conditions. Figure 2.1 presents the alphas by experimental condition. Neither the main effect of the physical arrangement nor the interaction was significant, though the main effect of the introduction was marginally significant ($F(2, \infty) = 2.39, p < .10$).

Figure 2.1. Plots of Cronbach's Alphas by Experimental Condition



Only the neutral introduction shows the hypothesized trend with the highest intercorrelations in the grid condition, followed by the same screen condition, and the separate screen condition (see the dotted lines in Figure 2.1). A χ^2 test of the linear trend is marginally significant ($\chi^2(2)=4.65, p=.098$), showing that the alphas are significantly different across the three physical arrangement conditions under the neutral introduction condition. This replicates the pattern reported by Tourangeau et al. (2004) and Couper et al. (2001). The same tests done for the facilitating and the inhibitory introduction conditions didn't yield significant differences across physical layouts.

The spread of the alphas across different introductions is largest under the grid condition; the χ^2 test of the differences across the three correlations is highly significant ($\chi^2(2)=9.23, p<.001$). Respondents seemed to be affected more by the introductions when the questions are laid out in a grid. However, the direction of this effect is contrary to my hypothesis. It seems that the facilitating introduction backfired, appreciably lowering the correlations among the items. This is surprising at the first sight. However, careful examination of the literature shows it might be reasonable.

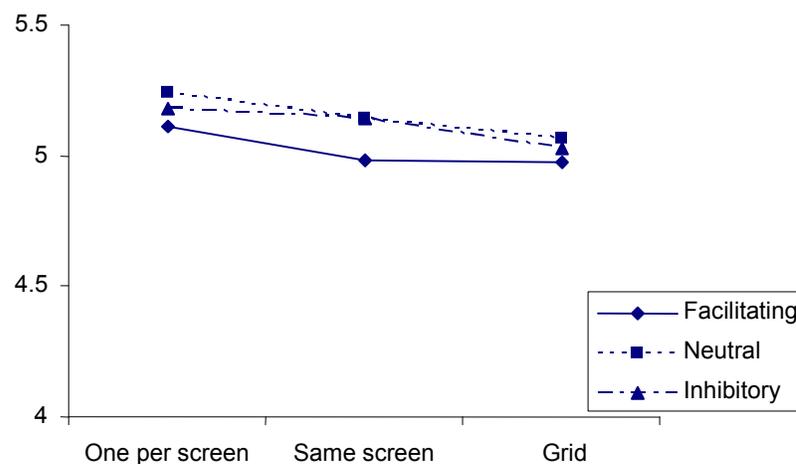
Sudman, Bradburn, and Schwarz (1996, p.122) pointed out that a conversational norm of nonredundancy (i.e., the maxim of quantity) may be invoked when related questions are perceived as belonging to the same conversational context. Among the variables that could trigger the norm of nonredundancy are the introductions to a group of related items. Schwarz, Strack, and Mai (1991) showed the importance of joint lead-in at triggering the application of the maxim of quantity, and Strack and his colleagues showed that presenting two questions in a single box – thus emphasizing their relatedness – elicited contrast effects (Strack, Schwarz, & Wanke, 1991). An earlier study, by Metzner and Mann (1953), demonstrated that the correlations among questions grouped together under a caption stating the topic, were smaller than those correlations among same questions when they were interspersed among questions on other topics almost as often as they were larger. Grouping questions about a subject does not invariably increase the correlations among the questions (Metzner & Mann, 1953).

In this study, then, when the introduction revealed nothing about the relatedness of the questions, the formal feature – i.e., arrangement of the questions in the grid – led respondents to infer that the questions were conceptually related, and higher intercorrelations were obtained. However, when the introduction told the respondents that the items were related, creating a shared conversational context, it seemed to trigger the maxim of quantity, leading respondents to look for distinctions among the items; as a result, the intercorrelations were much lower.

Inferences. I hypothesized that if respondents applied the maxim of relation, they would infer the strongest relation among the items when the items were presented in a grid, the next strongest when the items were on the same screen, and the weakest relation

when the items were spread across six separate screens. Furthermore, the facilitating introduction would reinforce the perception of relatedness while the inhibitory introduction would reduce it. I assessed respondents' inferences about the relation among the six items with a follow-up question administered after the six target questions.³ The follow-up question asked respondents to rate how related the six items were on a rating scale that ranged from 1 ("Very closely related") to 7 ("Not at all related"). I reverse coded the question so that higher numbers mean greater perceived relatedness. Figure 2.2 plots the mean perceived relatedness rating by experimental conditions. Neither the main effects nor the interaction is significant in a 2-way ANOVA of the means.

Figure 2.2. Mean Perceived Relatedness Rating by Experimental Condition



Contrary to my hypothesis, none of the lines exhibited the linear trend as hypothesized. It seems that the formal feature – the physical arrangement of the survey

³ The answers to the other two follow-up questions that asked respondents what the six target items were about were not reported here because the facilitating introduction contained a key phrase “risky behavior,” which influenced answers to these two questions (the same phrase was even included as a response alternative to the closed-end question). Thus, more respondents given the facilitating introduction listed or chose ‘risky behaviors’ in their answers to both questions, creating an artificial significant main effect of the introduction variable. The physical arrangement of questions didn’t have a significant main effect on responses to either question; neither did the interaction.

questions – didn't trigger the inference of relatedness as hypothesized. In addition, the introduction variable did not strengthen (or suppress) the inference effectively.

2.4 Conclusions

This experiment manipulated the physical arrangement of a set of six questions by presenting them one item per screen, all on the same screen, or in a grid on one screen. Contrary to the hypotheses, the visual cue of a grid did not lead respondents to see the items as more strongly related; they tended (nonsignificantly) to rate the items as *least* related when the items were displayed in a grid. On the other hand, the grid condition produced the highest intercorrelations (measured by Cronbach's alphas) among the six items, partially replicating the previous research on the effects of physical arrangement (see Couper et al., 2001; Tourangeau et al., 2004). However, the pattern for the intercorrelations by the physical arrangement of the items depended on introductions I gave respondents. When the introduction did not mention anything about the relatedness or unrelatedness of the items, the alphas followed the hypothesized trend – the highest intercorrelations were observed in the grid condition, the second highest in the same screen condition, and the lowest in the one item per screen condition. A different picture emerged when the introduction told the respondents that the items were related or unrelated. For the facilitating introduction in particular, the grid condition seemed to serve as a shared conversational context and invoked the maxim of quantity. Respondents seemed to make a distinction among the items rather than regarding them as related; this matches the findings by Strack and his colleagues (Strack et al., 1991). The reaction time data also suggested a switch of maxims – respondents were slower when

given a facilitating introduction (mean response time=51.7 seconds) than when given a neutral introduction (48.0 seconds) or an inhibitory introduction (49.1 seconds), though the difference didn't approach significance ($F(2,824)=1.38, p=.25$).

The differential effect of introductions on the correlations in the grid condition is worth further investigation. Although this particular experiment failed to produce conclusive evidence, we should be more careful about using grids since they may encourage respondents to exaggerate the similarities or discrepancies among a set of related items, depending on the introductions given. In addition, in the neutral introduction condition, the marginally significant differences in the correlations among target items across physical layouts should alert us to the different performance in the grid and the same screen conditions. The two physical layout conditions are similar in that respondents were able to read the whole set of questions before answering. However, the correlations obtained from these two layouts were not similar. This suggests that respondents draw different inferences drawn from the layouts.

On the whole, this experiment didn't bear out the hypotheses regarding the effects of the physical arrangements of survey questions and the introductions on correlations. The inference obtained didn't show convincingly that respondents applied the maxim of relation (or the "near means related" heuristic in Tourangeau et al., 2004). One possible reason that respondents' inferences from the physical layout of the questions were not in the hypothesized direction is that the follow-up question I used to capture respondents' inferences was flawed. Although I tried to balance the wording of the follow-up item (see Table 2.2 for the exact wording), the question "how much were the items related?" may carry the presupposition that the items are related (cf. Knauper, 1998; Sterngold,

Warland, & Herrmann, 1994; see Chapter 4 of this dissertation). In addition, the presupposition might carry a stronger weight under the one-item-per-screen condition. This layout is the default for web surveys and it may not carry any implication about the relatedness of the items; as a result, the respondents may give the question presupposition a greater weight. The fact that average responses to this relatedness question were above the scale midpoint – a rating of 4 – in all conditions (see Figure 2.2) is consistent with this speculation.

Chapter 3 The Maxim of Manner and Providing Definitions

3.1 Introduction

According to the survey response process framework, survey responding starts with comprehension of survey questions (Tourangeau, Rips, & Rasinski, 2000).

Incorrect understanding of survey questions will affect – directly or indirectly – the retrieval of relevant information, the estimation and judgment strategies used, and even the mapping of an answer to one of the response options (Conrad & Schober, 2000; Schober & Conrad, 1997). Survey researchers have known for some time that, despite their best attempts to write clear questions, respondents have problems comprehending survey questions (Belson, 1981; Conrad & Schober, 2000; Fowler, 1992; Schober & Conrad, 1997).

There are at least two types of comprehension problems, both of which may affect the validity of the survey data. First of all, different respondents may interpret the same question very differently. In a well-known study, Belson (1981) showed that respondents vary in their interpretation of even simple words such as “you.” For instance, while 84.6% of the respondents believed that the term “you” referred to the individual alone, 1.9% thought it meant the individual and his family. For another 3.8% of respondents, “you” was the individual and the spouse. The other 3.8% considered “you” as a combination of the individual plus at least one other family member. What is more alarming is that 5.8% of respondents simply overlooked the term (Belson, 1981). Such a lack of consensus on the meaning of key survey terms (and survey questions as a whole) could lead to systematic variations, affecting the comparability of data across respondents, particularly if the differences of interpretation are large and coincide with boundaries of

subgroups defined by culture, race, age or other characteristics (Martin, Campanelli, & Fay, 1991). It could also reduce survey researchers' ability to reach valid conclusions about relationships based on the data because of the error in measurement (see Fuller, 1987, on the effects of measurement error on regression coefficients and other statistics).

Second, respondents could be answering based on the same interpretation of the question, but one that does not fit the survey researchers' definitions. Respondents' interpretations may disagree with a survey definition by either being too broad or too narrow compared with the meaning intended in the survey. Suessbrick, Schober, and Conrad (2000) reported, for instance, that about 46% of respondents considered "smoking" in the first question in the Tobacco Use Supplement to the Current Population Survey ("Have you smoked at least 100 cigarettes in your entire life?") as "only puffs inhaled," whereas the survey's definition included all puffs, whether inhaled or not. Such a difference between respondents' interpretations and survey's definitions could contribute to bias in the survey estimates.

Both variability across respondents and systematic misinterpretation jeopardize survey quality. Offering respondents definitions to clarify unclear or ambiguous terms is one possible solution. There is empirical evidence that providing definitions helps improve respondents' comprehension. For example, Fowler (1992) revised seven questions from health surveys conducted by government agencies or academic survey organizations by offering definitions to clarify unclear terms. He found that the revised questions improved comprehension; the rates of requests for clarification and rates of inadequate answers declined (Fowler, 1992).

Similarly, Conrad and Schober demonstrated that uniformity of interpretation – and thus data quality – could be increased dramatically when respondents are provided with clarification about the meaning of the words in the questions (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 1999; Suessbrick, Schober, & Conrad, 2000). For instance, Schober and Conrad (1997) showed that allowing interviewers to provide definitions when the questions involved complicated mapping – that is, when the respondents’ situations didn’t map to the key terms in the questions in a straightforward way – greatly improved the accuracy of responses (87%) relative to interviewers who were not allowed to provide definitions (27% accuracy).

In another study, Conrad and Schober (2000) compared response changes in a reinterview due to different interviewing techniques. They observed that 22% of respondents changed their answers, on average, when the reinterview was conducted as a conversational interview in which interviewers provided definitions to respondents. By contrast, only about half as many changes occurred when the second interview was still standardized.

The existing research on the effects of definitions on survey responses has focused more on the mode of presenting definitions than on the content of the definitions themselves (see Tourangeau & Conrad, 2004, for an exception). These earlier studies have examined such aspects of offering definitions as the trigger for providing definitions (a request from the respondent, the interviewers’ own initiative, or an automated feature in a web survey) and the accessibility of definitions (always shown, shown when the respondent click or rollover, and so on) in web surveys (see, for example, Bloom & Schober, 1999; Conrad, Couper, Tourangeau, & Peytchev, in press; Lind, Schober, &

Conrad, 2001; Schober, Conrad, & Bloom, 2000; Schober, Conrad, & Fricker, 1999). These studies uncovered some problems with offering definitions to respondents to improve comprehension of the question.

One major problem is that respondents almost never requested or retrieved definitions either from the interviewer in interviewer-administered interviews or from the automated questionnaire in the web studies (Conrad, Couper, Tourangeau, & Peytchev, in press; Schober, Conrad, & Fricker, 1999; Suessbrick, Schober, & Conrad, 2000). One likely reason that respondents rarely requested a definition is that it didn't occur to the survey respondents (or the interviewers) that their interpretations of the questions might differ from anyone else's. Respondents seem to be following a "presumption of interpretability" in responding to survey questions (Clark & Schober, 1992; Schwarz & Oyserman, 2001). The presumption – which is operative in daily conversation – is that the utterance has been tailored to the listener and contains all the information needed to interpret it.

The interpretability presumption is entailed by Grice's Cooperative Principle, which requires speakers to contribute to the common conversational goal. Listeners, for their part, assume that both what is said and how it is said are relevant to the conversational goal and critical to understanding. In the survey setting, respondents assume that the survey researcher has chosen his wording so they will understand it as intended (Clark & Schober, 1992). Thus, respondents assume that the most obvious meaning is likely to be the correct one, and if they cannot find an obvious meaning, they will look to the immediate context of the question to determine one.

A second problem with the alternative of offering definitions to respondents has to do with the effort required to access a definition. If it requires much effort to access the definition, respondents are unlikely to bother (Conrad et al., in press). Conrad and his colleagues showed that few respondents requested a definition if retrieving it required more than one mouse click. This seems to lend support to the notion of “satisficing” (Krosnick, 1991) – respondents are unwilling to expend effort to achieve optimal responding.

The existing work also shows an increased interview time on average when definitions are requested, retrieved, or consulted (Conrad et al., in press; Conrad & Schober, 2000; Schober & Conrad, 1997). This seemed to provide further proof for the speculation that processing definitions requires efforts. In combination of the second problem described above, it is unclear whether respondents would be willing to spend the efforts to process the definition carefully even if they did access it.

Groves and his colleagues argue that there is a tension between explicitly defining terms in a question (in an attempt to eliminate ambiguity) and increasing the burden on the respondents to absorb the full intent of the question (Groves, Couper, Fowler, Lepkowski, Singer, & Tourangeau, 2004). Similarly, Tourangeau, Rips, and Rasinski (2000: Chapter 2) argue that attempts to clarify terms can lead to syntactically complex questions. The trade-off involves the amount of information to give in a definition so that it clarifies meaning without seeming redundant or adding too much complexity.

This tension becomes more acute when survey researchers define terms used everyday, terms such as “you,” “child,” “poultry,” and so on. Defining terms that everyone already understands violates the maxim of manner (Grice, 1989). That maxim

enjoins speakers to be brief, clear, and orderly, and to avoid unnecessary ambiguity and wordiness. Violations of the Gricean maxim tend to generate conversational implicature – an inference listeners work out to maintain the overarching assumption that the speaker is being cooperative. Defining everyday terms could suggest to respondents that the everyday terms in the survey questions are being used in some special or technical sense or that the definitions are intended for a subpopulation that needs them (e.g., non-native speakers). These implicatures may give rise more confusion than clarification.

