A closer look at web questionnaire design

Toepoel, V.

A Closer Look
at
Web Questionnaire Design

A Closer Look
at
Web Questionnaire Design

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof.dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op maandag 3 november 2008 om 14.15 uur door

VERA TOEPOEL,

geboren op 3 augustus 1980 te Tilburg.

Promotor: prof.dr. Arthur van Soest
Copromotor: dr. Marcel Das

Overige commissieleden: prof.dr. Mick P. Couper
prof.dr. Don A. Dillman
prof.dr. Arie Kapteyn
dr. Edith de Leeuw

HOC ERAT IN VOTIS

Acknowledgements

ALIQUIS IN OMNIBUS, NULLUS IN SINGULUS

When I finished my Master in Leisure Studies in 2001, I was a little bit disappointed. I had followed many courses, and learned a lot, but after my graduation I really thought the multidisciplinary character of the study caused that I knew nothing about everything; I knew something about economy, psychology, sociology, methodology, and even something about law, but I could not call myself an economist, psychologist etc. So what was I?

Working at CentERdata made me realize more and more that my background was something to be proud of. Although answering "Leisure Studies" to the question about my background raised many eye-brows, it was really an asset to have a multidisciplinary background. I experienced economists to see the world through economist-glasses, while psychologist see it through psychologist-glasses. My study helped me in my work as I was not biased by discipline; I knew that every focus is not all of it and probably the best way to approach scientific research is through multidisciplinary sun-glasses!

I started to work at CentERdata in 2003. My work at CentERdata was as divers as my studies; I knew something about programming (questionnaires) and something about research, but not enough to feel secure about my knowledge on each subject. That's why I really wanted to write a dissertation: after a PhD I thought I could call myself a researcher in the true sense of the word.

I am very grateful that Marcel Das gave me the opportunity to write a dissertation. He invested time and money in me, as both my director as my co-promotor, and this dissertation would not have been possible without his consent. I found him to be an excellent co-promotor. I have the tendency to forget to look at minor details, and Marcel always found the faults I forgot to detect. Also, Marcel never failed to read anything before a meeting, and since I know how busy he is, this is really something. Another person who was crucial in this project of becomming a PhD is Arthur van Soest. I think he made a huge step to supervise someone with my background, who knows nothing about econometrics. And although we sometimes had some kind of a discipline gap, he never gave me the idea I knew too little (although I obviously do). His knowledge of statistics and writing papers helped me enormously. In addition to Marcel and Arthur, I really need to thank my colleagues for their support during the last years: Corrie, Miquelle, Marije, and Marika, and also Josette, Hendri, and Marius helped me each in their own way. Further, it is really an honour to have the crème-de-la-crème in the field of web questionnaire design in my PhD committee. Mick Couper, Don Dillman, Arie Kapteyn, and Edith de Leeuw form a committee one could only dream about. Thank you for giving me the honour!

I make use of this opportunity to thank my parents for all they have done for me. They have always supported me in the way that's best for me. My mother always tries to slow me down, since I have the tendency to speed through life. Mama, thank you for always knowing what's best for me. Daddy, thank you for always encouraging me to sport. Sporting clears my head and is absolutely necessary for me to keep sane. Of course I would also like to thank my three babies, Eddy, Grizzly, and Aitor, for their unconditional love. And last but not least: Ramon. You always tried to stimulate me to solve (statistical) problems myself, instead of just giving me the answers and choose the easy way. Either intentionally or not, you are the main reason I pursued the act of becoming a doctor.

Contents

1 Introduction

IN PRIMIS

1.1 Introduction

People get influenced by information in all things they do. For example, people use information about prior contacts in their ordinary conversations; they use information available to decide which product they are going to buy, etc. Information can also be used to achieve certain goals. Governments sometimes choose to provide imperfect or incorrect information to influence public opinion; directors may choose which information they distribute to influence employees and stockholders, etc. Similarly, survey respondents use information to decide which answer they are going to report. They use information available in the questionnaire, question, and answer. They also use information they got in prior surveys, as well as information they have obtained in ordinary life. All this information is used to fill out a questionnaire. Researchers influence which information is available to respondents. They leave, either intentionally or not, cues in a questionnaire which are used by respondents to decide which answer to select.

When designing a survey, a researcher has to make many choices. Particularly in web surveys, little is known about the consequences of these choices on the quality and interpretability of the data. It has only been a decade since systematic research started on how visual layout of questions in surveys influences respondent answers. This research has been motivated, and sustained, in part, by the development of Internet survey methods, and the desire to relate research findings to those collected through other modes of administration. More specifically, one of the difficulties faced by surveyors in the early 21st century is that most individual survey modes suffer from specific problems, e.g. inadequate coverage or poor response rates making it necessary to mix survey modes. These developments make it imperative that we develop a better understanding of how questionnaire design affects the quality and interpretability of survey data. In particular data collected through the web, since the effects of web questionnaire design on respondents' answers are relatively less known. Research is warranted to know more about the consequences of design choices in web surveys in order to develop a better understanding of the factors influencing the question-answering process. These factors are discussed below.

First of all, a researcher has to make a decision on which kind of sample he or she wants to use. Different samples may yield different respondents' answers. For example, trained respondents may answer questions differently than fresh respondents: they may have higher content and procedural knowledge in answering surveys. Second, the mode of data collection is an important design decision. The way of communicating (paper, telephone, computer), the presence or absence of an interviewer and the use of visual information or not can influence respondents' answers. In addition, design choices about the questionnaire, questions, and answer categories send cues that influence respondents in their choice to select an answer. Respondents' personal characteristics such as age, gender, education, and personality factors may also influence the question-answering process. The goal of this dissertation is to analyze how some key web design choices can influence respondents' answers.

This dissertation contributes to the knowledge of web questionnaire design in several ways. First, the few studies that have been done are mostly based on non-representative samples for an entire population (e.g. student populations with high education level or volunteer opt-in panels with Internet access), and it may not be possible to generalize the results of these studies to an entire population. The studies in this dissertation, however, are based on a probability sample of the Dutch population (without the need to have access to the Internet) which makes it possible to generalize the results. Second, despite the (still) increasing popularity of online surveys, little research has examined the key factors influencing the question-answering process: the web as mode of administration and its effects on data quality (visual cues), a panel as type of sample (with its effects of re-interviewing), the questionnaire characteristics (questionnaire, question, and answer type) and respondents' personal characteristics (background characteristics as well as personality traits). This dissertation aims at gaining deeper insights into which factors matter and how they influence the quality of the survey data.

In six chapters this dissertation addresses the following design choices, all in the context of a web survey: interface design (paging versus scrolling), the impact of response categories, the effect of layout, the use of a panel as sample and its possible threats to data quality (panel conditioning and attrition bias), the relation between panel conditioning and web survey design, and the relation between panel conditioning and question type. All these concepts are related to the question-answering process. In the remainder of this chapter this process and influencing factors are explained in more detail. The chapter ends with the objectives of the different chapters in this dissertation.

The question-answering-process 1.2

Papers on response effects draw on social information-processing models of how people answer questions. Interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, and reporting it are the main psychological components of a process that starts with respondents' exposure to a survey question and ends with their report (Strack and Martin 1987; Sudman et al. 1996; Tourangeau et al. 2000). In the next subsections these steps will be discussed in more detail.

Step 1: Interpreting the question

The first step in the question-answering process is to understand what is meant by the question. There must be a shared meaning between the researcher and the respondent with respect to each of the words in the question as well as the question as a whole. To comprehend the question, the respondent considers the question and attempts to understand what information is requested. In doing so, the respondent is lead by information in the questionnaire, such as verbal and non-verbal cues.

Step 2: Retrieving information

Given the respondent's understanding of the question, the respondent then retrieves whatever information is necessary to respond. Information needed to formulate a response is retrieved from memory. Some questions do not require the retrieval of factual data, but information may still be retrieved from memory in the form of feelings, viewpoints, positions on issues and so on (Biemer and Lyberg 2003). The amount in which the respondent searches information for answering the question may differ because of the respondent's cognitive activity in answering the survey. People with a high need for retrieving information undergo different processes in formatting an answer than people with a low need for retrieving information. People with a high need tend to seek more information and think more carefully about it than people with a low need. People with a low need are more easily influenced by peripherical cues.

Step 3: Generating an opinion

In the third step of the question-answering process, the respondent is generating an opinion on the subject. This stage includes the process of reflecting on the issues raised by the questions in order to arrive at a report, attitude, belief, or opinion. Petty and Jarvis (1996) suggest that people with a low need for generating opinions are expected to be more susceptible to various low effort biases than people with a high need for generating opinions, such as being influenced by cues in a survey suggesting one response over another. On the other hand, Tormala and Petty (2001) found that people with a high need for generating opinions formed attitudes in a spontaneous, on-line fashion, whereas people with a low need for generating opinions formed them in a less spontaneous, more memory-based fashion. From this perspective, people with a high need for generating opinions could be more susceptible to verbal and non-verbal cues in a survey.

Step 4: Formatting a report

Following the opinion-stage, the next stage of the response process is referred to as the response formatting process. Answers to survey questions have to be reported in a format that is provided by the survey researcher. This format contains verbal and nonverbal cues that influence respondent behavior. Nonverbal cues include numerical, symbolic, and graphical languages that convey meaning in addition to the verbal language (Dillman and Christian 2002). Verbal and nonverbal cues can independently and jointly influence answers to questions.

1.3 Factors influencing the question-answering process

In this section the most important design aspects for (web)surveys influencing the question-answering process are discussed[1]: the sample, mode of administration, questionnaire characteristics, and respondents' personal characteristics.

1.3.1 The sample

The recruitment of samples for (web)surveys can be divided in probability and non-probability approaches. For a detailed description of types of (online) surveys see Couper (2000). Probability-based web surveys often restrict to populations with access to the web. Since people who have access differ systematically from people with no access to the Internet, this introduces measurement error. Therefore, some types of samples use traditional methods, such as telephone and face-to-face interviews, to reach a broader sample of the population. There are few organisations that provide Internet access, and if necessary a computer or a similar device, to those reached without a computer or Internet. Examples of these organisations are Knowledge Networks in the U.S. and CentERdata in the Netherlands.

[1] Since this dissertation discusses web questionnaire design, the factors influencing the question-answering process will be restricted to online surveys.

All our studies make use of this kind of probability sampling. Appendix A explains the sample methods for the two panels used in this dissertation: the CentERpanel and the LISS-panel. Using a probability sample for our studies without the need for Internet access is a big plus, since almost none of the research on web questionnaires design makes use of representative samples (most studies use students or volunteer opt-in panels). Therefore we can generalize conclusions to the adult population of the Netherlands. But using a representative sample also may introduce measurement error. Saris and Gallhofer (2007) argue in their inventory of design choices that quality differences in data could be the composition of the sample used in the study. They mention that lower educated and older people - groups normally missed in non-probability sampling - may produce lower quality data.

One of the most important developments in the social sciences over the last decade has been the increasing use of (web) panel surveys at the household or individual level. Panel data have important advantages for research, such as creating the possibility to analyze changes at the micro-level, to disentangle permanent from transitory characteristics, to distinguish between causal effects and individual heterogeneity, efficiency gains, etc. Two potential drawbacks compared to, e.g., independent cross-sections are attrition bias and panel conditioning effects. Attrition bias can arise if respondents drop out of the panel non-randomly, i.e., when attrition is correlated to a variable of interest. Panel conditioning arises if responses in one wave are influenced by participation in the previous wave(s). The experience of the previous interview(s) may affect the answers of respondents in a next interview on the same topic, such that their answers differ systematically from the answers of individuals who are interviewed for the first time. Trained respondents may answer questions differently than those with little or no experience in a panel. This can result in different responses with regard to content (e.g. because of increasing knowledge on topics) as well as the procedure (question-answering process).

The mode of administration 1.3.2

Each mode of data collection has its own associated errors. Web surveys are conducted since the last decade. However, little is known about effects in questionnaires using the computer. Although web questionnaires may draw on the principles of paper questionnaires, they also have new elements that require independent testing (use of mouse/screen, possibility for online routing and checks, use of different browsers and screen resolutions, etc.).