Applied work in the field of computer-human interaction provides some empirical evidence for the confusions that can be caused by apparent violation of the Cooperative Principle. Young (1999) compared instruction descriptions generated according to Grice's CP to an "exhaustive plan," which gives the most detailed (and most redundant) instructions on every single step of the task, and to a "primitive plan," which describes only the lowest-level steps in the task. In other words, the three instructions differed sharply in the amount of information they contained. Subjects were asked to carry out a task described by the instructions within a computer simulation. Young (1999) found that subjects given the instructions produced according to the CP committed fewer errors and achieved more of their top-level goals than subjects who got either of the other two sets of instructions. Specifically, the "exhaustive plan" led to the highest failure rate compared to the other two (Young, 1999). Giving more information than was necessary didn't improve performance.

A similar finding was reported by Gerber and her colleagues, who explored how to convey the notion of "residence" to respondents in an effort to improve the accuracy of their responses to roster questions like those used in the decennial census (Gerber,

Wellens, & Keeley, 1996). Gerber and her coauthors found that providing definitional information about census rules on residence resulted in fewer correct answers for questions on simple and straightforward living situations. In two of five instances of straightforward living arrangements, the decreases were fairly substantial (15-18% decreases in correct answers). Gerber and her colleagues speculated that accuracy decreased because respondents regard the presentation of definitional rules which they already “know” as redundant; respondents may, as a result, reinterpret these rules in an effort to make sense of them (Gerber, Wellens, & Keeley, 1996).

The findings of Young (1999) and Gerber et al. (1996) raise the question of how much information to include in a definition or a set of instructions. This question is important since any definition can appear self-evident and redundant to those portions of the target population who share the same definition as the researcher. It will become a bigger issue especially as surveys are designed to accommodate populations that vary more in these linguistic backgrounds.

This study examines the potential costs of offering explicit definitions for everyday terms. It focuses on the effects of offering definitions on comprehension and data quality when the definitions fail to provide new information to respondents. This will often happen in practice since, for many respondents, their definitions of these terms will coincide with those of the survey. For these respondents, the definitions will seem unnecessary, violating the maxims of manner. The current study employs definitions that were designed not to provide new information.

The main hypotheses concern three dependent variables – responses to the questions, the response times, and respondents’ inferences based on the definitions.

When respondents are offered self-evident definitions for everyday terms, I predict that they will recognize that the definitions aren't necessary; accordingly, they will become confused, try to work out an inference to explain why a definition was provided, and shape their responses based on the inference. I tested the following hypotheses:

- 1) Giving a definition to an everyday term leads to different survey responses from when no definition is offered. Specifically, the variance and covariance matrix for the items will differ.
- 2) Respondents who are given unneeded definitions will take longer on average to arrive at an answer than those who are not offered the definitions.
- 3) Respondents who are given definitions to everyday terms will infer that these terms are used in some technical sense or that the definitions are intended for some special population.

3.2 Method

Overview. The experiment was one of several experiments embedded in a questionnaire administered via audio computer-assisted self-interviewing (ACASI). The questionnaire covered a range of topics, most of them political topics. The questionnaire included several experiments; most of them involved response order. We recruited 160 participants from the College Park, Greenbelt, and Silver Spring, MD area. We placed recruitment flyers at local libraries, advertised the study in local papers, and sent an e-mail invitation to the staff and graduate students at the University of Maryland. Since we desired a heterogeneous sample, we restricted the number of undergraduate students to

less than 40 (one fourth of the targeted number of completes). Participants came to the Joint Program in Survey Methodology, where they first completed a 40-45 minute survey on a computer. The questionnaire was programmed in Blaise. Participants could listen to the questions (and response options) via earphones, read them displayed on the computer screen, or both. As I note below, the speed of the voice reading the questions was systematically varied. Participants indicated their answers by typing in the number corresponding to one of the answer options or by typing in text (in response to the open-ended questions). They then completed a paper-and-pencil questionnaire assessing the “Big Five” personality traits as well as their need for cognition (Cacioppo, Petty, & Kao, 1984). These questions took about 10 minutes. Participants were paid \$25 upon completion of two questionnaires. The experiment ran from January 10, 2005, to February 28, 2005.

Experimental manipulation. My experiment was placed in the middle part of the questionnaire. Participants were randomly assigned to one of the two experimental conditions. In one condition, redundant definitions for everyday terms were embedded in the question text and provided to participants for all four key survey terms (poultry, fat, vegetable, and red meat). For example, here is the poultry item with its accompanying definition:

“We will first ask you about how much poultry you eat. We define poultry as domestic fowl raised for meat. During the last 6 months, how much poultry did you typically consume?”

In the other condition, participants were asked the same questions with the same key terms. However, no definitions were given. An example of the wording is given below:

“We are interested in studying Americans’ consumption of poultry. We will first ask you about how much poultry you eat. During the last 6 months, how much poultry did you typically consume?”

The other experimental variable is related to the larger A-CASI study that included my experiment. We used synthesized voices that permitted us to systematically vary the rate of speech.⁴ There were three speed conditions: slow speed, fast speed, and fast speed with pauses between the response options. Participants were randomly assigned to one of the three speed conditions. This randomization was independent of the random assignment to definition conditions. Although this manipulation is not directly relevant to my experiment, some of my analyses control for voice speed. Table 3.1 summarizes the number of participants in the two (definition or no definition) by three (slow speed, fast speed, or fast speed with pauses) experimental design.

Table 3.1. Number of Participants Assigned to Experimental Conditions

	Given Definition	No Definition	Total
Slow Speed	27	25	52
Fast Speed	22	30	52
Fast Speed with Pauses	26	30	56
Total	75	85	160

Target questions. Respondents were asked about four food categories (poultry, fat, vegetable, and red meat). For each food category, respondents were asked about their typical consumption (e.g., “During the last 6 months, how much poultry did you typically consume?”) and whether they tried to consume or avoid that food category (e.g., “Thinking about the food you eat, is poultry something you actively try to include in your diet, something you actively try to avoid, or something you do not think about either way?”). For some categories, the respondents were also asked to judge whether a

⁴ The ACASI voice was a synthesized voice, generated by the AT&T Natural Voices® software, developed by AT&T Laboratories.

specific food belonged to that food category (e.g., “Do you consider Cornish hens to be poultry?”).

Follow-up questions. I assessed respondents’ inferences about the food categories with two follow-up questions. The first asked respondents whom they thought the study was intended for and the second asked whether respondents believed these terms were used in a technical sense or their ordinary sense. In addition, respondents also rated themselves on their diet consciousness and health consciousness along seven-point scales (where the endpoints were labeled).

3.3 Results

The analyses focus on three outcome variables – the interrelations among the responses, response times, and respondents’ inferences. I begin my analyses with the effect of definitions on responses to the target questions.⁵

Effects on the responses. Providing self-evident definitions to everyday terms breaches the maxim of manner. If respondents noticed the violation and used it to make an inference about the meaning of the questions, then their answers might be affected. I first compared the variance in responses to the four key survey questions when definitions were offered versus when they were not. Table 3.2 shows an apparent trend for smaller variances when definitions were offered than when they were not offered. *F*-tests showed that providing definitions significantly reduced the variance for two of the four target questions (consumption of poultry and consumption of red meat). The

⁵ Answers to the 14 target questions didn’t vary by whether self-evident definitions were offered or not. Responses to one (out of 14) question were affected significantly by voice condition. The two experimental factors had significant interaction effects on responses to two other questions.

reduction in variance indicates only that responses were more uniform across respondents; without true values to compare the answers to, it is hard to tell whether the reduction in variance also represents a reduction in measurement error.

Table 3.2. Variances of Responses to Four Target Questions By Definition Condition

	Definitions offered	Definitions not offered	<i>F</i> values
Consumption of poultry	0.488	0.853	$F(74,84)=1.75, p=.01$
Consumption of fat	0.907	0.727	$F(84,74)=1.24, p=.16$
Consumption of vegetable	0.728	0.766	$F(74,84)=1.05, p=.41$
Consumption of red meat	0.916	1.333	$F(74,84)=1.46, p=.05$

I also looked at whether the definitions helped respondents classify borderline instances of a category. I asked respondents about three borderline instances—whether eggs are poultry; whether potatoes are vegetables; and whether ham is red meat. I also asked about more prototypical exemplars of the target categories – Cornish hen, broccoli, and steak – and nearly everyone classified these correctly with or without a definition. For the borderline instances, responses were somewhat evenly divided (see Table 3.3). The definitions didn’t systematically move respondents either way. Furthermore, the definitions didn’t significantly influence the speed with which respondents responded to the questions.

Table 3.3. Percentage of “Yes” Responses to Borderline Instances by Definition Condition

	Given Definition (%)	No Definition (%)	χ^2 test results
Eggs	34.1	46.0	$\chi^2=2.31 p=.13$
Potatoes	63.5	61.3	$\chi^2=.08 p=.77$
Ham	47.6	44.0	$\chi^2=.21 p=.64$

For each of the four key target terms, I also examined the relation of responses to the target question and several related questions. I fit multiple regression models with the target question as the dependent variable and the related questions as the predictors. If

unnecessary definitions confuse respondents, their answers may be more error prone, reducing their correlation with the predictors. Thus, the focus of these analyses is the interaction effects between the definition and the other predictors. Table 3.4 displays the regression coefficients from the four multiple regressions. The bolded estimates are significant at the $\alpha=.05$ level. The p values are given for marginally significant coefficient estimates, but not for non-significant estimates. Overall, four out of the 24 interactions between the predictors and the definition variable were significant, indicating that the definition altered the relation between the predictor and the target dietary consumption variable. However, of the four significant interactions, two are positive and two are negative. Thus, there didn't seem to be a consistent trend as to the size and direction of the significant interaction effects. Although giving definitions to everyday terms did have some effect on the underlying covariance matrix, these effects were neither consistent nor strong.

Response times. I hypothesized that the total response time would be longer for respondents given definitions to the everyday terms than for those who didn't get definitions. We recorded the time from the moment the audio started to read the question until the moment when the respondent chose an answer. Thus, our measure of the response time encompasses the reading time for the question by the audio plus the time respondents took to answer the question. This measure poses a problem since the respondents' actual response time is confounded with the voice speed for the audio. The reaction times would almost necessarily be longer in the slow-speed voice condition than in the fast-speed voice conditions. Table 3.5 presents the results.

Table 3.4. Regression Coefficients from Multiple Regression Models by Target Item

Dependent Variable		Independent Variable			
		Consumption of Poultry (β)	Consumption of Fat (β)	Consumption of Vegetable (β)	Consumption of Red Meat (β)
Definition	Main Effect	1.86 ($p=.08$)	-0.52	-1.44	-3.63 ($p=.01$)
Consumption of poultry	Interaction effect with definition	---	-0.15	.031	0.37 ($p=.07$)
Consumption of fat	Interaction effect	-0.31 ($p=.03$)	---	-0.08	0.05
Consumption of vegetable	Interaction effect	-0.01	-0.19	---	0.14
Consumption of red meat	Interaction effect	0.21 ($p=.09$)	0.12	0.09	---
Diet conscious	Interaction effect	-0.18 ($p=.03$)	0.10	-0.03	0.36 ($p=.00$)
Gender	Interaction effect	-0.12	0.19	0.01	0.04
Education	Interaction effect	-0.18	0.15	0.38 ($p=.01$)	0.26

Table 3.5. ANOVA Results on Total Response Time

Voice Condition	No Definition	Given Definition
Slow speed	210.4	213.6
Fast speed	177.3	178.2
Fast speed with pauses	199.6	185.5
ANOVA Results		
Definition	$F(1,154)=0.14$	ns
Voice Condition	$F(2,154)=4.96$	$p=0.01$
Interaction	$F(2,154)=0.39$	ns

Results from a two-way ANOVA confirmed that voice condition made a big difference in the total response time (see Table 3.5). Attempting to tease apart the variation in response times caused by the experimental manipulation of the voice speed, I standardized the response times under each voice condition and analyzed these standardized times. Since response times are usually highly skewed (cf. Ratcliff, 1993), I replaced the standardized response times that were above three standard deviations with the value three standard deviations above the mean (cf. Ratcliff, 1993; Van Zandt, 2002).

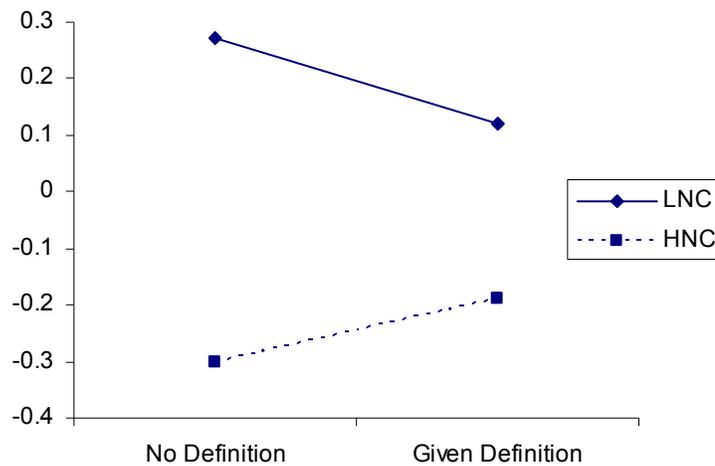
A one-way ANOVA on the total response time for this block of questions (with the factor being whether respondents were given a definition or not) revealed that respondents who were given the definitions turned out to be nonsignificantly faster (mean overall standardized response time = -0.030) than those who were not given a definition (0.027) ($F(1,158) < 1$, ns).

Additional analyses of the data showed that one respondent characteristic affected the response time – the need for cognition. According to Petty and Jarvis (1996), people with a high need for cognition (HNC) would normally process the questions more carefully than those with a low need for cognition (LNC respondents). Applying the conversational maxims requires cognitive effort; the HNC group is more likely to notice apparent violations of the conversational maxims and to draw implicatures based on apparent violations than their LNC counterparts. A study by McCabe and Brannon (2004) confirmed such a role for the need for cognition. They reported that HNC respondents applied the maxim of quantity to part-whole questions and displayed an attenuated correlation between the items, but not LNC respondents. Their finding suggested that the

conversational norm to avoid redundancy is not automatically applied in the survey context; only those with a high need for cognition seemed to apply the maxim.

Based on this reasoning, I compared average response times for these two groups (HNC respondents vs. LNC respondents) with and without definitions. Figure 3.1 plots the mean standardized response times for the two groups.

Figure 3.1. Mean Standardized Response Times for HNC and LNC Respondents



The figure partially confirms the hypothesis. Within the HNC group, those who got a definition did take longer (mean standardized response time=-.19) than their counterparts who were not given a definition (-.30). Even though the difference didn't approach significance ($F(1,71)<1$, ns), the direction was consistent with my hypothesis and replicated the finding by McCabe and Brannon (2004). On the other hand, the LNC group was slower on average (mean standardized response time=.19) than the HNC group (-.24).⁶ The difference in response times between the two need for cognition groups was significant in the no definition condition ($F(1,73)=5.90$, $p<0.02$), but not in

⁶ The longer response time by the LNC group could be explained by the positive correlation between education and need for cognition. The LNC group contains significantly more respondents who had less than a college degree. Thus, the LNC respondents are probably slower readers than the HNC respondents.