Saris and Gallhofer (2007) operationalized the data collection method into three possibilities: (1) computer-assisted data collection or not, (2) interviewer administered or not, and (3) visual information used or not. Their meta-analysis on the quality of survey questions, using a Multi-Trait-Multi-Method design, suggests the following order in quality: (1) mail, (2) Computer-Assisted Self-

Interviewing (CASI), (3) telephone, (4) Computer-Assisted Personal Interviewing (CAPI). The differences between mail (1) and CASI (2) are minimal, while differences between these two and telephone (3) and CAPI (4) are large. Since other quality criteria in the mode of data collection, such as non-response and possible (online) checks should also be considered, these results suggest that CASI may be the most reliable mode of data collection.

Key advantages of online surveys are the low cost structure, low variable costs, fast responses, no entry costs, possibility of online checks and routing, possibility of forced answers, use of audio-visual material, etc. The main disadvantage of online surveys is the relatively large coverage error. More and more people have access to the Internet every day, although the difference in characteristics between Internet users and non-Internet users becomes more problematic (also known as the Couper-paradox). After all, a person who is 91 years old with Internet is not representative for all 91 year old persons.

Although each mode has its own faults, design plays a much larger role in web surveys compared to other modes of administration. More tools are available, but every tool can introduce its own specific errors (e.g. the adding of pictures may cause a different interpretation of the verbal language of the question). Further, because of differences in browsers and screen resolutions, researchers never know exactly how a questionnaire appears on a screen. Since web surveys make use of visual information, visual cues can influence respondents (in addition to verbal cues). The origins of research on visual layout go back to observations made by Smith (1995), Jenkins and Dillman (1997), and Christian and Dillman (2004). A summary of the findings on visual design in web surveys is as follows:

1. Answer spaces send signals to respondents on what is or is not expected (Christian et al. 2007; Couper et al. 2000).

2. Larger answer spaces for open-ended questions elicit more information than smaller spaces (Christian and Dillman 2004; Smyth et al. forthcoming).

3. Placing appropriate labels on answer categories and using visual signals can increase significantly the number of people who provide what the researchers want (Christian et al. 2007).

4. It’s better to use visualness to convey scales, i.e. list scalar categories rather than having people carry the information from the stem of a question to the place where e.g. a "Number" is requested in a box. (Christian and Dillman 2004).

5. Shape and spacing of a scale influence respondent answers (Schwarz et al. 1998; Tourangeau et al. 2004;Tourangeau et al. 2007).

6. People expect more positive categories, or ones that express a greater degree of satisfaction or some other opinion to have higher labels (Tourangeau et al. 2004).

7. People expect categories to appear from positive to negative (Tourangeau et al. 2004).

8. Scales should appear in linear layouts rather than multiple columns or rows of categories, because multiple columns result in higher measurement error (Christian 2003; Christian and Dillman 2004; Dillman and Christian 2005).

9. Instructions need to precede answers if they are to be used that way. (Christian et al. 2005).

10. Symbols convey special meaning to respondents (Christian and Dillman 2004).

11. Pictures may change people's answers (Couper et al. 2004).

12. Visible categories are more likely to be chosen than are categories that require manual displays, e.g. in drop-down menus (Couper et al. 2000).

13. Changing color/hues for each end of a scale influences scalar responses when only polar point labels, and no numbers, are used. There may be a hierarchy of features that respondents attend to, with verbal labels taking precedence over numerical labels and numerical labels taking precedence over purely visual cues like color (Tourangeau et al. 2007).

14. Placing more items on a single screen increases correlations and item non-response (Tourangeau et al. 2004; Couper et al. 2000; Lozar Manfreda et al. 2002a).

1.3.3 Questionnaire characteristics

Questionnaire characteristics for web surveys can be divided in interface design, question type, and answer type. For web questionnaires, interface design varies in terms of the division of questions on screens and the navigation methods used. At one end are form-based designs that present questionnaires as one long form in a scrollable window, at the other end of the design continuum are screen-by-screen questionnaires that present only a single item at a time (Norman et al. 2001). Schonlau et al. (2002) suggest that scrolling questionnaires can become a burden to respondents and lengthy web pages can give the impression that the survey is too long to complete. On the other hand, scrolling questionnaires can lead to shorter completing times, help preserve the context of items, and avoid the problem of arbitrary page breaks. An advantage of the screen-by-screen design is the focusing on single items, but here the disadvantage is the loss of context and the large amount of operations needed to navigate through the survey. Presenting several items per screen fits somewhere in the middle of the design continuum; it reduces the number of screens without the need for scrolling.

In addition to the interface design, the question type might influence respondents' answers. Questions may be about facts, attitudes, knowledge, expectations, evaluations, etc. A researcher can then decide to use simple or complex assertions: ask an opinion or the strength of the opinion. A researcher has many options for how to formulate a question. Saris and Gallhofer (2007) give an extended elaboration of the choice of the formulation of the request. Choices, such as the use of absolute or comparative statements, balanced or unbalanced response alternatives, stimulation to answer in the request, use of extra information and arguments etc. are made either intentionally or not. But they all can influence respondents' answers.

The type of answer is also an important design decision a researcher has to make. One of the most basic decisions is whether to use open or closed questions. From a cognitive perspective, open questions present a free-recall task to respondents whereas closed questions present a recognition task. Closed questions may fail to provide an appropriate set of meaningful alternatives in substance or wording. Furthermore, respondents are influenced by the specific closed alternatives given. One can expect an answer closer to the correct value if the respondent must produce an answer himself (Schuman and Presser 1981). Schwarz (1996) and Schwarz et al. (1985) recommend asking questions in an open response format. However, open questions also have disadvantages. They may be affected by rounding (Tourangeau et al. 2000), and are therefore affected by estimation strategies. Furthermore, respondents find open questions more difficult to answer and (item) non-response tends to be higher compared to (item) non-response in closed questions (Blumberg et al. 1974; Crawford et al. 2001).

Respondents' personal characteristics 1.3.4

The extent to which personal characteristics, such as gender, age, and education affect respondents' performance is relatively unknown. Couper (2000) argues that design may interact with the type of web survey conducted and the population at which the survey is targeted. Although some research suggests effects (mainly caused by working memory capacity) other research found no variation in response effects caused by personal characteristics. McFarland (2001) did not find evidence that personal characteristics interact with the ordering of questions. The effects of question order were consistent for both sexes and across education levels. Tourangeau et al. (2007) observed no consistent variation in the impact of the layout of a response scale by gender, age group, or education group. Krosnick and Alwin (1987), on the other hand, find respondents with less education and more limited vocabularies to be influenced more by different answer categories. Knauper et al. (2004) also found differences due to working memory capacity. They found that the generally poorer memory of older adults results in increased scale effects, although some behavior is more salient to older people and may therefore be better remembered. Their results show that respondents' need to rely on estimation strategies is a function of their general memory performance and the extent to which the request is salient to them.

A respondent's personality might also influence the question-answering process. There are many personality scales that indicate respondents' differences in personality. A respondent's Need for Cognition and Need to Evaluate are discussed in more detail in this dissertation since these concepts measure two steps in the question-answering process: retrieving information and formatting an opinion.

The extent to which the respondent searches information for answering the question may differ for the respondent's cognitive activity in answering the survey. Cacioppo and Petty (1982) developed a scale to measure the need for cognition (see Appendix B for the Need for Cognition Scale). Need for Cognition (NFC) represents the tendency for individuals to engage in and enjoy thinking. They reasoned that when respondents are motivated (such as when the topic is of high relevance to the respondent) respondents are more eager to think than when their motivation is low (such as when the topic is of low interest). According to them, not only situational factors determine how much thinking occurs. Individual differences in intrinsic motivation to engage in cognitive activity are also likely to affect the effort a respondent is willing to make. People with a high need for cognition (HNC) undergo different processes in formatting an answer than people with a low need for cognition (LNC). People with HNC tend to seek more information and think more carefully before making an evaluation than people with LNC, who are more easily influenced by peripherical cues.

Jarvis and Petty (1996) developed a measure to assess individual differences in the propensity to engage in evaluation, the Need to Evaluate Scale (NES, see Appendix C for the questions used for this scale). Although attitudes are a fundamental concept in psychology, little research exists on how the process of reflecting on issues can be used to predict meaningful mental and behavioral processes. Bizer et al. (2004) found that respondents high in need to evaluate (HNE) reported their answers more quickly than those low in the need to evaluate (LNE). Petty and Jarvis (1996) suggest that people with a LNE are expected to be more susceptible to various low effort biases than people with a HNE, such as being influenced by cues in a survey suggesting one response over another. On the other hand, Tormala and Petty (2001) found that HNE individuals formed attitudes in a spontaneous, on-line fashion, whereas LNE individuals formed them in a less spontaneous, more memory-based fashion. From this perspective, people with a HNE could be more susceptible to verbal and nonverbal cues in a survey. Evaluation does not require effortful thought. The relation between the NES and the NFC was tested by Jarvis and Petty (1996) and was found moderate and positive (r=.35, p<.001).

1.4 Objectives of the different chapters

The central research question of this dissertation focuses on design choices in web surveys. To study some of the effects of these design choices on respondents' answers, six interrelated studies are conducted. The next paragraphs briefly describe the studies and their objectives.

Objectives Chapter 2
For web questionnaires, interface design varies in terms of the division of questions on screens and the navigation methods used. At one end are form-based designs that present questionnaires as one long form in a scrollable window, at the other end of the design continuum are screen-by-screen questionnaires that present only a single item at a time. Presenting questions in a matrix (several questions per screen) fits somewhere in the middle of the design continuum. Matrix questions are frequently used either to save space on the screen or to reduce the number of screens. It is relatively unknown, however, whether the questionnaire design affects the quality of the data. The objective of this study is to examine the effect of the number of items per screen. Measurement, nonresponse, time to complete the questionnaire, and respondents' evaluation of the questionnaire are taken into account, as well as the effect of personal characteristics.

Objectives Chapter 3
Studies about the cognitive and communicative processes underlying question answering in surveys suggest that the choice of response categories can have a

significant effect on respondent answers. The objective of this study is to explore the impact of response categories on the answers respondents provide in web surveys. An open-ended question format as a benchmark is added. The study uses a full range of question possibilities that vary in the difficulty of information processing. The heterogeneous nature of our sample makes it possible to measure the effect of personal characteristics and personality factors on survey responses and category effects. To indicate a respondent's motivation for giving correct answers, indexes for the respondent's need to think and need to evaluate are included in the analysis.

Objectives Chapter 4
Response categories can be presented in various ways: in a single column or in multiple columns, in rows, with labels for all categories or the endpoint categories only, with radio buttons or an answer box, etc. Differences in layout can yield detectable differences on responses to survey questions. Layout includes graphical, numerical, and symbolic languages that convey meaning in addition to the verbal language (Dillman and Christian 2002). A conceptual framework for explaining how visual languages may influence respondent behavior has been provided by Jenkins and Dillman (1997). Verbal and nonverbal cues can independently and jointly influence answers to questions. The objective of this study is to examine how visual languages influence answers to web surveys. These languages are individually manipulated on a rating scale. This study reports results focusing on respondents with different characteristics, which have received little attention so far.

Objectives Chapter 5
On the web, there has been an increasing use of panel surveys at the household or individual level, instead of using independent cross-sections. Panel data have important advantages, but there are also two potential drawbacks: attrition bias and panel conditioning effects. Attrition bias can arise if respondents drop out of the panel non-randomly, i.e., when attrition is correlated to a variable of interest. Panel conditioning arises if responses in one wave are influenced by participation in the previous wave(s). It is relatively unknown how re-interviewing influences respondents' answers. In addition, it is difficult to disentangle the total bias in panel surveys due to attrition and panel conditioning into a panel conditioning and an attrition effect. The objective of this study is to develop a test for panel conditioning allowing for non-random attrition.