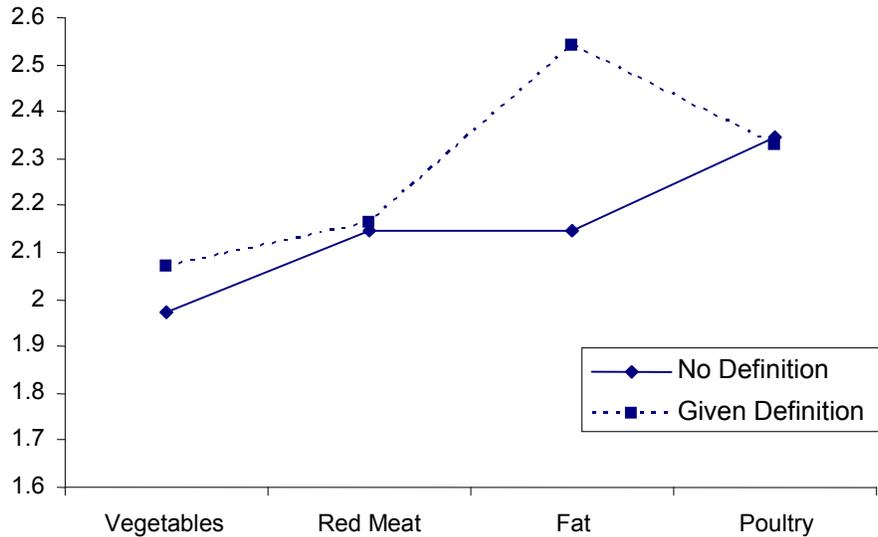
the definition condition ($F(1,83)=2.35, p=.13$). A three-way interaction between voice speed, definition condition, and need for cognition is marginally significant ($F(2,148)=2.22, p=.11$).

An item-level analysis of response times showed the same picture as the total response times. In general, the HNC respondents tended to be slower with definitions than without definitions, but the reverse was true for LNC respondents. None of the item-level interactions were significant either.

Inferences. I predicted that respondents who got the unnecessary definitions would conclude that the definitions were not intended for them or that the terms were not being used in their ordinary sense. I examined responses to the two follow-up questions to test for such inferences. One question specifically asked the respondents whom the survey was intended for. Respondents were somewhat more likely to think that the survey was intended for a special population rather than the general public when they got the unneeded definitions (7.0% of respondents given definitions vs. 3.6% of those without definitions), but the difference was not significant ($\chi^2=.72, p=.40$).

I also asked the respondents whether they believed the four key survey terms (poultry, vegetable, fat, and red meat) were used in their ordinary sense or in a special technical sense. All four of these follow-up items used the same scale, in which 1 meant the ordinary sense and 5 the technical sense. The mean responses are displayed in Figure 3.2. Higher numbers indicate ratings more in the direction of the technical sense.

Figure 3.2. Mean Ratings of Whether Survey Terms Are Used in Technical Sense by Definition Condition



It is clear from Figure 3.2 that when given definitions, respondents tended to regard the survey terms as used in a more technical way than their counterparts who were not shown definitions. The difference between the two groups of respondents was marginally significant for the term “Fat” ($F(1,158)=3.04, p=0.08$). Thus, the study offers limited support for the notion that respondents found the definitions unnecessary and drew inferences to account for them.

3.3 Conclusions

This study investigated the effects of offering definitions for everyday terms in surveys. I predicted that defining terms that don’t need definitions would violate the maxim of manner and create a linguistic anomaly; cooperative respondents recognize the anomaly and work out inferences to account for it. The results presented here lent at best weak support for the original hypotheses.

First, offering definitions to respondents did seem to influence their responses. Providing definitions reduced the variances of the responses, but the effects on regressions involving the key items are not easy to interpret and do not show a consistent picture. There is no clear evidence about whether offering redundant definitions improves or reduces data quality.

Second, some respondents did apparently process the definitions. Respondents with a high need for cognition were slower in answering the target questions when they were given definitions than when they were not (see Figure 3.1). Still, these differences in response times were not significant.

Third, respondents who got the definitions were somewhat more likely to infer that the survey was targeted at a special population and that the terms were used in a technical sense than those who didn't get the definitions. Again, though the direction was right, the trend was not significant.

The major limitation of this study lies in its small sample size, which inevitably reduces the power of the study. Even though offering definitions along with survey questions could promote more uniform interpretations of the questions, we need to be careful with the level of detail and the amount of information to be included in definitions. Respondents in this experiment noticed the self-evident definitions and seemed to have worked out inferences to account for them; the definitions also affected the variability in their responses. More systematic research should be carried out to investigate the effects of offering definitions on survey data quality. In the meanwhile, a thorough pretesting of survey concepts should go beyond detecting simple cognitive problems to investigate potential pragmatic issues related to offering definitions.

Chapter 4 Three Maxims

This chapter discusses four experiments that examine the maxims of relation, quantity, and quality. The four experiments examine respondents' use (or misuse) of these maxims in responding to web survey questions.

4.1 The Maxim of Relation and Numerical Values of Rating Scales

4.1.1 Introduction

Of Grice's four maxims, the maxim of relation is potentially the most powerful one. It enjoins participants to make their contribution relevant to the aims of the ongoing conversation. This maxim implies that listeners can assume that the speaker's contribution to a conversation is relevant to its goal, unless it is explicitly marked as irrelevant. If information has been included, it should be relevant; otherwise, why would the speaker have mentioned it (Hilton, 1995; Schwarz, 1996)? At the extreme, the maxim implies that every feature of an utterance is to be interpreted.

In survey research settings, Schwarz and Tourangeau and their colleagues have demonstrated that respondents make use of various visual features of survey questions in interpreting the questions (as shown in Chapter 1). Respondents consider the visual features of the questions to be relevant to survey responses; that is, they see the visual features not as style elements as the researchers may intend, but as task elements (cf. Couper, Tourangeau, & Kenyon, 2004). This inference of meaningfulness is based on the maxim of relation (cf. Grice, 1989). For example, the numerical values assigned to a rating scale are one feature that seems to be taken as a task element in the survey response process. Several studies by Schwarz and his colleagues demonstrated that the

numerical values assigned to the scale points affect the distribution of the responses (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991; Schwarz & Hippler, 1995; Schwarz, Grayson, & Knauper, 1998). Table 4.1 summarizes the main features of these studies.

In one experiment, Schwarz and his colleagues asked respondents in a face-to-face interview to evaluate their success in life along an 11-point rating scale, with one endpoint labeled “not at all successful” and the other “extremely successful” (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark 1991, Experiment 1). The scale was presented on a show card in the form of a ladder and ranged either from 0= “not at all successful” to 10= “extremely successful,” or from -5= ‘not at all successful’ to 5= “extremely successful.” Respondents were randomly assigned to one of the two numerical value conditions. Schwarz and colleagues found a mean shift in rating towards the higher end of the scale when the scale ran from -5 to 5 as compared to 0 to 10. Respondents were particularly unlikely to endorse a value between -5 and 0. A second experiment replicated the effect of numerical values with a self-administered questionnaire and demonstrated the effect both for self-reports and reports about one’s parents.

The third experiment provided more direct evidence of the effects of numerical values on the interpretation of the scale end labels. Respondents were asked in an open-end question to provide their interpretation of someone else’ reports given along two different scales. The findings indicated that respondents drew more extreme inferences from reports given on a -5 to 5 scale than from formally identical reports given on a 0 to 10 scale (Schwarz, Knauper, et al., 1991, Experiment 3).

In a fourth experiment, Schwarz and Hippler (1995) compared the effects of numerical values in two modes of administration – telephone interviews and mail questionnaires. Respondents were asked to evaluate politicians along an 11-point scale, running from 0 to 10 or from -5 to 5, in both modes of administration. In both modes, they found the usual mean shift to the higher end of the scale with the -5 to 5 scale (Schwarz & Hippler, 1995).

O’Muircheartaigh and his colleagues replicated the work of Schwarz, Knauper, et al. (1991). They demonstrated that numerical values induced a mean shift in responses whether or not the numerical values assigned to the rating scales were explicitly mentioned in the question stem and whether the scale labels were bipolar or unipolar (O’Muircheartaigh, Gaskell, & Wright, 1995).

The prior studies all used rating scales. Schwarz and his colleagues also investigated the effects of different numerical values attached to a frequency scale (Schwarz, Grayson, & Knauper, 1998, Experiment 1). The frequency scale ranged either from 0 to 10 or from 1 to 11. The end labels remained “rarely” for 0 or 1 and “often” for 10 or 11. Again, the numerical values of the scales influenced the responses – respondents reported higher frequencies when the scale ranged from 0 to 10 than when it ranged from 1 to 11. Schwarz and colleagues speculated that the end label “rarely” indicated a lower frequency when combined with value 0 than with value 1; as a result, the scale running from 0 to 10 shifted the means to the higher end of the frequency (Schwarz et al., 1998).

Table 4.1. Experimental Conditions of Studies on Numerical Values of Rating Scales

Study	Mode of Administration	Polarity of End Label	Domain of Content	Special feature
Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991, Experiment 1)	Face-to-Face/Show Card	Unipolar (“not at all successful”-“extremely successful”)	Success in life	--
Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991, Experiment 2)	Self-Administered Questionnaire	Bipolar (“unsuccessful”-“successful”) vs. Unipolar (“not so successful”-“very successful”)	-Success in life -Happiness of childhood	Self-reporting vs. Proxy reporting
Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991, Experiment 3)	Self-Administered Questionnaire	Bipolar (“dissatisfied”-“very satisfied”) Unipolar (“not so successful”-“successful”)	-Health satisfaction -Success on academic exams	Evaluation of third party
Schwarz and Hippler (1995)	Telephone Interview vs. Mail Questionnaires	Bipolar (“don’t think very highly of this politician” – “think very highly of this politician”)	Evaluation of politicians	--
O’Muircheartaigh, Gaskell, and Wright (1995, Experiment 1)	Face-to-face/Show Card	Bipolar (“much more entertaining than the programmes” – “much less entertaining than the programmes”)	Evaluation of advertisements	Explicit mentioning scale in question stem vs. not
O’Muircheartaigh, Gaskell, and Wright (1995, Experiment 2)	Face-to-face/Show Card	Bipolar (“given much less power” – “given much more power”) vs. Unipolar (“not given any more power” – “given much more power”)	Evaluation of the Advertising Standards Authority	--
Schwarz et al. (1998, Experiment 1)	SAQ	Unipolar (“rarely” – “often”)	Frequency of behaviors	Scales of 0 to 10 vs. 1 to 11

Across the various studies in Table 4.1, the numerical values assigned to scale points consistently affected survey responses. The effect is found across modes of administration (telephone interview vs. mail surveys vs. face-to-face interviews), for both unipolar and bipolar scales, for various domains, and for self and proxy reports. One limitation of the studies is that only one (out of the eight reported) sought direct evidence that the response changes induced by the numerical values were due to respondents' utilization of the maxim of relation during the survey response process.

My first experiment in this web study aimed to fill this gap, seeking further direct evidence that respondents use the maxim of relation when answering rating scales questions with numerical labels. Unlike the previous studies, this one attempted to create conditions in which the maxim of relation wouldn't apply. The numerical values assigned to scale points were displayed in a distinct font that was much fainter than the font used for the question text and the verbal label. Such fonts are typically used in paper questionnaires for information that is *not* intended for the respondents. I thought that presenting the numerical values in a faint font might lead respondents to discount their relevance.

I hypothesized that the mean shift associated with negative scale values would be reduced when the numbers were in the faint font, and the inferences about the meaning of the end labels associated with different numerical values would converge in the faint font condition.

4.1.2 Method

Overview. This experiment, together with three other experiments, was embedded in a web survey conducted by MS Interactive. Survey Sampling Inc. (SSI) selected the sample for this study from its opt-in Web panel (Survey Spot) of over one million persons who have signed up online to receive survey invitations. SSI selected 17,362 e-mail addresses for this study and sent out e-mail messages inviting the recipients to take part in “a study of attitudes and lifestyles.” The e-mail invitations included the web address (URL) for the survey web site and a unique identification number (which prevented respondents from completing the survey more than once). The survey ran from May 24 to June 2, 2005. Of the 17,362 invited to participate in the survey, 1,071 completed the entire survey (and 146 others got part way through) for a response rate (AAPOR [2000] RR1) of 6%. The questionnaire included questions on a range of topics, most of them attitudinal. The 18-item need for cognition scale (Cacioppo, Petty, and Kao, 1984) was included in the last section, together with demographic questions. This experiment came first in the questionnaire.

Experimental manipulation. This experiment manipulated both the numerical values assigned to the scale points (replicating the earlier studies) and the appearance of the scale values in a 2 (numerical values: 0 to 6 vs. -3 to 3) x 2 (appearance: normal font vs. faint font) factorial design. The faint font version of the scales shaded the numerical values for each scale point so that the numbers were legible, but not as obvious and distinct as other texts in the same screen (see Figure 4.1). Table 4.2 displays the number of completes per experimental condition.

Table 4.2. Experiment 4: Number of Completes Per Experimental Condition

	0 to 6	-3 to 3	Total
Normal Font	259	271	530
Faint Font	271	270	541
Total	530	542	1071

Target questions. Respondents were asked to rate their success in life, their moodiness, their nervousness, and optimism along one of the four randomly assigned scales. Respondents got the same numerical labels for all of the target items.

Follow-up questions. The follow-up questions asked respondents about their use of the scale values and the inferences they drew about the scale end labels. The exact wordings of the target questions and follow-up questions are given in Table 4.3.

Figure 4.1. Example of a Faded Scale

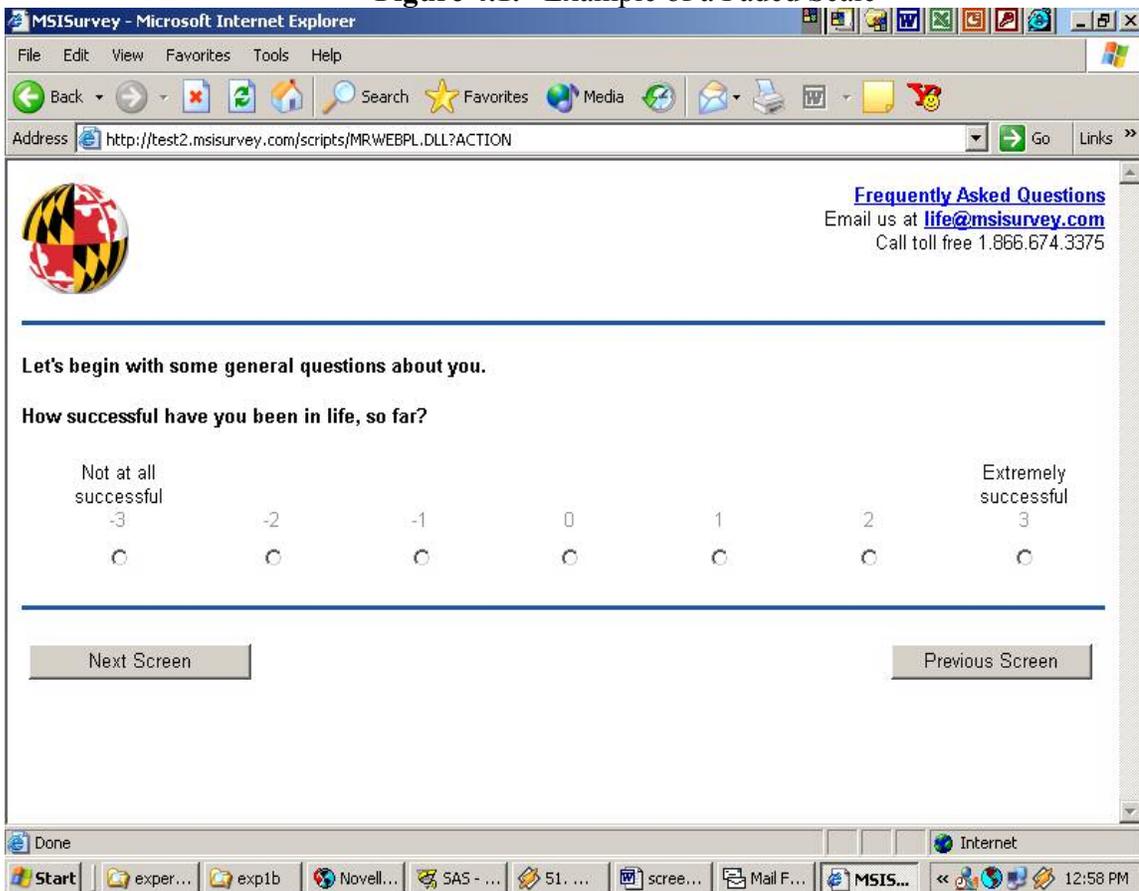


Table 4.3. Questions Used in Experiment 4 on Maxim of Relation

Target Questions:
How successful have you been in life, so far?
Overall, how moody would you say you are?
In general, how nervous do you think you are?
How optimistic would you say you are?
Follow-up Questions:
To your best recollection, what numerical value do you remember was assigned to the leftmost scale point (the starting value) in the scale you have used for the last four questions?
How much attention did you pay to the numerical values of the scales when you answered the questions?
How useful did you find the numerical values in helping you answer the questions?
Now let us focus on one particular question you answered just now. That is, ‘Overall, how successful would you say that you have been in your life?’ What do you think the scale label ‘Not at all successful’ means to most people like you? (OPEN-END)
Which one of these options comes closest to the meaning of “Not at all successful?”
1 Modest Accomplishment
2 Little accomplishment
3 Absence of Significant Accomplishment
4 Little failure
5 Modest failure
6 Utter failure

4.1.3 Results

I begin by presenting the analyses of responses to the four target questions, followed by analyses on respondents’ use of and inferences about the scale.