Objectives Chapter 6
Trained respondents may answer questions differently than those with little or no experience in answering surveys. This can result in different responses with regard to content (increasing knowledge on topics in a survey) as well as proce-

dure (the question-answering process). The objective of this study is to examine the effect of panel experience on the question-answering process. Trained respondents may react differently to web survey design choices than inexperienced respondents. Because of their experience they may be able to process more information on a screen, e.g. make fewer errors when more items are placed on a single screen. In addition, they may be more or less susceptible to social desirability bias and more or less reluctant to select a response category that seems unusual in the range of responses. They also may be used to a particular question layout so that changing that layout (e.g. from disagree-agree to agree-disagree) may not be noticed. With this in mind, the objective of this study is to explore differences in web design effects between trained and fresh respondents.

Objectives Chapter 7
This chapter relates to the previous chapter in the sense that the study compares trained and fresh respondents. However, instead of questionnaire design this chapter focuses on which type of question is sensitive to repeated interviewing. The following types of questions are considered: knowledge, behavior, attitudinal, and factual questions. In addition, it is investigated whether respondents' characteristics (age, gender, education) interact with panel experience.

The chapters are based on the following papers

1. Toepoel, Vera, Marcel Das, and Arthur van Soest (2009), Design of Web Questionnaires: The Effects of the Number of Items per Screen, Field Methods, 21 (2).

2. Toepoel, Vera, Corrie Vis, Marcel Das, and Arthur van Soest (2009), Design of Web Questionnaires: an Information-Processing Perspective for the Effect of Response Categories, Sociological Methods and Research, Special Issue on Web Surveys.

3. Toepoel, Vera, Marcel Das, and Arthur van Soest (2006), Design of Web Questionnaires: the Effect of Layout in Rating Scales, CentER Discussion Paper 2006-30, CentER, Tilburg University.

4. Das, Marcel, Vera Toepoel and Arthur van Soest (2007), Can I use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys, CentER Discussion Paper 2007-56, CentER, Tilburg University.

5. Toepoel, Vera, Marcel Das and Arthur van Soest (2008), Design Effects in Web Surveys: Comparing Trained and Fresh Respondents, CentER Discussion Paper 2008-51, CentER, Tilburg University.

6. Toepoel, Vera, Marcel Das and Arthur van Soest (Forthcoming), Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents, Survey Research Methods.

and can therefore be read independently from each other. This implies that parts of the introductions may have some overlap. Table 1.1 shows the relation between the chapters and the factors influencing the question-answering process. The studies are interrelated and built upon each other.

In Chapter 5, 6, and 7 it is investigated whether using a panel as sample influences the answers provided by respondents. In Chapter 5 it is examined if re-interviewing causes panel conditioning. In addition, the influence of using a panel as sample is addressed by comparing trained and fresh respondents with regard to procedural knowledge (Chapter 6) and knowledge on content (Chapter 7).

The effects of the web as mode of administration are discussed in Chapter 4. Web surveys contain visual cues and this chapter investigates whether these cues influence the response process. In Chapter 6 the influence of visual cues is compared between trained and fresh respondents.

Questionnaire characteristics are discussed in all chapters. Interface design is investigated in Chapter 2, where the interface is varied between a scrolling and screen-by-screen design. In Chapter 6 the number of items per screen is varied in order to compare trained and fresh respondents. The question type is examined in Chapter 3 where we use questions that differ in difficulty to detect response category effects. In Chapter 5 and 7 it is investigated which type of question (e.g. knowledge, attitude, behavior, fact) is sensitive to repeated interviewing. In Chapter 6 the relation between question type (difficult versus easy) and answer type (high versus low response scale) is related to respondents' experience in answering surveys. Answer types are discussed in Chapter 3, 4, and 6. The influence of closed and open questions on respondents' answers are discussed in Chapter 3, the influence of the layout of response categories is discussed in Chapter 4, and in Chapter 6 these issues are related to panel experience.

The influence of demographics (gender, age, and education) is discussed in Chapter 2 to 4. In addition, in Chapter 3 the influence of personality factors is analyzed.

Table 1.1: Relation between the Chapters in this Dissertation and the Four Factors Influencing the Question-Answering Process

Chapter	Study	Factor
1	(Introduction)	
2	Design of Web Questionnaires: the Number of Items per Screen	*Questionnaire Characteristics:* Interface Design, *Personal Characteristics:* Demographics
3	Design of Web Questionnaires: an Information-Processing Perspective for the Effect of Response Categories	*Questionnaire Characteristics:* Question Type and Answer type, *Personal Characteristics:* Demographics and Personality Factors
4	Design of Web Questionnaires: Layout in Rating Scales	*Questionnaire Characteristics:* Answer type, *Personal Characteristics:* Demographics, *Mode of Administration:* Verbal and Visual Cues
5	Can I Use a Panel? Panel Conditioning and Attrition in Web Surveys	*Sample:* Panel, *Questionnaire Characteristics:* Question Type
6	Design Effects in Web Surveys: Comparing Trained and Fresh Respondents	*Sample:* Panel, *Questionnaire Characteristics:* Interface Design, Question Type, Answer Type, *Mode of Administration:* Verbal and Visual Cues
7	Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents	*Sample:* Panel, *Questionnaire Characteristics:* Question Type
8	(Conclusion)	

1.5 Appendix A: Description of the panels

This appendix gives some details about the trained panel (CentERpanel) and the fresh panel (LISS panel) used in this dissertation. Both panels are administered by CentERdata, a research and data collection institute affiliated with Tilburg University, The Netherlands. For the trained panel we will in particular focus on the recruitment of new members (to correct for attrition) and for the fresh panel we will provide some details on the original set-up.

CentERpanel (see also http://www.centerdata.nl/en/CentERpanel)

The CentERpanel was established in 1991 and exists about 2000 households. The panel is aimed to be representative of the Dutch-speaking population in the Netherlands. Panel members complete questionnaires at home every week through the Internet. Although the CentERpanel is an Internet-based panel, there is no need to have a personal computer with an Internet connection. The households that do not have access to Internet when recruited, are provided with a so-called Net.Box, with which a connection can be established via a telephone line and a television set. If the household does not have a television, CentERdata provides that too.

The recruitment of new panel members is done in three stages. In the first stage, a random sample (landline numbers) of candidates is interviewed by telephone. In the first telephone interview a number of questions are asked about demographic characteristics of the household. The interview ends with the question whether the person would like to participate in survey research projects. If so, the household is included in a database of potential panel members. If a household drops out of the panel, a new household is selected from the database of potential panel members. This is done on the basis of demographic characteristics (such that the panel will remain representative of the Dutch-speaking population). The selected household is asked whether the members of the household would like to become panel members. If so, a number of additional questions are asked and, if necessary, equipment is provided.

LISS panel (see also http://www.centerdata.nl/en/LISSpanel)

The LISS panel was established in 2007 and exists about 5000 households. At the time of the study presented in this chapter recruitment was not completely finished yet, but the first questionnaires were fielded. Panel members complete questionnaires at home every month through the Internet. As with the CentERpanel, Internet access is not a prerequisite for participation in the panel. If a household does not have Internet access at the time of recruitment into the panel, he or she is provided with a so-called SimPC (a basic PC with the ability to surf the Internet and some other basic functionalities).

The LISS panel is representative of the Dutch speaking population in the sense that the first recruiting of respondents was based on a random, nationwide sample of 10.600 addresses drawn from the community registers in cooperation with Statistics Netherlands. In a first step, all households in the sample receive an announcement letter and a brochure explaining the nature of the panel study. A prepaid incentive of 10 euro is added. Next, households are contacted by an interviewer, either by telephone or face-to-face, depending on whether a landline number is available. In a 10-minute recruitment interview some basic information is collected and at the end, the request to participate in the panel is made.

Within one to two weeks after the interview the respondents who agree to participate in the panel receive a confirmation e-mail and letter with login code, an information booklet and an answer card. Respondents without Internet or computer can confirm their willingness to participate by sending back the signed answer card, and the necessary equipment will be installed in their home. Respondents with Internet access can choose to confirm in the same way or to confirm online with the login code provided in the letter. In the latter case they can immediately start the first interview. This confirmation procedure ensures the double consent of each respondent.

Respondents who are initially not reached are re-contacted a number of times, first by phone (in case a landline number is available) and, if still not successful, face-to-face. If they are not reached after 15 face-to-face visits either, they receive a new invitation letter including a link to the Internet version of the recruitment interview, or a shortened paper version of the questionnaire.

The attempt is made to convert (soft) refusals into participation by a tailored procedure, depending on the refusal type. For example, older individuals who feel a bit unsure are offered a video demonstration in their home with a clear explanation of how the SimPC works.

Appendix B: Need for Cognition Scale 1.6

The Need for Cognition Scale (Cacioppo and Petty 1982) is a scale designed to measure the tendency for individuals to engage in and enjoy thinking. The list of 34 items is presented below.

1. I really enjoy a task that involves coming up with solutions to problems.
2. I would prefer a task that is intellectual, difficult, and important to one that's somewhat important but does not require much thought.
3. I tend to set goals that can be accomplished only by extending considerable mental effort.
4. I am usually tempted to put more thought into a task than the job minimally requires.
5. Learning new ways to think doesn't excite me very much.*
6. I am hesitant about making important decisions after thinking about them.*
7. I usually end up deliberating about issues even when they do affect me personally.
8. I prefer to let things happen rather than try to understand why they turned out that way.*
9. I have difficulty in thinking in new and unfamiliar situations.*
10. The idea of relying on thought to get my way to the top does not appeal to me.*
11. The notion of thinking abstractly is not appealing to me.*
12. I am an intellectual.
13. I only think as hard as I have to.*
14. I don't reason well under pressure.*
15. I like tasks that require little thought once I've learned them.*
16. I prefer to think about small daily projects to long-term ones.*
17. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
18. I find little satisfaction in deliberating hard and for long hours.*

19. I more often talk with other people about the reasons for and possible solutions to international problems than about gossip of tidbits of what famous people are doing.
20. These days, I see little chance for performing well, even in ’intellectual’ jobs, unless one knows the right people.*
21. More often than not, more thinking just leads to more errors.*
22. I don’t like to have the responsibility of handling a situation that requires a lot of thinking.*
23. I appreciate opportunities to discover the strengths and weaknesses of my own reasoning.
24. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.*
25. Thinking is not my idea of fun.*
26. I try to anticipate and avoid situations where there is a likely chance I’ll have to think in depth about something.*
27. I prefer watching educational to entertainment programs.
28. I think best when those around me are very intelligent.
29. I prefer my life to be filled with puzzles that I must solve.
30. I would prefer complex to simple problems.
31. Simply knowing the answer rather than understanding the reasons or the answer to a problem is fine with me.*
32. It’s enough for me that something gets the job done; I don’t care how or why it works.*
33. Ignorance is bliss.*
34. I enjoy thinking about an issue even when the results of my thoughts will have no outcome on the issue.

*=item is reverse worded

Answer format: 1 extremely uncharacteristic - 5 extremely characteristic

Appendix C: Need to Evaluate Scale 1.7

The Need to Evaluate Scale (Jarvis and Petty 1996) is a scale designed to measure individual differences in the propensity to engage in evaluation. The list of 16 items is presented below.

1. I form opinions about everything.
2. I prefer to avoid taking extreme positions.*
3. It is very important to me to hold strong opinions.
4. I want to know exactly what is good and bad about everything.
5. I often prefer to remain neural about complex issues.*
6. If something does not affect me, I do not usually determine if it is good or bad.*
7. I enjoy strongly liking and disliking new things.
8. There are many things for which I do not have a preference.*
9. It bothers me to remain neutral.
10. I like to have strong opinions even when I am not personally involved.
11. I have many more opinions than the average person.
12. I would rather have a strong opinion than no opinion at all.
13. I pay a lot of attention to whether things are good or bad.
14. I only form strong opinions when I have to.*
15. I like to decide that new things are really good or really bad.
16. I am pretty much indifferent to many important issues.*

*=item is reverse worded

Answer format: 1 extremely uncharacteristic - 5 extremely characteristic

2 Design of Web Questionnaires: The Effects of the Number of Items per Screen

ABSTRACT This chapter analyzes the effects of an experimental manipulation of the number of items per screen in a web survey with forty questions aimed at measuring arousal. We consider effects on survey answers, item non-response, interview length, and the respondents' evaluation of several aspects of the survey (such as layout). Four different formats were used, with one, four, ten, and forty items and headers on a screen. We also analyze how the design effects vary with personal characteristics. We found no effect of format on the arousal index, but we found that item non-response increased with the number of items appearing on a single screen. Having multiple items on a screen shortens the duration of the interview, but negatively influences the respondent's evaluation of the questionnaire layout. Grouping effects are generally similar for different demographic groups, though there are some differences in magnitude and significance level. For example, grouping affects item non-response for men, older respondents, and respondents with low education.