Responses. For all four scale conditions, I coded the responses from 1 to 7, where 1 corresponded either to 0 or -3, and 7 to 6 or 3. To compare responses to scales with different numerical values, I examined the mean ratings of the 0 to 6 scales and of the -3 to 3 condition (see Table 4.4). Two-way ANOVAs were conducted on all four target questions, showing that numerical values had a significant effect on responses for three out of the four items (see Table 4.5. for ANOVA results).

Table 4.4. Experiment 4: Mean Ratings By Experimental Condition

	Normal Font		Faint Font	
	0 to 6	-3 to 3	0 to 6	-3 to 3
Successful	4.92	5.16	4.79	5.09
Moody	3.53	3.75	3.55	3.83
Nervous	3.17	3.47	3.32	3.40
Optimistic	5.02	5.24	5.11	5.06
Average of last 2 items	4.10	4.35	4.22	4.23

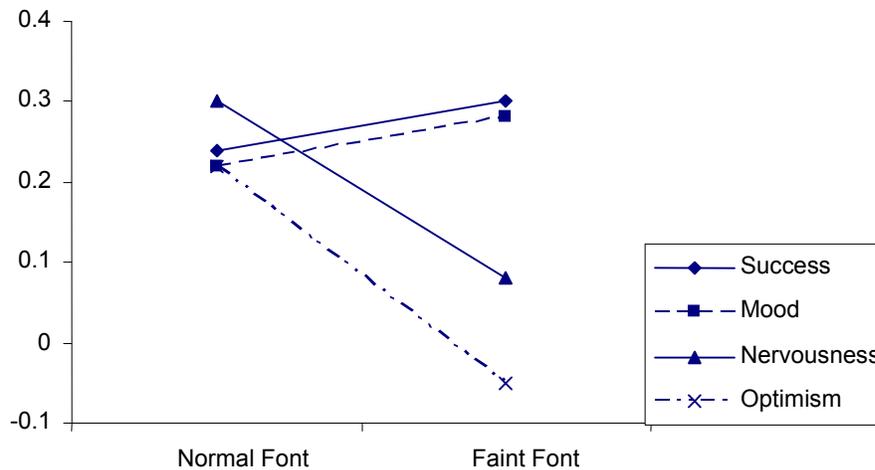
Table 4.5. Experiment 4: Two-way ANOVA Results

	Scale Values	Scale Font	Interaction (Values x Font)
Success	$F(1,1059)=12.89$ $p<.001$	$F(1,1059)=1.8$ ns	$F(1,1059)=.13$ ns
Mood	$F(1,1062)=7.13$ $p<.01$	$F(1,1062)=.26$ ns	$F(1,1062)=.12$ ns
Nervousness	$F(1,1059)=3.95$ $p=.05$	$F(1,1059)=.19$ ns	$F(1,1059)=1.45$ ns
Optimism	$F(1,1063)=1.02$ ns	$F(1,1063)=.27$ ns	$F(1,1063)=2.83$ $p<.10$
Average of last 2 items	$F(1,1055)=5.76$ $p=.02$	$F(1,1055)=0$ ns	$F(1,1055)=4.77$ $p=.03$

There was some evidence that the effect of the numerical labels was dampened when the faint font was used. I calculated the difference between the average rating for the two scales for each item and font condition. Positive numbers indicated that the negative scale values induced a mean shift in ratings to the higher end of the scale.

Figure 4.2 presents the results.

Figure 4.2. Mean Shifts Due to Negative Numbers by Their Font Across 4 Items



As evident from the figure, the negative scale values produced a mean shift in ratings in almost all situations – this is apparent from the positive values in all conditions except for the optimism item in the faint font condition. Furthermore, the mean shifts caused by the negative scale values were clearly reduced for two of the items (nervousness and optimism) when the numerical labels were in the lighter font. Taking an average of the last two items as the dependent variable, I ran another two-way ANOVA, which revealed a significant interaction effect between the numerical values and the font (see last row in Table 4.5). Further analysis showed that, for this average, the simple main effect of the numerical values was significant only when the numbers were in the normal font ($F(1,1055)=10.40, p=.001$), but not when they were in the lighter font ($F(1,1055)=.02, p=.88$). A planned contrast was also significant, showing that the faint font dampened the effects of negative scale values. The faint font seemed to prevent respondents from perceiving the numerical values as relevant for these two items.

Inferences. I assessed the inferences respondents drew based on their answers to follow-up questions. According to Grice, conversational implicatures are worked out when maxims are flouted but the CP is still presumably being observed (Grice, 1989). The extra effort needed to work out an implicature should produce better recall of the numbers that triggered the interpretative maxim. In addition, respondents should have paid more attention to the numbers and considered the numerical values more useful when they used them in interpreting the response scale.

I examined the percentage of respondents who recalled the leftmost scale value correctly by the scale values and the font of the numbers. Figure 4.3 indicates that more respondents recalled the number correctly when presented with the -3 to 3 scale labels

than with the 0 to 6 labels. A two-way logit analysis revealed a marginally significant main effect of numerical values on the percentage of respondents who accurately recalled the numbers ($\chi^2=3.18, p=0.07$). Neither the main effect of the font ($\chi^2= 1.27, p= 0.26$) nor the interaction between the numbers and fonts ($\chi^2= 0.57, p=0.45$) were significant. Still, the pattern is consistent with the prediction – the negative values were recalled better overall and the difference in recall is less marked when the font was lighter.

Figure 4.3. Percent Correctly Recalling Numerical End Point by Scale Condition

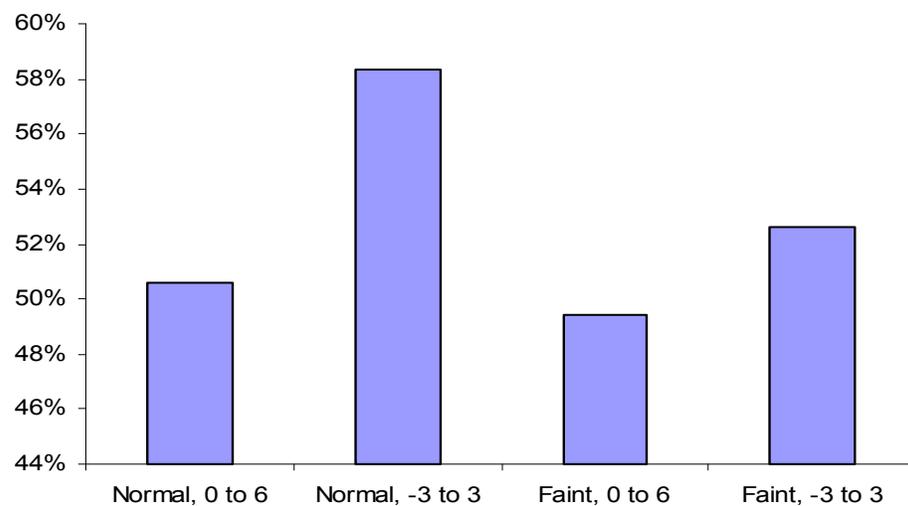
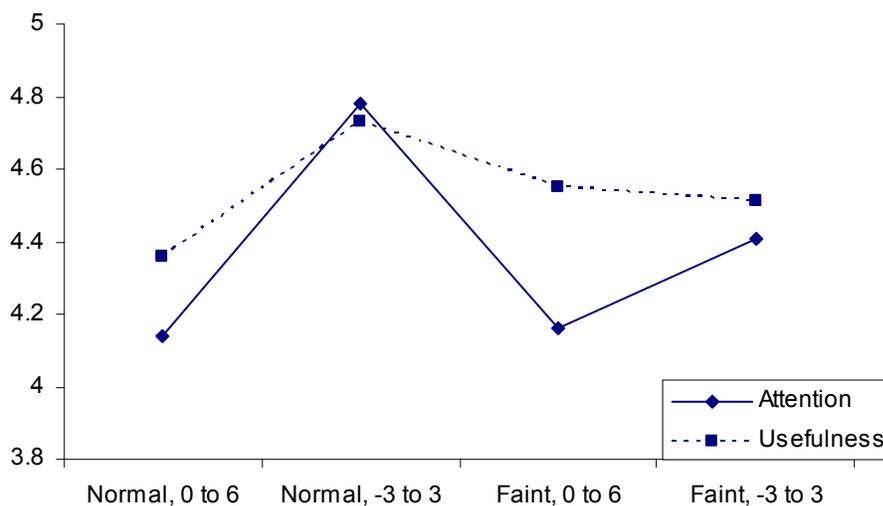


Figure 4.4. Mean Ratings of ‘Attention’ and ‘Usefulness’ by Scale Condition



Another two follow-up items asked respondents how much attention they paid to the numerical values attached to the scale and how useful they considered those numbers. Figure 4.4 plots the mean ratings of attention and usefulness across the four scale conditions. Higher numbers indicate more attention and higher usefulness ratings

Figure 4.4 shows that the numerical labels seemed to play an important role in the attention ratings; respondents tended to pay more attention when the scale started with a negative number (-3) than with zero ($F(1,1066)=15.65, p<.0001$), which confirmed my prediction. There is also a marginally significant interaction effect of the numerical values with the font ($F(1,1066)= 2.94, p<.09$), indicating that the faint font reduced the effect of the numbers – respondents paid less attention on average to the negative end labels when it was in the lighter font.

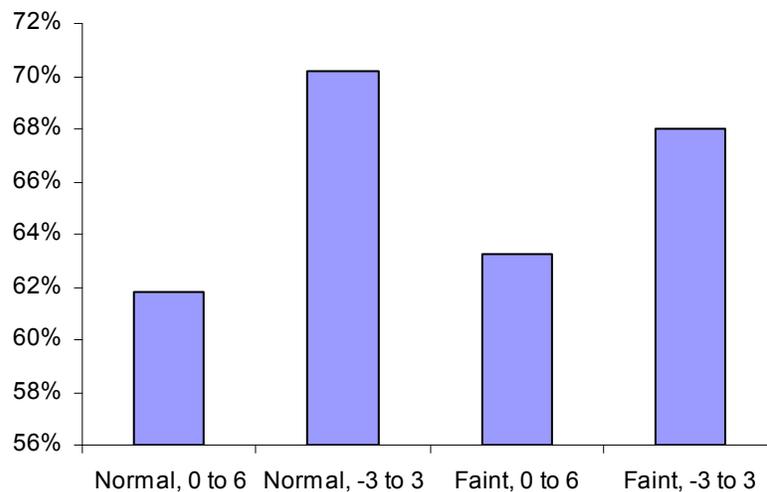
As for the usefulness rating, neither the numerical values ($F(1,1059)=2.19, p< 0.14$) nor the font ($F(1,1059)=.02, p<.89$) had significant main effects. The interaction effect is marginally significant at $p<.07$ ($F(1,1059)=3.30$). With the fainter fonts, respondents rated the numerical labels equally useful.

To determine what respondents inferred from the numerical values, the last follow-up question asked respondents what the scale label “not at all successful” meant to them.⁷ There were six answer categories to this question (see bottom panel of Table 4.3 for the exact wordings of the six answer categories). I collapsed the answer categories into two groups – one group represents the absence of success and the other group the presence of failure. Figure 4.5 plots the percentage of people inferring “presence of failure” by scale condition.

⁷ I present here only the results based on the closed-end question presented above. Analyses of the open-ended responses were similar and didn't change the conclusions reported here.

The result supports the conjecture of Schwarz and his colleagues (1991) about how respondents interpret the scales with different numerical labels. Significantly more respondents interpreted the scale label “not at all successful” to mean the presence of failure when the numerical labels ran from -3 to 3 than when they ran from 0 to 10. The logit analysis revealed a significant main effect of scale numerical values ($\chi^2=5.13$, $p=.02$), but no effect of scale appearance and no interaction between the two variables. Respondents did take the numerical values into consideration when they constructed their answers. Once again, the effect of the label seems reduced when the numbers were presented in the lighter font (see Figure 4.5 below) but the relevant interaction is not significant.

Figure 4.5. Percentage of Respondents Inferring “Presence of Failure” by Scale Condition



Response times. I also examined the time respondents took to complete the four target questions. On average, respondents took longer to complete the block when given a scale running from -3 to 3 (mean response time=54.1 seconds) than when presented a scale running from 0 to 6 (mean response time=49.8). The difference is marginally

significant ($F(1,1067)=3.49, p=.06$), suggesting again that respondents did notice the negative values and took them into account in forming their responses. The font main effect and the interaction between the numerical values and the font were not significant.

4.1.4 Conclusions

This experiment replicated the findings of Schwarz, Knauper, and colleagues (1991) and demonstrated that the numerical values of the scales affected survey responses. Scale appearance (i.e., the font of the numerical values) moderated the effects of the numerical values, especially for the last two of the four questions. Further research is needed to investigate under what circumstances factors such as scale appearance are able to reduce or eliminate the perceived relevance of the numerical values to the meaning of the scale.

Follow-up questions supported the view that respondents applied the maxim of relation when they encountered a scale with negative values and that they generated different inferences based on the scale numbers (Schwarz, Knauper, et al., 1991). When respondents were presented a scale with negative numerical labels, they took more time to respond, were significantly more likely to say they paid attention to the numerical values, and were (nonsignificantly) more likely to say they considered the numbers useful. They were also able to correctly recall the numerical value assigned to the leftmost scale point more often, and more likely to infer that the scale label “not at all successful” associated with -3 meant “presence of failure.” Thus, this experiment presented a variety of evidence that the CP and, in particular, the maxim of relation were being applied. Unfortunately, it seems that the effect of the numerical labels was quite robust. Even

when the numbers were presented in a distinctive lighter font, they still had an effect on respondents' answers and their inferences about the meaning of the response scale.

4.2 The Maxim of Quantity and Similar Items

4.2.1 Introduction

The maxim of quantity concerns the informativeness of an utterance in a context. It requires speakers to make their contribution as informative as is required, but no more informative. In survey research settings, this maxim invites respondents to provide all the information the survey researcher seems interested in, but not other information that may come to mind. On the other hand, it also discourages providing information that has already been given earlier or information that “goes without saying” because, as a rule, speakers should not burden hearers with information they are already likely to know (Hilton, 1995; Schwarz, 1996, 2000; Schwarz & Oyserman, 2001). Thus, for utterances to conform to the maxim of quantity, speakers (both researchers and respondents) must be able to provide all that their addressees want to know without being redundant.

A few studies have demonstrated situations where survey respondents appeared to apply the maxim of quantity in answering seemingly repetitive questions. One set of studies asked respondents a specific question (e.g., happiness with marriage or romantic life) and then a more general question (e.g., happiness with life as a whole). The correlation between the two questions was found to be lower when the questions were administered in the specific-general order than the general-specific (cf. Kalton, Collins, & Brook, 1978; Mason, Carlson, & Tourangeau, 1994; McCabe & Brannon, 2004; Schwarz, Strack, & Mai, 1991; Strack, Martin, & Schwarz, 1988; Tourangeau, Rasinski, &

Bradburn, 1991). The most common explanation is that survey respondents applied the maxim of quantity in answering the items. Respondents interpreted the general question as excluding the domain mentioned in the specific question and took the general item to mean something like “aside from your marriage” (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991). Factors that prevented respondents from applying the maxim in this way included increasing the number of specific questions, rephrasing the general question to include the specific items, and placing the general and specific question in different conversational contexts (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991).

Another example of apparently redundant questions occurs in a study by Strack and his colleagues (Strack, Schwarz, & Wanke, 1991). They asked respondents to report both their happiness and their satisfaction with life as a whole. They varied the introduction to the two questions and the appearance of the questions, so that, in one condition, the two questions were placed in the same conversational context, whereas in the other, the two questions belonged to two different contexts (because they were in different questionnaires). Strack and colleagues observed a lower correlation between the two items when they were in the same conversational context than when they were in different contexts (Strack et al., 1991). They attributed the difference to respondents’ application of the maxim of quantity when the two items were in the same conversational context. Two highly similar questions in the same conversational context appear repetitive in that answers to them would have been quite alike; to uphold the Cooperative Principle, respondents inferred that there must be differences – minute or not – between

the two apparently similar items and responded with that fine distinction in mind (Strack et al., 1991).

I carried out two experiments (Experiments 4 and 5) to clarify when respondents apply the maxim of quantity and when they do not. The experiments varied the introductions, the wording, and the total number of related items to determine which factors govern inferences about redundancy.

An introduction usually signals the start of a new block of questions and the beginning of a new conversational context. If the introduction gives respondents permission to report redundant information by explaining why two similar questions appear in the same conversational context, then a higher correlation between the two items should be observed – the introduction makes it clear that the maxim of quantity is no longer applicable.

The perceived redundancy of two highly similar items comes mainly from semantic overlap between the concepts, but the appearance of redundancy may be heightened by syntactical similarity. If Gricean implicatures are triggered by conceptual overlap alone, then a superficial change, such as reversed wording, shouldn't change the correlation between the two items. However, if syntactic similarity between the items also contributes, then the magnitude of correlation should be different when an item is reverse worded.

Two similar items are found to trigger the maxim of quantity (Strack et al., 1991). However, increasing the number of similar items to four might actually boost the correlation between the same two items because it is hard to apply the maxim of quantity when there are multiple items and to make distinctions among them.

4.2.2 Method

Experiments 4 and 5 were embedded in the same web survey as Experiment 3. Experiment 4 was the second experiment overall starting from question 9 in the web questionnaire; Experiment 5 was the fourth one in the questionnaire, beginning at question 20. Both experiments employ a 2 x 2 factorial design, with independent randomizations.

Experiment 4 design. Experiment 4 varied the introduction and question wording independently. The introduction used in the study by Strack and colleagues (1991) was used to create a conversational context for the two questions that followed: “Now we have two questions about your life.” I refer to this as the “general” version of the introduction.

The second version (the “specific” version) highlighted the overlap in meaning across the two questions: “The next two questions are to measure your outlook on life.” By highlighting their shared purpose, the specific introduction would, I thought, give respondents permission to provide redundant information.

To test for the effect of syntactic similarity, I varied the wording of one of the pair target questions. One half of the respondents got the normal wording:

“How happy would you say you are with your life as a whole?”
“How satisfied would you say you are with your life?”

The other half got the reversed wording:

“How unhappy would you say you are with your life as a whole?”
“How satisfied would you say you are with your life?”

The number of completed interviews by experimental condition for experiment 4 (and experiment 5) is shown in Table 4.6.

Table 4.6. Experiments 4 and 5: Number of Completes per Experimental Condition

	General Introduction	Specific Introduction	Total
Experiment 4			
Normal Wording	276	266	542
Reversed Wording	257	272	529
Total	533	538	1071
Experiment 5			
2 Items	268	260	528
4 Items	252	291	543
Total	520	551	1071

Design of experiment 5. Experiment 5 again varied the introduction to the target items in the same way as Experiment 4. The second factor varied the number of target items included in the block. Half of the respondents got two similar items. The other half got the same two items plus two new items about the same topic. All four questions were taken from a scale assessing anxiety. The items are displayed in Table 4.7. The number of completes in each experimental cell is given in the bottom panel of Table 4.6.

Table 4.7. Items for Experiment 5

Two-item	Four-item
How easily do you get upset?	How easily do you get disturbed? How easily do you get stressed out?
How easily do you get irritated?	How easily do you get upset? How easily do you get irritated?

Target questions. The main target questions for Experiment 4 are the two questions asking respondents to rate their happiness and satisfaction with life. The target questions for Experiment 5 are the two common items, asking respondents about how easily they get upset and irritated.

Follow-up questions. The same set of follow-up questions was used in both experiments:

- “Think about the last two questions you have just answered. How repetitive did you think these two questions were?”
- “To what extent did meanings of these two questions differ?”

- “Based on your understanding of the two questions, did you think they were supposed to measure the same thing or different things?”
- “Now thinking about how you came up with answers for the last two questions, did you think about the same things or different things when you answered these two questions?”

4.2.3 Results

The analyses begin with an examination of the responses to the target items, with a focus on the correlations between the two target items in both experiments.⁸ The second part of the analysis examines the inferences respondents drew.

Correlations. Table 4.8 displays the correlation coefficients by experimental conditions for Experiment 4. The reverse-worded items were recoded so that they are in the same direction as the normal wording items with higher numbers indicating higher happiness. It is clear from the table that the introductions did not have a significant effect on the correlation coefficients ($F(1,\infty)=.85, p=.36$), regardless of the wording condition. This was contrary to the original hypothesis that the specific introduction would discourage respondents from applying the maxim of quantity since it highlighted the purpose of the redundancy. The size of correlations Experiment 4 produced across introductions fall between those observed in the same conversational context ($r=.75$) in the study by Strack et al. (1991) and those obtained in the different conversational context ($r=.96$). Thus, it looked like both versions of the introduction seemed to trigger the maxim of quantity, but the effect was not as strong as that observed by Strack and colleagues. The specific introduction, whose wording was probably too similar to that of the general introduction, was not successful at halting the maxim of quantity.

⁸ I also checked whether the means of the answers varied as a function of the experimental factors. Reverse wording significantly affected the mean responses to the two target items for Experiment 4 whereas the introduction variable interacted with the number of items on the mean responses for Experiment 5.

Reversing the wording of the happiness item had a large impact on the correlation; it lowered the correlation between the pair target items significantly. In other words, respondents seemed to have differentiated between the target pair more when one of the items was reversely worded⁹.

Table 4.8. Experiment 4: Correlation Coefficients by Experimental Condition

Experiment 4 (Happiness and Satisfaction)	
General introduction – Normal Wording	0.87
Specific introduction – Normal Wording	0.85
General introduction – Reversed Wording	0.51
Specific introduction – Reversed Wording	0.47

The results from Experiment 5 displayed a different picture from those of Experiment 4 (see Table 4.9). First, the number of items had a significant effect on the correlations obtained. Consistent with my hypothesis, the correlation between the two key items was significantly higher when they were asked with two other items about the same topic than when they were asked by themselves ($F(1,\infty)=9.83, p<.01$). Respondents appeared to have stopped trying to differentiate among the items where there were four items bundled together. It is likely that at this point the maxim of relation was at work rather than the maxim of quantity. This is consistent to the part-whole literature; when the number of specific questions was increased to three, respondents took the general question as asking for an average or sum over the three specific questions rather than as excluding the specific question when there was only one specific question preceding the general one (Schwarz, Strack, & Mai, 1991). In addition, the introduction had a

⁹ An alternative explanation for the low correlation is that “unhappy” and “happy” are not direct opposites; thus, asking about unhappiness is not exactly the opposite to asking about happiness. This explanation is plausible. However, responses to a later follow-up question seemed to rule out this alternative. More respondents considered the target pair as measuring the same thing when “unhappy” was used in place of “happy.” And this inference seemed to have fit in the “avoid redundancy” hypothesis (see Inference section for more detailed discussion).

marginally significant interaction with the number of items ($F(1, \infty)=3.16, p<.10$). The specific introduction reduced the correlation when there were only two items (though this was not significant), but significantly increased the correlations between the two key items when two other related questions preceded them ($Z=-2.07, p=.04$).

Table 4.9. Experiment 5: Correlation Coefficients by Experimental Conditions

Experiment 5 (Getting upset and Getting irritated)	
General introduction – 2 items	0.70
General introduction – 4 items	0.74
Specific introduction – 2 items	0.67
Specific introduction – 4 items	0.81

Inferences. The first two follow-up questions were intended to assess the perceived redundancy of the target questions. They asked respondents to rate how repetitive and how different the target items seemed on a seven-point scale. In both experiments, the responses to these two follow-up questions were highly correlated – the coefficients in the high .50s. Thus, I combined responses to the two questions into a redundancy index, with higher numbers indicating higher perceived redundancy.

For the two target questions in Experiment 4, neither the introduction nor the wording affected respondents’ perceptions of how redundant the two target questions were. The two experimental variables didn’t significantly interact either. In Experiment 5, though, the introduction had a marginally significant effect on respondents’ perception of redundancy. Respondents tended to rate the two target items to be more redundant when they had been given the general introduction (mean rating=3.60) than the specific introduction (3.43) ($F(1,1059)=3.07, p<.10$). Increasing the number of items also increased the average redundancy rating. The mean redundancy rating was 3.42 when the two target questions were asked by themselves, but it went up to 3.60 when the two

target questions followed two similar items also assessing anxiety. The difference was only marginally significant, however ($F(1,1059)=3.43, p<.10$).

The last two follow-up questions asked respondents whether they thought the target items were intended to measure the same thing or not, and whether they used the same experiences to answer the two similar questions or not. For experiment 4, logit analysis showed more respondents (48%) thought the target questions were supposed to measure the same thing when one target item was reverse worded than when the two items were worded in the same direction (36%). The difference was significant ($\chi^2=5.09, p<.03$). In other words, reverse wording didn't break the conceptual overlap. This inference explained the lower correlations I obtained in Experiment 4; the inference that the two items were supposed to measure the same thing might have led respondents to further distinguish between the items when they form answers to the questions; thus, bringing down the correlations.

For experiment 5, there was no significant effect of either experimental variable on whether respondents thought the target items were intended to measure the same thing or not. In addition, whether respondents reported using the same experience or different experiences in forming their answers to the target items was not affected by any of the experimental factors or their interactions.

4.2.4 Conclusions

These two experiments attempted to replicate and extend the work by Strack and colleagues (1991) on how respondents deal with apparently redundant questions. The Gricean tendency to 'avoid redundancy' seemed to be robust across introductions in

Experiment 4. Respondents seemed to have differentiated between the two similar items whichever introduction they were given, but the effect was not so strong as those observed in Strack et al. (1991). The introduction variable, however, did interact with the number of items in affecting correlations in the second experiment (Experiment 5, Table 4.9), though the effect was only marginally significant.

The effects of reverse wording on responses and inferences were surprising. When I reversed the wording of one of the two target items, it significantly reduced the correlation between them, because more respondents believed the two items were supposed to measure the same thing. Reverse wording had no impact on ratings of the redundancy of the two items.

Adding two more items boosted the correlations between the two target items and it also caused people to regard them as significantly more redundant.

These two experiments demonstrated that respondents do not necessarily apply the maxim of quantity when answering similar questions. Some factors (e.g., the number of similar items to be asked at the same time) seem to suppress the maxim of quantity; others (e.g., reversing the wording of one of the items) seem to strengthen the need for the maxim of quantity. Sometimes, factors interact with each other (e.g., introduction and the number of items).

4.3 The Maxim of Quality and Presuppositions

4.3.1 Introduction

The maxim of quality concerns the likely truth value of an utterance. A cooperative communicator is expected to speak the truth rather than telling lies (Grice,

1989). As a result, a hearer usually considers the probable truth value of an utterance to be high unless he or she has legitimate reasons to doubt the speaker's sincerity, reliability, or knowledge. This maxim applies not only to what is said, but also to what is presupposed by the utterance.

A presupposition can be regarded as simply any condition that the speaker and the hearer normally assume to hold for an utterance to be contextually appropriate (Stalnaker, 2002). Linguists have taken a number of different views on presuppositions. One perspective is the conversational view proposed by Stalnaker (Stalnaker, 1974, 2002; see also Simons, 2001, 2003). According to Stalnaker, presupposition is primarily a property of speakers and listeners, not a formal property of sentences. A speaker's presuppositions are, roughly, those propositions which he/she believes to constitute the accepted background information for the conversation in which he/she is engaged. Stalnaker thus links presupposition to the common ground the speaker and listener share (Stalnaker, 2002). Presupposition is treated as a restriction on the common ground – the set of propositions constituting the current context. The failure or non-satisfaction of the presupposition makes a given utterance inappropriate in a given context.

Since speakers and listeners have their own set of beliefs about which presuppositions are in the common ground, ideally the presuppositions of the speaker should match those of the listener, producing a nondefective context (Stalnaker, 2002). However, when the hearers detect a discrepancy between their presupposition and those of the speakers, they may do one of several things. If the listeners consider the speaker unreliable, they are most likely to respond with an explicit rejection of the presupposition. But if the listeners consider the speaker reliable on this point, they might add the

presupposition to the set of propositions in the common ground. Or if the listeners consider the speaker unreliable with respect to the proposition but don't have any interest in challenging its truth, they might merely decide to go along with the speaker's presupposition.

In a survey research setting, survey respondents cannot usually express their rejection of the presupposition conveyed by questions unless the researcher provides a response option (e.g., "Does not apply"). But if respondents choose to accept the researcher's presupposition and add it to the common ground, their subsequent responses can be affected.

There is evidence showing that presuppositions carried by questions can modify respondents' memory about past events. For example, Lipscomb, McAllister, and Bregman (1985) looked at the effect of using unmarked modifiers (versus marked modifiers) on respondents' numerical estimates in eyewitness reports. The unmarked modifier of a dimension has a nominal or neutral use that refers to the whole dimension whereas the marked modifier designates the absence of such a property and has a lower bound of zero (Harris, 1973). Thus, as the unmarked modifiers such as "how long" or "how heavy" presuppose no upper limit with regard to the actual length/weight, respondents tended to provide higher numerical estimates as compared to the use of marked modifiers such as "how short" and "how light" (Lipscomb, McAllister, & Bregman, 1985; see also Harris, 1973).