NE QUID NIMIS

2.1 Introduction

In web surveys, the questionnaire format is usually chosen for reasons that have little to do with increasing the accuracy of the answers or reducing item non-response. For example, matrix questions are frequently used either to save space on the screen or to reduce the number of screens. But does the use of a particular questionnaire format affect the answers to the survey questions? Does it have an effect on item non-response or interview duration? And do respondents evaluate various formats differently?

The existing literature provides evidence that words and graphics combine in ways that influence how people answer survey questions (e.g. Couper et al. 2000; Dillman et al. 2006; Lozar Manfreda et al. 2002b; Sanchez 1992), but little is known about the nature of the effects and the underlying mechanisms.

This chapter discusses the impact of one feature of web survey design: the number of items presented on each single screen. Some existing studies are described in Section 2.2. A questionnaire of forty items aimed at constructing an arousal scale is presented to respondents in a web survey that is broadly representative of the Dutch population. The experimental design is described in Section 2.3. The format is randomized: each respondent is presented one, four, ten, or all forty questions on one screen. We then analyze the effect of the questionnaire format on the mean of the forty answers (the arousal score) and the variance (Section 2.4.1), item non-response (Section 2.4.2), the time the respondent needs to complete the questionnaire (Section 2.4.3), and the respondent's evaluation of the questionnaire (Section 2.4.4). In addition, we investigate if any affects are tied to personal characteristics (Section 2.5), an issue that is at the frontier of web survey methodology (Dillman 2007; Stern et al. 2007). Section 2.6 concludes.

2.2 Background

The interface design of web questionnaires varies in terms of how questions are grouped per screen and of the navigation methods used. At one end of the design continuum are form-based designs that present questionnaires as one long form in a scrollable window, at the other end are screen-by-screen questionnaires that present only a single item at a time (Norman et al. 2001). Schonlau et al. (2002) suggest that scrolling can be a burden to respondents and that lengthy web pages can give the impression that the survey is too long to complete. On the other hand, scrolling can also reduce completion times, help preserve the context of items, and avoid the problem of arbitrary page breaks. An advantage of a screen-by-screen design is the focus on each single item, but a disadvantage is the loss of context and the large number of mouse clicks (or other actions) needed to navigate through the survey. Presenting questions in a matrix may be a good compromise: it reduces the number of screens without the need for scrolling.

The visual layout of the scale is an important source of information that respondents use when deciding which answer to select (Christian 2003). Tourangeau et al. (2004); (2007) argue that respondents use several visual heuristics to interpret a question. For example, items that are close to each other are seen as similar. Grouping principles from Gestalt Psychology also address this issue. The Law of Proximity (placing objects closely together will cause them to be perceived as a group), but also the Laws of Similarity (objects sharing the same visual properties will be grouped together) and Common Region (elements within a single closed region will be grouped together) suggest that grouping items on one page affects the answering process (Dillman 2007).

Items are more likely to be seen as related if grouped on one screen, reflecting a natural assumption that blocks of questions bear on related issues, much as they would during ordinary conversations (Schwarz 1996; Sudman et al. 1996). Couper et al. (2000) concluded that correlations are consistently higher among items appearing together on a screen than among items separated across several screens, but the effect is small and differences between pairs of correlations are insignificant. Tourangeau et al. (2004) replicated the above findings and found significant differences between correlations. They concluded that respondents apparently use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. Peytchev et al. (2006) found no significant differences between measurements using a paging and a scrolling design. Bradlow and Fitzsimons (2001) found that when items are not labeled or clustered, respondents base their responses on the previous item to a greater degree, regardless of whether the items are intrinsically related.

According to Bowker and Dillman (2000), the questionnaire format may increase partial non-response if respondents dislike it so much that they fail to complete the survey. Lozar Manfreda et al. (2002b) find no evidence of differences in partial non-response between one and multiple page designs, but they do find that a one-page design results in higher item non-response.

Interview duration may affect response rates, as well as the willingness to cooperate in (future) surveys (Deutskens et al. 2004). Couper et al. (2000), Lozar Manfreda et al. (2002b), and Tourangeau et al. (2004) all find evidence that a multiple-item-per-screen design takes less time to complete than a one-item-per-screen design. Moreover, Deutskens (2006) finds that respondents who are highly motivated have a higher willingness to participate in (follow-up) surveys, and that one of the strongest factors driving motivation is how much the respondent enjoyed answering the questions. Particularly in panel surveys where reducing attrition is an important concern, this makes it worthwhile to analyze the effect of questionnaire format on how the respondent evaluates the attractiveness of the survey.

Visual design theory hardly makes any reference to respondent characteristics (Dillman 2007) and few empirical studies have analyzed how the effects of questionnaire format vary with respondent characteristics. Tourangeau et al. (2007) find no systematic variation in the impact of the layout of a response scale in relation to gender, age, or education group. Stern et al. (2007) also show that the layout of survey questions affects different demographic groups in similar ways. Krosnick and Alwin (1987), on the other hand, find that using different categories affects respondents with less education and less extensive vocabularies more than other groups. Knauper et al. (2004) and Borgers et al. (2004) find that the generally poorer memory of older adults' results in increased design effects. Deutskens et al. (2004), Dillman et al. (2000), and Stern et al. (2007) all conclude that further research is needed on the effects of questionnaire format for different populations.