In several experiments, Loftus demonstrated that presuppositions affected answers to subsequent questions, even to questions asked some time afterwards (Loftus & Palmer, 1974; Loftus, 1975). For example, Loftus and Palmer (1974) showed that the

question, “About how fast were the cars going when they smashed into each other?” consistently elicited a higher estimate of speed than when “smashed” was replaced by “collided,” “bumped,” “contacted,” or “hit.” The differences in the estimates of speed reflected the process of including the extra information presupposed by the different verbs into the common ground (see also Loftus, 1975).

Survey researchers have documented the influence of presuppositions on survey responses for at least two decades. Schuman and Presser (1981), for instance, observed a large percentage of respondents offered opinions on obscure issues; the act of asking a question presupposes that the issue exists and is important. This presupposition apparently put pressure on respondents to give opinions about nonexistent topics or ones they were unfamiliar with (see also Bishop, Tuchfarber, & Oldendick, 1986, Bishop, Oldendick, & Tuchfarber, 1983). More respondents opted for a “Don’t Know” or “No Opinion” option when they were presented first with a filter question that asked whether they know anything about the topic; the filter carried the implication that not knowing was a possible response.

In a similar way, direct questions such as “How many times did you do X?” or “How concerned are you with Y?” presuppose that one did X or that one should be concerned about Y, which may or may not be appropriate. Nonetheless, assuming that the survey designers are cooperative and reliable communicators, the respondents may take the presuppositions to be true and base their responses on this premise or reinterpret the questions to fit the presuppositions. For instance, Knauper (1998) showed that questions asking respondents “In the past 10 years, how many times did you witness a crime?” elicited significantly more reports of witnessing crimes, and fewer reports of no

instances. Knauper (1998) argued that, because the question “how many times did you witness a crime?” presupposes that one has witnessed (at least) a crime, respondents inferred the category “crime” included less serious incidents. A filtered version of the question that asked first whether respondents had witnessed a crime, on the other hand, did not carry this presupposition. Respondents reported fewer (and more serious) crimes. Sterngold, Warland, and Herrmann reported similar findings (1994). They showed that the direct degree-of-concern questions (“How concerned are you with ___?”) might be leading questions, encouraging respondents to overstate their concerns. By contrast, the use of filter questions (“Are you concerned with ___ or not?”) produced a higher percentage of “not concerned” responses and fewer responses at the upper end of the response scale (Sterngold, Warland, & Herrmann, 1994). Thus, filter questions seem to eliminate the presupposition that one should be concerned.

My final experiment (Experiment 6) examined the impact of presuppositions on people’s opinion about general political issues. Since the issues used in this experiment (such as environment and agriculture) are vaguely defined and broad in scope, respondents were left to decide which specific aspects of the issues they were being asked to evaluate. This experiment examined two possible methods to foster a presupposition of importance. Since asking a question itself presupposes that an issue exists and is important, respondents will consider an issue more important when a previous question asks about it than when this issue is not mentioned in the earlier item. Furthermore, respondents will give a higher importance rating to an issue when they are asked a direct question about it, such as “How important is ___?” than when they were first asked a filter question. The hypothesis is that respondents will adopt the presupposition of importance

when a direct question is asked or when the same issue is mentioned in an earlier question and will express higher concern about the issue.

4.3.2 Method

This experiment was included in the same web survey as Experiments 3, 4, and 5.

Experimental variables. The first factor was whether or not the issue appeared in a previous block of questions. Respondents were randomly assigned to one of two blocks of prior questions, asking them to evaluate the level of government spending on various programs. Table 4.10 gives the exact wording of the two blocks of questions.

Respondents were then asked to judge the importance of four target issues – two of which came from each block – and to rate their concern about each of these issues.

Table 4.10. Experiment 6: Questions in the Spending Block

There are many problems in this country, none of which can be solved easily or inexpensively. For each of the following problems, please tell me whether you think we're spending too much money on it, too little money, or about the right amount.	
Block A	Block B
Space exploration	The environment
Highways and bridges	Parks and recreation
Agriculture	Medical research
Health	Welfare

Note: The issues in bold were the four target issues whose importance respondents subsequently rated.

The level-of-spending questions carry the presupposition that the issues are important ones; as a result, respondents might give a higher importance rating to these issues. By contrast, when the target issue was not included among those in the spending questions, the absence of the presupposition would lead to a lower importance rating.

The second factor varied how the importance questions were asked. Questions were either asked in a direct way (“How important do you think the environmental issues are to the country?”) or a filter was used (Do you think the environmental issues are important to the country or not?). The direct question had five response categories (“very important,” “somewhat important,” “neither important nor unimportant,” “somewhat unimportant,” or “very unimportant”). The filter question offered three response options (“important,” “neither important nor unimportant,” and “unimportant”). If respondents chose either “important” or “unimportant,” a second question asked them to indicate the extent of importance (or unimportance) by selecting one of the two response categories (“very important/unimportant,” and “somewhat important/unimportant”). My hypothesis was that respondents would rate an issue as more important when the importance question was posed without a filter. Table 4.11 displays the number of completed interviews by experimental condition.

Table 4.11. Experiment 6: Number of Completes by Experimental Condition

	Block A	Block B	Total
Direct Questions	304	255	559
Filtered Questions	264	248	512
Total	568	503	1071

Target questions. All respondents were asked to express their concern about the four target issues after the importance rating questions (e.g., “How concerned are you about the environmental issues then?”). The concern question had four response options (“very concerned,” “somewhat concerned,” “slightly concerned,” and “not concerned at all”). The concern question was included to assess the impact of presuppositions on survey responses.

4.3.3 Results

The main outcome variables of interest were the responses to the importance and concern questions.

Importance ratings. In this experiment, the key inferences involved the presuppositions carried by the two triggers (i.e., the fact of asking a prior question about the target issue and the direct question format). Responses to the filter version of the importance questions were recoded to the corresponding five response categories as the direct questions. Higher numbers indicate higher importance ratings. To test the hypotheses that respondents would see the issue as more important when it was included in the block of spending questions, and when the importance question was posed without a filter, I examined the mean importance ratings by experimental conditions. Results from two-way ANOVAs showed that the format of importance questions significantly influenced the mean importance rating for three out of the four target issues (see Table 4.12).

Table 4.12 shows that the hypothesis regarding the format of importance questions are not completely supported. The direct question elicited a higher mean importance rating for one of the issues (the environment) than the corresponding filtered question, but the difference was only marginally significant ($p < .08$). The question format made no significant difference at all on the mean importance rating for the issue related to parks and recreation. For issues related to highway and bridges and agriculture,

Table 4.12. Importance of Four Issues by Question Format

	Direct %	Filtered %	Significance Tests
Environmental Issues			
Very Important	52.7	48.1	
Somewhat Important	35.4	33.5	
Neither Important Nor Unimportant	7.0	15.4	
Somewhat Unimportant	3.7	1.6	
Very Unimportant	1.2	1.4	
Total	100.0	100.0	
N	512	559	
Mean	4.35	4.25	$F(1,1067)=3.24, p<.08$
% Choosing Important Options	88.1%	81.6%	$\chi^2=8.49, p=.004$
Issues Related to Highway and Bridges			
Very Important	34.8	52.2	
Somewhat Important	54.5	30.8	
Neither Important Nor Unimportant	7.4	15.2	
Somewhat Unimportant	2.7	1.4	
Very Unimportant	0.6	0.4	
Total	100.0	100.0	
N	512	559	
Mean	4.20	4.33	$F(1,1067)=7.5, p<.01$
% Choosing Important Options	89.3%	83.0%	$\chi^2=6.61, p=.01$
Agricultural Issues			
Very Important	47.4	66.0	
Somewhat Important	38.2	20.9	
Neither Important Nor Unimportant	10.4	12.3	
Somewhat Unimportant	3.2	0.5	
Very Unimportant	0.8	0.2	
Total	100.0	100.0	
N	508	559	
Mean	4.28	4.52	$F(1,1063)=24.3, p<.0001$
% Choosing Important Options	85.6%	86.9%	$\chi^2=0.13, ns$
Issues Related to Parks and Recreation			
Very Important	19.9	30.3	
Somewhat Important	46.9	27.8	
Neither Important Nor Unimportant	21.5	36.0	
Somewhat Unimportant	9.2	4.1	
Very Unimportant	2.5	1.8	
Total	100.0	100.0	
N	512	558	
Mean	3.72	3.81	$F(1,1066)=1.89, ns$
% Choosing Important Options	66.8%	58.1%	$\chi^2=8.54, p=.0004$

the direct questions produced significantly *lower* importance ratings than the filtered versions, which is opposite to the hypothesis. Examination of the univariate distribution of the two importance questions (see Table 4.12) revealed that for these two issues, the filtered version yielded a higher percentage of respondents choosing the “very important” option than the “somewhat important” option, pushing up the mean ratings.

I next collapsed the five-point importance scale so that the upper end of the scale (“very important” and “somewhat important”) constituted one category and the remaining three options a second category. I conducted logit analyses on these dichotomized variables. The results revealed that whether a filter is used or not has significant effect on the percentage of respondents who chose the important end of the scale for three out of four issues (see χ^2 values reported in Table 4.12). Consistent with the finding by Sterngold and colleagues (1994), respondents gave fewer responses at the upper end of the scale when first presented with a filter question than with a direct question.

Whether the issue was asked in the previous block of spending questions had a significant impact on the ‘highway and bridges’ and ‘agriculture’ issues, but not on the other two target issues (see Table 4.13). It is clear from Table 4.13, the hypothesis regarding the block variable was not supported. For environmental issues and issues related to parks and recreation, whether respondents regarded the target issue as important seemed to be independent of whether the target issue was mentioned in the previous block. However, for the other two issues (highway and bridges, and agriculture), the effect of including the issues among the spending questions was significant, but opposite from the predicted direction – when the target issue appeared in the previous block, fewer respondents regarded them as important. Since both issues were included in

the spending questions along with “space exploration” and “health,” it is possible that respondents could have contrasted health and the two target issues (highway and bridges, and agriculture) when evaluating spending on them and this contrast effect may have affected the later importance ratings.

Table 4.13. Mean Importance Ratings (and Percent Selecting Options Above Midpoint) by Whether Issue Included In Prior Block

	Asked in Previous Block	Not asked in Previous Block	<i>F</i> Values	<i>Pr</i> > <i>F</i>
	Mean Rating (% selecting important options)			
Environment	4.29 (84.7%)	4.30 (84.7%)	<i>F</i> (1,1067)=.04	ns
Highway and bridges	4.21 (83.1%)	4.32 (89.3%)	<i>F</i> (1,1067)=5.16	<i>p</i> =.02
Agriculture	4.34 (82.7%)	4.49 (90.4%)	<i>F</i> (1,1063)=9.18	<i>p</i> =.003
Parks and recreation	3.80 (64.1%)	3.74(60.6%)	<i>F</i> (1,1066)=.87	ns

There were no significant interaction effects for the two manipulations for any of the issues.

Responses to concern questions. Respondents were asked how concerned they were with all four target issues after they rated the importance of each issue. When the presuppositions resulting from the two triggers (asking the issue in a previous block and the direct question format) suggested to respondents that the issue was an important one, the respondents might include this presupposition in the common ground and use it when answering the concern items. Specifically, when respondents think an issue is important, they might feel that they ought to be concerned with that issue. Table 4.14 displays the correlation between the importance items and the concern items, with the correlation coefficients ranging from .55 to .63. It seems that the presupposition of importance alone can explain 30 to 40% of the variance of the concern items.

Table 4.14. Correlations Between Importance Items (Inference Questions) and Concern Items

Issues	Correlation
Environment	.63
Highway and Bridges	.60
Agriculture	.55
Parks and recreation	.62

4.3.4 Conclusions

The final experiment examined the effects of question format and prior questions on respondents' inferences about the importance of issues and on responses to subsequent concern questions. Consistent with the findings by Knauper (1998) and Sterngold et al. (1994), direct questions in the format of "How important do you think ___ is?" were found to carry a presupposition that the issue is an important one. Respondents were more likely to choose the upper end of the response scale (the important options) when presented with a direct than a filtered question. On the other hand, the mean ratings of importance showed reversals for two of the four items.

The very act of asking a question in a previous block of questions itself didn't carry the presupposition as originally hypothesized. It turned out that the previous block containing the target issues created a contrast effect, which affected respondents' inferences about the importance of the items.

The effect of the presupposition of importance can be seen through the significant correlations between the importance questions and the concern questions. The importance ratings explained about 30 to 40% of the variance in responses to the concern questions.

Linguists have long noted that presuppositions triggers differ in their ability to be neutralized. They distinguish hard triggers from soft triggers (see Abbott, 2005). Soft triggers are easily neutralized in a given context. One reason that the manipulation of including an issue in a previous block did not come out as expected is probably because that trigger is a soft one, whose presupposition can be easily canceled or suspended. The contrast effect lends some support to this speculation.

Chapter 5 Conclusions and Discussion

5.1 Summary of Results

This dissertation included six experiments that sought direct evidence that the four Gricean maxims are at work in the survey response process. Table 5.1 summarizes the experimental conditions and the main findings from the six experiments.

Experiment 1 varied the physical arrangement of survey questions on web pages, presenting a set of six loosely related items one question per screen, all on the same screen, or all in a grid on a single screen. But the physical arrangement did not seem to lead respondents to see the items as more (or less) strongly related. Although the relatedness ratings didn't provide direct evidence that respondents applied the maxim of relation, the responses (i.e., inter-item correlations) did differ by physical arrangement. When the introduction didn't mention anything about the relatedness (or unrelatedness) of the items, the grid condition produced the highest intercorrelations (measured by Cronbach's alphas) among the six items, partially replicating the previous research on the effects of physical arrangement (see Couper et al., 2001; Tourangeau et al., 2004). However, when the introduction told the respondents that the items were related, the grid yielded the lowest intercorrelations. It seemed that a grid format triggered different inferences and led to different results depending on the introduction given to the items.

Table 5.1. Summary of Results

Chapter	Maxim	IVs	DVs	Results
Chapter 2, Experiment 1	Relation	-Physical arrangement of survey questions on web screens (Grid vs. Multiple items on same screen vs. single item per screen) -Introduction	-Intercorrelations -Inferences (perceived relatedness ratings)	-Physical arrangement had no significant effect on intercorrelations; effect of introduction marginally significant; no significant interaction effect -Neither physical arrangement nor introduction had significant effects on perceived relatedness; interaction also non-significant -Hypotheses not supported
Chapter 3, Experiment 2	Manner	-Self-evident definitions for everyday terms vs. no definitions	-Response errors -Response times -Inferences	-Giving definitions to everyday terms significantly reduced variances of some questions; significantly changed the covariance matrix of some questions; didn't improve or reduce the accuracy of classifying instances to food categories -Giving definitions slowed down respondents with a high need for cognition, but not those with a low need for cognition; differences not significant -Giving definitions didn't have a significant effect on respondents' inference about the purpose and meaning of the terms -Some hypotheses were supported, though not strongly
Chapter 4, Experiment 3	Relation	-Scale values (-3 to 3 vs. 0 to 6) -Font of numbers (normal vs. faded)	-Mean shift in responses -Inferences (Accuracy of recall, attention to and usefulness of numbers, interpretation of end labels) -Response times	-Negative scale values produced significant mean shift -Fainter font moderated the effect of negative scale values for two of the four questions -Marginally significantly more recall of correct numbers given scale running from -3 to 3 -Significantly higher level of attention to scale number for -3 to 3 scales -Significantly more respondents associated -3 with presence of failure -Marginally significant longer response time when given -3 to 3 scale -Hypotheses regarding numerical values were supported; some hypotheses regarding fainter font were supported

Chapter	Maxim	IVs	DVs	Results
Chapter 4, Experiments 4 and 5	Quantity	<ul style="list-style-type: none"> -Introduction -Wording (one item reverse worded vs. none reverse worded) -Number of items (2 vs. 4) 	<ul style="list-style-type: none"> -Correlations -Inferences 	<ul style="list-style-type: none"> -Introduction didn't have a significant effect on correlations in experiment 4, but had a marginally significant interaction effect with the number of items in experiment 5 -Reversed wording significantly reduced the correlations -Four items significantly increased the correlations compared to two -Hypotheses partially supported
Chapter 4, Experiment 6	Quality	<ul style="list-style-type: none"> -Issues raised in a previous block of questions or not -Question format (direct vs. filtered) 	<ul style="list-style-type: none"> -Correlations -Inferences 	<ul style="list-style-type: none"> -Direct question led to a higher percentage of responses at the upper end of the scale -Whether an issue was mentioned previously didn't increase the importance ratings -Presupposition of importance explained about 30% to 40% of total variance of responses to concern items -Hypotheses partially supported

The use of grids in computerized surveys has been controversial; grids have been shown to be challenging to new users (Couper, 2000) but efficient once users have oriented themselves to the grid layout (Couper, 2000; Fuchs, 1999). Grids can increase the correlations among attitude items (Couper et al., 2001; Tourangeau et al., 2004), though this may represent increased error (Tourangeau et al., 2004). Despite general screen complexity of grids, respondents were found to be faster with them (Couper, 2000; Couper et al., 2001). The present experiment showed that grids are sensitive to introductions, pointing to an additional source of variability associated with grids. These results ought to alert survey researchers to use grids with caution.

The second experiment provided unnecessary (and not very informative) definitions for everyday terms like “poultry” in one condition, contrasting this with a condition in which respondents did not get any definitions. Instead of being confused by the apparently redundant definitions for terms, respondents seemed to have followed the “interpretability presumption” (Clark and Schober, 1992), believing that what survey questionnaires say and how they say it are meaningful. They seemed to incorporate the definitions into their interpretation of the survey questions. A larger percentage of respondents inferred that the definitions were intended for someone else or the terms were used in a more technical sense when they were given the redundant definitions than when they were not given them. They also seemed to use those redundant definitions; their responses to (some of) the key items had significantly smaller variances when redundant definitions were present than when they were not. The covariance structures of responses to key items and to related items were changed as well when definitions were provided.

This experiment had a small sample size ($n=160$) and low power, contributing to the weak evidence that respondents generated inferences based on the maxim of manner. The finding that the seemingly redundant definitions reduced variances of responses, though contradictory to the “more is less” phenomenon described in Young (1999) and Gerber et al. (1996), is encouraging news to survey researchers. Nonetheless, a larger sample size might be needed before we can draw definitive conclusions regarding usefulness of definitions for familiar terms.

The experiment on the numerical labels for scale points and the font in which they were presented provided a variety of evidence that respondents worked out an inference from the numerical values, drawing on the maxim of relation, and based their responses on the inference. The shift in responses induced by the numerical scale values was unexpectedly robust; when a scale started with a negative number, it pushed the responses to the right or positive end of the scale across items and across fonts. Process measures such as the recall task, response times, and retrospective probes confirmed that respondents paid attention to the numerical labels on the scales, carefully processed the negative numbers, and worked out inferences to interpret the verbal labels on the end points of the scale. This experiment provided one more piece of evidence that respondents pay attention to peripheral cues (the style elements in Couper et al. 2004); in designing questions, we should pay more attention to what we put on the screen or on the page.

Increasing the number of specific questions, rephrasing the question to explicitly include the prior specific item, and placing the two questions in different conversational contexts seem to prevent respondents from using the maxim of quantity when a general

question follows a specific one (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991). Experiments 4 and 5 were intended to examine the factors that may have a similar role when two highly similar items at the same level of generality are asked together. I replicated the work of Strack and colleagues (1991) on similar items in Experiment 4. In that study, respondents used the maxim of quantity when two similar items were placed in the same conversational context and lower correlations were the result. The tendency to “avoid redundancy” was robust across introductions in Experiment 4. Furthermore, reversed wording seemed to cause people to focus more on the conceptual overlap rather than structural dissimilarity; significantly more respondents thought the two target items were supposed to measure the same thing when one of the items was reverse-worded than when two items were worded in the same direction, leading to appreciably lower correlations. On the other hand, respondents didn’t apply the maxim of quantity in all circumstances. When the number of similar items was increased to four (as in Experiment 5), the correlations between the target items rose as well, suggesting that respondents stopped applying the “avoid redundancy” rule when they saw multiple items bundled together.

Experiment 6 on the effects of presupposition lent partial support to the existing literature that direct questions without a filter carry a presupposition that can influence respondents’ subsequent responses. Experiment 6 examined two methods of conveying the presupposition of “importance” by varying whether a particular issue was included in a prior block of questions and whether the importance question was asked in a direct way without a filter. Asking about an issue in a prior block turned out to be a weak trigger of the presupposition, whose effects were cancelled by the presence of other salient issues in

the same block. By contrast, direct questions without filters have a complex impact on survey responses; on the one hand, asking the importance questions in a direct way induced significantly more responses at the upper end of the scale – presumably because of the presuppositions direct questions carry. On the other hand, the mean responses to the direct questions were significantly lower than those to the filtered questions for two of the four issues.

Close examination of the univariate distributions of responses to these two issues revealed that filter questions managed to move respondents away from the “somewhat important” option to either the “very important” option or “neither important nor unimportant” option (see Table 4.12. in Chapter 4). In the direct question condition, answer categories were listed vertically with “very important” as the first answer category on the top. Responses to the direct questions were roughly evenly split between “very important” and “somewhat important” options. In the filter condition, a filter question first asked respondents to indicate whether they thought a certain issue was important or not. If they selected the “neither important nor unimportant” option, they would skip the next question; otherwise, they would be asked a degree-of-importance question, with the “very important” option being the first answer option on the top. The filter condition seemed to cause respondents either to turn neutral by selecting the middle category on the first screen or to become polarized in their answers by selecting the “very important” option.

There seem to be at least two completing explanations for this shift of the answers in the filter condition. The Gricean account predicts that the filter question encourages more answers on the lower end of the scale because it doesn’t carry a presupposition of

importance; the shift down to the middle category seems to fit this prediction. The follow-up question in the filter condition starts with “How important...?” and again carries a presupposition of importance that might have served as a reinforcement to those who have selected “important” on the previous screen; thus, leading to polarization.

A second explanation for this shift of answers is based on Krosnick’s notion of satisficing (Krosnick, 1991, 1999). According to Krosnick, primacy effects (selecting earlier response options) and status-quo responses (that is, selecting the middle category) are both manifestations of respondents’ selecting responses that reduce the level of effort needed to answer the questions. The observed shift to the middle category and the first responses in the filter condition seemed to fit this satisficing account as well. Both accounts are possible; Experiment 6 was not set up to tease apart these two competing explanations. Future experiments are needed to explain the effects on responses of filter questions versus direct questions. In addition, the conflicting evidence of Experiment 6 should alert us to the importance to examine response distributions before moving to aggregate level analysis such as tests of means and/or correlations.

5.2 Discussion

Among the six experiments aiming to retrieve direct evidence that the Gricean maxims are at work during the survey response process, some experiments didn’t provide strong evidence that the hypothesized maxims were applied. However, this doesn’t necessarily mean that respondents did not use the maxims. For instance, Chapter 2 showed that the same screen condition performs quite differently from the grid condition, despite the fact that in both conditions respondents were able to read the whole set of

questions before answering. This strongly suggests that the differences between the two formats are not driven by a noninterpretive process (e.g., differences in the accessibility of information used to answer the earlier items), since both the questions and the question order were constant. Only the layout differed. Thus, it seems premature to rule out an account based on interpretive inferences simply because this experiment didn't yield direct evidence for the predicted inference.

The question remains why some experiments failed to obtain evidence that respondents made the Gricean inferences as they formulated their answers. One reason could be respondents' limited ability to report on their cognitive processes. Nisbett and Wilson (1977) demonstrated that people are not always aware of the existence of stimuli that affect their responses or of the stimulus' impact on their responses. Nisbett and Wilson (1977) argued that higher order cognitive processes can occur outside of awareness; people had little or no access to their cognitive processes, and, thus, they couldn't report them.

5.2.1 Automatic versus Controlled Processing

If inferences based on the Gricean maxims were also the product of automatic process outside of respondents' awareness and control, then it would explain the absence of significant evidence from some of the experiments. There is evidence from different fields suggesting that at least some linguistic processing is automatic. The famous "Stroop effect" (in which the meaning of a color word makes it harder to say what color it is printed in) is one example. The Stroop effect shows that extracting word meaning is automatic – it happens without our intent and it is impossible to stop. Paradis (1998,

2002) argues more generally that linguistic competence is acquired incidentally, stored implicitly, used automatically, and subserved by procedural memory.

Research in the field of neurolinguistics and cognition, involving ERP (event-related brain potentials), indicates that people go through different cognitive processes when conducting syntactical and semantic analysis. For instance, Hahne and Friederici showed that the first pass of syntactic analysis – the initial representation of structure – is mostly automatic (Hahne & Friederici, 1999). However, infrequent or incorrect syntactical structures tend to call for a more controlled second passing (Hahne & Friederici, 1999). Hahne and Friederici (2002) further presented that semantic aspects of sentence comprehension involve controlled processes. Both papers demonstrated that people went through different processes with syntactically (or semantically) correct sentences versus syntactically (or semantically) incorrect sentences. In another study indicating that both automatic and controlled processes are involved in comprehension, Hill and his colleagues showed that access to semantic memory by spreading activation is automatic, but that integrating prime and target words into a semantic context is a controlled process (Hill, Strube, Roesch-Ely, & Weisbrod, 2002).

There is not much direct evidence on the role of automatic and controlled processes in pragmatic processing. But a few findings from Gibbs' work on conversational implicatures and indirect speech acts are worth mentioning (Gibbs, 1981, 1986, 1997). First, people recognize a distinction between what speakers say and what they implicate in particular contexts (Gibbs, 1997). Similarly, people know which indirect requests fit in different situations and view the one that fits the situation best as conventional (Gibbs, 1981, 1986; see Austin, 1962, on his original postulations of direct

versus indirect speech acts). Furthermore, people are generally faster in understanding conventional indirect requests than unconventional ones in a particular context (Gibbs, 1981, 1986). Gibbs's work suggests that conventional utterances are understood via an automatic process, even if they involve indirect requests. This is analogous to Hahne and Friederici's finding that correct sentences are automatically processed (Hahne & Friederici, 1999, 2002).

To sum up, linguistic processing in general seems to employ both types of process. Language processing starts automatically as we draw on our pool of vocabulary, grammar, and common knowledge to understand and respond to sentences (or utterances). However, we can switch to controlled processes when situations call for it. Incorrect use of words, infrequent syntactic structures, and apparently inappropriate fit to contexts all fall into this category. In terms of conversational inferences based on Gricean maxims, I would argue for the same dual set of processes. Most of the conversational implicatures such as generalized conversational implicatures, metaphors, and sarcasm, are processed automatically. People pick up the implicature instantly without intention, control, and awareness. As a matter of fact, people pick up the "what is implicated" so automatically that some metaphors or implicatures have lost their novelty and become clichés. On the other hand, people switch to controlled processes when they encounter surprising information or when automatic processes failed to yield a reasonable meaning to a sentence or an utterance.

This dissertation didn't provide much direct evidence to substantiate these assertions regarding automatic versus controlled processing of Gricean implicatures; this was not my focus. However, some experiments produced indirect evidence regarding

process. For instance, in the experiment on the numerical labels of rating scales (Experiment 3 in Chapter 4), response time measures showed that people were generally slower when they were given a scale starting with a negative number (-3) and they were more likely to recall that numerical label than their counterparts who were given a scale beginning with 0. A scale starting with 0 is less surprising and the processing of it is probably automatic; this would explain the poorer recall of that label. By contrast, a negative number is quite novel and surprising information, which caught people's attention and triggered more controlled processing. This experiment provides some evidence that Gricean effects could be both automatic and controlled depending on the context.

5.2.2 Gricean Effects or Satisficing

Tourangeau, Rips, and Rasinski (2000) outline the typical cognitive processes involved when respondents answer survey questions: comprehension of survey questions, retrieval of relevant information, judgment and estimation, and reporting answers. They note that respondents may adopt different response strategies, slacking on certain cognitive steps and skipping others. Krosnick dubbed such respondents "satisficers" (Krosnick, 1991, 1999; see also Tourangeau, Rips, and Rasinski, 2000 for a review).

Krosnick distinguished two types of satisficing (Krosnick, 1991, 1999). The weak satisficers execute all four components of the response process, but do not execute them as thoroughly as they should. They settle for merely satisfactory answers rather than the most accurate ones (Krosnick, 1999: p548). By contrast, the strong satisficers skip the retrieval and judgment and estimation components altogether and base their

responses on peripheral cues. Strong and weak satisficing each produce certain types of measurement errors (see Krosnick 1991, 1999 for a review). For instance, agreeing with assertions is one manifestation of strong satisficing whereas picking earlier responses in a visual mode is a manifestation of weak satisficing.

Do the results obtained from the six experiments here reflect Gricean inferences or respondents' satisficing? As a matter of fact, results such as the mean shift caused by a negative number, or the different response distribution produced by variations in question order were first regarded as measurement errors ("artifacts") on the part of respondents. Superficially, it looks like satisficing respondents could be at fault. However, as Schwarz (1996) pointed out, these response patterns are not errors caused by lazy or poorly motivated respondents. Rather, they are due to respondents' efforts to help out by being good respondents. Respondents use what they know – including conversational maxims or rules – in answering survey questions. Even Krosnick acknowledge that respondents applied conversational rules in survey or experimental setting in his research (Holbrook, Krosnick, Carson, & Mitchell, 2000; Krosnick, Li, & Lehman, 1990). As discussed in Chapter 1, it is now widely accepted that respondents bring conversational maxims to the survey research setting (Schwarz, 1996).

Just as not everyone satisfices in survey response, not everyone uses conversational rules. McCabe and Brannon (2004) demonstrated that only people with a high need for cognition adhered to the "avoid redundancy" rule and described the use of conversational maxims as optimizing behaviors. That seems to put Gricean's conversational rules in opposition to Krosnick's satisficing account. However, I disagree with this view of satisficing and application of Gricean maxims as polar opposites.

Krosnick proposed a continuum indicating the degrees of thoroughness in the response process with strong satisficers as the one end and optimizers as the other (Krosnick, 1991, 1999). Pragmatic processing involves both low effort automatic processes and high effort controlled processes. Thus, both the satisficers and optimizers may draw automatic inferences based on Gricean maxims, but only the optimizers will carry out the more controlled processes that require extra effort. It is unwise to use a black-and-white model to integrate satisficing and Gricean effects, given the complexity of human language use.

To conclude, this dissertation provided limited evidence that respondents used the Gricean maxims in survey response processes and observed some changes in responses under different manipulation conditions. More systematic research is required to pursue the topic further. Additional experiments, including ones that incorporate other techniques such as eye tracking or cognitive interviews may help to uncover the mechanisms affecting survey responses.

Reference

- Abbott, B. (2005). "Where have some of the presuppositions gone?" In B. Birner & G. Ward (Eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*. Philadelphia: John Benjamins.
- Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.
- Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot: Gower.
- Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly*, 47, 528-546.
- Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, 50, 240-250.
- Bloom, J. E., & Schober, M. F. (1999). Respondent cues that survey questions are in danger of being misunderstood. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 992-997). Alexandria, VA: ASA.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals of language*. Cambridge: Cambridge University Press.
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment*, 48, 306-307.
- Cannell, C. F., & Kahn, R. (1968). Interviewing. In G. Lindzey & E. Aronson (Eds.),

- The Handbook of social psychology*, Vol 2. (pp. 526-595). Reading, MS: Addison-Wiley.
- Cannell, C. F., Miller, P. V., & Oksenberg, L. F. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389-437). San Francisco: Jossey-Bass.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions* (pp. 15-48). New York: Russell Sage.
- Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley and Sons.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Conrad, F. G., Couper, M. P., Tourangeau, R. & Peytchev, A. (In press). Use and non-use of clarification features in web surveys. *Journal of Official Statistics*.
- Conrad, F. G., & Schober, M. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, 64, 1-28.
- Converse, P. E. (1970). Attitudes and no-attitudes: Continuation of a dialogue. In E. R. Tufté (Ed.). *The quantitative analysis of social problems*. Reading, Mass: Addison-Wesley.
- Couper, M. P. (2000). Usability evaluation of computer assisted survey instruments. *Social Science Computer Review*, 18, 384-396.
- Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, 68, 255-266.
- Couper, M. P., Traugott, M., & Lamias, M. (2001). Web survey design and

- administration. *Public Opinion Quarterly*, 65, 230-253.
- Davis, W. A. (1998). *Implicature: intention, convention, and principle in the failure of Gricean theory*. Cambridge: Cambridge University Press.
- Dillman, D. A., & Bowker, D. K. (2001). The web questionnaire challenge to survey methodologists. In U-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 159-177). Lengerich, Germany: Pabst Science Publishers.
- Fillmore, C. J. (1999). A linguistic look at survey research. In Sirken et al. (Eds.), *Cognition and survey research* (pp. 183-198). New York: Wiley.
- Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly* 56, 218-231.
- Fuchs, M. (1999). Screen design and question order in a CAI instrument effects on interviewers and respondents. Paper presented at the 54th Annual Conference of the American Association for Public Opinion Research, St. Pete Beach, Florida.
- Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.
- Gaskell, G., O'Muircheartaigh, C.A., & Wright, D. B. (1994). Survey questions about the frequency of vaguely defined events: The effects of response alternatives. *Public Opinion Quarterly*, 58, 241-254.
- Gendall, P., & Carmichael, V. (1997). A test of the conversational logic analysis model of question order effects. *Marketing Bulletin*, 8, 41-53.
- Gerber, E. R., Wellens, T. R. & Keeley, C. (1996). Who lives here? The use of vignettes in household roster research. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 962-967). Alexandria, VA.
- Gibbs, R. W., Jr. (1981). Your wish is my command: Convention and context in

- interpreting indirect requests. *Journal of Verbal Language and Verbal Behavior*, 20, 431-444.
- Gibbs, R. W., Jr. (1986). What makes some indirect speech acts conventional? *Journal of Memory and Language*, 25, 181-196.
- Gibbs, R. (1997). Pragmatics in understanding what is said. *Cognition*, 62, 51-17.
- Graf, L. (2002). Assessing Internet Questionnaires: The Online Pretest Lab
In B. Batinic, U-D. Reips, & M. Bosnjak (Eds.). *Online Social Sciences* (pp. 49-68). Hogrefe & Huber Publishers.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax And Semantics, 3: Speech Acts* (pp. 41-58), Academic Press: New York.
- Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.
- Groves, R. M. (1989). *Survey error and survey cost*. New York: John Wiley and Sons.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.
- Haddock, G., & Carrick, R. (1999). How to make a politician more likeable and effective: Framing political judgment through the numeric values of a rating scale. *Social Cognition*, 17, 298-311.
- Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, 11, 193–204.
- Hahne, A., & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, 13, 339–356.
- Hansen, M. H., Hurwitz, W. N., & Bershad, M. A. (1961). Measurement errors in

- censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. (1953). *Sample survey methods and theory*. New York: John Wiley and Sons.
- Harris, R. J. (1973). Answering questions containing marked and unmarked adjectives and adverbs. *Journal of Experimental Psychology*, 97, 399-401.
- Hill, H., Strube, M., Roesch-Ely, D., & Weisbrod, M. (2002). Automatic vs. controlled processes in semantic priming – Differentiation by event-related potentials. *International Journal of Psychophysiology*, 44, 197-218.
- Hilton, D. J. (1995). The social context of reasoning: conversational inference and rational judgment. *Psychological Bulletin*, 118, 248-271.
- Hippler, H., & Schwarz, N. (1986). Not forbidding isn't allowing: The cognitive basis of forbid-allow asymmetry. *Public Opinion Quarterly*, 50, 87-96.
- Hippler, H. J., Schwarz, N., & Sudman, S. (Eds.). (1987). *Social information processing and survey methodology*. Springer-Verlag.
- Holbrook, A., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating conversational conventions disrupts cognitive processing of attitude questions. *Journal of Experimental Social Psychology*, 36, 465-494.
- Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q- and R-based implicatures. In D. Schiffrin (Ed.). *Meaning, Form, and Use in Context* (pp. 11-42). Washington DC: Georgetown University Press.
- Horn, L. R., & Ward, G. (Eds.) (2004). *The Handbook of pragmatics*. Blackwell.
- Jabine, T. B., Straf, M. L., Tanur, J. M. & Tourangeau, R. (1984). *Cognitive aspects of*

- survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.
- Ji, L. J., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin*, 26, 586-594.
- Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion questions. *Journal of the Royal Statistical Society (Series C)*, 27, 149-161.
- Kalton, G., & Schuman, H. (1982). The effects of questions on survey responses: A review. *Journal of the Royal Statistical Society (Series A)*, 145, 42-73.
- Kihlstrom, J. F. (1995). *From the subject's point of view: The experiments as conversation and collaboration between investigator and subject*. Invited address presented at the meetings of the American Psychological Society, New York.
- Knauper, B. (1998). Filter questions and question interpretation-Presuppositions at work. *Public Opinion Quarterly*, 62, 70-78.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 213-236.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50: 537-567.
- Krosnick, J., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgment. *Journal of Personality and Social Psychology*, 59, 1140-1152.
- Leech, G. (1983). *Principles of Pragmatics*. London: Longman.
- Levinson, S. (1983). *Pragmatics*. Cambridge University Press.
- Levinson, S. (1995). Three levels of meaning. In F. R. Palmer (Ed.), *Grammar and*

- meaning* (pp. 90-115). Cambridge: Cambridge University Press.
- Lind, L. H., Schober, M. F., & Conrad, F. G. (2001). Clarifying question meaning in a web-based survey. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.
- Lipscomb, T. J., McAllister, H. A., & Bregman, N. J. (1985.) Bias in eyewitness accounts: The effects of question format, delay interval, and stimulus presentation. *Journal of Psychology*, 119, 207-212.
- Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7, 550-72.
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13: 585-89.
- Martin, E. A., Campanelli, P. C., & Fay. R. E. (1991). An Application of Rasch Analysis to questionnaire design: Using vignettes to study the meaning of ‘Work’ in the Current Population Survey. *The Statistician*, 40, 265-276.
- Mason, R., Carlson, J., & Tourangeau, R. (1994). Contrast effects and subtraction in part-whole questions. *Public Opinion Quarterly*, 58, 569-578.
- McCabe, A., & Brannon, L. (2004). Application of conversational norms to the interpretation of survey results as a function of participants’ need for cognition. *The Journal of Psychology*, 138, 91-94.
- Metzner, H., & Mann, F. (1953). Effects of grouping related questions in questionnaires. *Public Opinion Quarterly*, 17, 136-141.

- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, 29, 1011-1020.
- O'Muircheartaigh, C. (1999). CASM: Successes, failures, and potential. In M. G. Sirken, D. J. Hermann, S. Schetcher, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and Survey Research* (pp. 39-64). New York: Wiley.
- O'Muircheartaigh, C., Gaskell, G., & Wright, D. B. (1995). Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics*, 11, 295-307.
- Paradis, M. (1998). The other side of language: Pragmatic competence. *Journal of Neurolinguistic*, 11, 1-10.
- Paradis, M. (2002). Neurolinguistics of bilingualism and the teaching of languages. Talk given at the *Pluridisciplinary Colloquium on the Multimodality of Human Communication*. May. Toronto. (URL: <http://www.semioticon.com/virtuals/talks/paradis.pdf>)
- Petty, R. E., & Jarvis, W. (1996). An individual differences perspective on assessing cognitive processes. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 221-258). San Francisco: Jossey-Bass Publishers
- Ratcliff, R. (1993). Methods for dealing with response time outliers. *Psychological Bulletin*, 114, 510-532.
- Reips, U-D. (2002). Context effects in web surveys. In B. Batinic, U-D. Reips, and M.

- Bosnjak (Eds.). *Online Social Sciences* (pp. 69-79). Hogrefe & Huber Publishers.
- Schaeffer, N. C. (1991). Conversation with a purpose—or conversation? Interaction in the standardized interview. In P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 367-391). New York: Wiley.
- Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.
- Schober, M. F., Conrad, F. G., & Bloom, J. E. (2000). Clarifying word meanings in computer-administered survey interviews. In L. R. Gleitman, & A. K. Joshi (Eds.) *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 447-452). Mahwah, NJ: Lawrence Erlbaum
- Schober, M. F., Conrad, F. G., & Fricker, S. S. (1999). When and how should survey interviewers clarify question meaning? In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 986-991). Alexandria, VA: ASA.
- Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and congruity in survey questionnaire. *Public Opinion Quarterly*, 47, 112-115.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
- Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Mahwah, NJ: Erlbaum.
- Schwarz, N. (1998). Warmer and more social: Recent developments in cognitive psychology. *Annual Review of Sociology*, 24, 239-264.

- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54: 93-105.
- Schwarz, N. (2000). Social judgment and attitudes: Warmer, more social, and less conscious. *European Journal of Social Psychology*, 30, 149-176.
- Schwarz, N. (2003). Culture-sensitive context effects: A challenge for cross-cultural surveys. In J. Harkness, et al. (Eds.), *Cross cultural survey methods*. (pp. 93-100). New York: Wiley.
- Schwarz, N., & Bienen, J. (1990). What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology*, 4, 61-72.
- Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference: The impact of frequency range on diagnostic judgment. *Applied Cognitive Psychology*, 5, 37-50.
- Schwarz, N., Grayson, C. E., & Knauper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research*, 10, 177-184.
- Schwarz, N., & Hippler, H.-J. (1991). Response alternatives: The impact of their choice and ordering. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 41-56). New York: Wiley.
- Schwarz, N., & Hippler, H. J. (1995). The numeric values of rating scales: A comparison of their impact in mail surveys and telephone surveys. *International Journal of Public Opinion Research*, 7, 72-74.
- Schwarz, N., & Hippler, H. J. (1995b). Subsequent questions may influence answers to

- preceding questions in mail surveys. *Public Opinion Quarterly*, 59, 93-97.
- Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and subsequent judgment. *Public Opinion Quarterly*, 49, 388-395.
- Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570-582.
- Schwarz, N., & Oyserman, D (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction, *American Journal of Evaluation*, 22, 127–160.
- Schwarz, N., & Scheuring, B. (1988). Judgment of relationship satisfaction: Inter- and intraindividual comparison strategies as a function of questionnaire structure. *European Journal of Social Psychology*, 18, 485-496.
- Schwarz, N., Strack, F., Hilton, D. J., & Naderer, G. (1991). Judgmental biases and the logic of conversation: The contextual relevance of irrelevant information. *Social Cognition*, 9, 67-84.
- Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, 55, 3-23.
- Schwarz, N., Strack, F., Muller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition*, 6, 107-117.
- Simons, M. (2001). On the conversational basis of some presuppositions. In R.

- Hastings, B. Jackson, & Z. Zvolensky (Eds.), *Proceedings of Semantics and Linguistic Theory* 11, Ithaca, NY: CLC Publications
- Simons, M. (2003). Presupposition and accommodation: Understanding the Stalnakerian picture. *Philosophical Studies*, 112, 252-278.
- Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (Eds.). (1999). *Cognition and survey research*. New York: Wiley.
- Smith, T. W. (1995). *Little things matter: A sampler of how differences in questionnaire format can affect survey responses*. Paper presented at the annual meeting of the American Association for Public Opinion Research. Ft. Lauderdale, FL.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. (2nd Edition). Oxford: Blackwell.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. (1st Edition). Cambridge, MA: Harvard University Press.
- Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz & P. Under (Eds.), *Semantics and Philosophy* (pp. 197–213), New York University Press, New York.
- Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy*, 25, 701–721.
- Sterngold, A., Warland, R. H., & Herrmann, R. O. (1994). Do surveys overstate public concerns? *Public Opinion Quarterly*, 58: 255-263.
- Strack, F. (1994). Response process in Social Judgment. In R. Wyer, Jr. & T. Srull (Eds.), *Handbook of social cognition*. 2nd edition, vol1. (pp. 287-317). Hillsdale, N.J. : Laurence Erlbaum Associates.
- Strack, F., Martin, L.L., & Schwarz, N. (1988). Priming and communication: Social

- determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, 18, 429-442.
- Strack, F., Schwarz, N., & Wanke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, 9, 111-125.
- Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of American Statistical Association*, 85, 232-241.
- Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.
- Sudman, S., Bradburn, N. & Schwarz, N. (1996). *Thinking about answers: The Application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass Publishers.
- Suessbrick, A. A., Schober, M. F., & Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 907-912). Alexandria, VA: ASA.
- Tourangeau, R. (1999). Casting a wider net: Contributions from new disciplines. In Sirken et al. (Eds.), *Cognition and survey research* (pp. 177-182). New York: Wiley.
- Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research*, 15, 3-7.
- Tourangeau, R., & Conrad, F. G. (2004). Everyday concepts and reporting errors. Paper Presented at the 59th Annual Conference of the American Association for Public Opinion Research. Phoenix, AZ.

- Tourangeau, R., Couper, M., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68, 368-393.
- Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299-314.
- Tourangeau, R., Rasinski, K., & Bradburn, N. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly*, 55, 255-266.
- Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
- Tourangeau, R., & Smith, T. W. (1996). Asking sensitive information: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60, 275-304.
- Young, M. (1999). Cooperative plan identification: Constructing concise and effective plan descriptions. In *Proceedings of the National Conference of the American Association for Artificial Intelligence*, Orlando, FL.
- Van Zandt, T. (2002). Analysis of response time distributions. In H. Pashler & J. Wixted (Eds.), *Stevens 'handbook of experimental psychology (3rd ed.)*, Vol. 4: *Methodology in experimental psychology* (pp. 461-516), New York: John Wiley & Sons.
- Wanke, M. (2002). Conversational norms and the interpretation of vague quantifiers. *Applied Cognitive Psychology*, 16, 301-307.
- Wanke, M., Schwarz, N., & Noelle-Neumann, E. (1995). Asking comparative

questions: The impact of the direction of comparison. *Public Opinion Quarterly*, 59, 347-472.

Winkielman, P., Knauper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, 75, 719-728.