2.3 Design and implementation

The experiment was conducted in the CentERpanel: an online household panel consisting of more than 2,000 households, administered by CentERdata. The panel aims to be representative of the Dutch-speaking population in the Netherlands, including those without Internet access. The households that do not have access to Internet when recruited are provided with a so-called Net.Box, enabling a connection via a telephone line and a television set. If the household does not have a television, CentERdata will provide that too. The recruitment of new panel members is performed in three stages. In the first stage, a random sample of potential panel members is interviewed by telephone. The interview ends with the question whether the person would like to participate in survey research projects. If so, the household is included in a database of potential panel members. If a participating household drops out of the panel, a new household is selected from the database of potential panel members. This is done on the basis of demographic characteristics, such that the panel will remain representative (see www.centerdata.nl/en/CentERpanel for details).

Our experiment compared several layout options for a questionnaire of 40 items based on a scale developed by Mehrabian and Russell (1974) to measure arousal. All items were answered on a 5-point Likert scale (totally disagree - totally agree).

We divided respondents randomly into seven groups. The first group answered each item on a single screen (see Appendix A, format 1). The second group answered four items per screen (format 2, using less than half of the screen), while the third group answered ten items per screen (format 3, using the whole screen). The fourth group answered all 40 items on one single screen (scrollable design, see Appendix A, format 4).

Because of the height of the screen, people with 40 items per screen had to scroll in order to fill in all the items. Scrolling down made it impossible to see the header (totally disagree-totally agree) when this was only shown at the top of each screen. We formed three additional groups, with four, ten, and forty items on one screen, but with a header displayed at each item (resulting in four, ten, and forty headers per screen, as opposed to one header per screen). For most of the analyzes we combined the single header and multiple header per screen formats[1], as we found hardly any differences between them. Wherever we did find differences, these will be discussed explicitly. We therefore speak of four different formats. Note that the exact appearance of the questionnaire on the respondent's screen depends on screen settings (e.g. height and resolution). In particular, what respondents with a Net.Box see is somewhat different from what others see. Still, we did not find significant differences between Net.Box and other respondents.

The experiment was fielded in December 2004; 2565 respondents (69% of selected panel members) completed the questionnaire. See Table 2.1 for the number of respondents in each format.

Table 2.1: Number of Respondents Who Completed the Questionnaire for the Different Formats

Format		N		Consists of:	N
1	1-item-per-screen	352			
2	4-items-per-screen	727	a	single header	353
			b	multiple headers	374
3	10-items-per-screen	768	a	single header	370
			b	multiple headers	398
4	40-items-per-screen	718	a	single header	359
			b	multiple headers	359
	Total	2565			

Results 2.4

We looked at the effects of questionnaire format on the outcome variables (Section 2.4.1), item non-response (Section 2.4.2), duration of the interview (Section 2.4.3), and respondents' subjective evaluations of the questionnaire (Section 2.4.4). Effects of personal characteristics are presented in Section 2.5.

[1] As a result, the one-item-per-screen format has half the number of respondents compared to the other formats.

2.4.1 Effects on the outcome variables

To summarize the forty outcome variables, we considered the mean and the variance of the forty item scores, encoding the answers 1 (totally disagree), 2, 3, 4 or 5 (totally agree) (or, for reverse worded items, the reverse). The mean is Mehrabian and Russell (1974) arousal score and the variance can be seen as a measure of internal consistency. We were able to reject neither the null hypothesis that questionnaire format has no effect on the mean score (F=1.84; p=0.14), nor the null hypothesis that format has no effect on the variance (F=1.34; p=0.26).

The existing literature (Section 2.2) suggests that grouping items on one screen may increase the correlations among the answers. We find small differences between inter-item correlations when the items were presented (1) 1-item-per-screen (Cronbach's alpha of .8801), (2) 4-items-per-screen (alpha of .8849), (3) 10-items-per-screen (alpha of .8871), or (4) all-items-per screen (alpha of .8788). Values of Cronbach's alpha for the sets of items that were placed on one screen (e.g. 10 times for 4 items or 4 times for 10 items) revealed no differences between formats either.

2.4.2 Effects on item non-response

Based on the results of Lozar Manfreda et al. (2002b), we expect item non-response to increase as more items are placed on a single screen. We analyze both the number of missing items per respondent (with a mean of 0.20 in the complete sample) and the dummy variable that distinguishes respondents with no missing items (0) from those with at least one missing item (1; 15% of the complete sample). Table 2.2 shows the results.

In line with existing findings, we find a positive and significant effect of the number of items per screen on item non-response. Not only does the number of missing items increase as more items appear on a single screen, but so does the probability that at least one answer is missing.

2.4.3 Effects on the time needed to complete the interview

When items are presented on a grid, a respondent has to perform fewer physical actions (mouse clicks) than when items are presented on separate screens. Existing studies show that presenting multiple items per screen shortens the duration of the interview, so we expect a negative effect of the number of items per screen on the time it takes to complete the questionnaire. We indeed found a significant and monotonically decreasing effect (F=4.20, p=.006): the duration was longest for the one-item-per-screen format (median[2]=384 seconds), followed by the 4-items-per-screen format (median=316 seconds), the 10-items-

[2] We present median scores because the distribution of response times is skewed.

Table 2.2: Regression Results on Dummies for the Different Formats (Format 1 is taken as reference); (A) Linear Regression of the Number of Missing Items; (B) Logit Regression of the Dummy Indicating Zero (0) or More Missing Items(1)

A) Linear regression			
		Coefficient	p-value
	constant	0.142	<0.01
Format	2: 4-items-per-screen	0.027	0.52
	3: 10-items-per-screen	0.077	0.06
	4: 40-items-per-screen	0.103	0.01

B) Logit regression			
		Coefficient	p-value
	constant	-1.07	<0.01
Format	2: 4-items-per-screen	0.058	0.70
	3: 10-items-per-screen	0.231	0.11
	4: 40-items-per-screen	0.454	<0.01

per-screen format (median=303 seconds), and the all-items-per-screen format (median=302 seconds).

Effects on the evaluation of the questionnaire 2.4.4

At the end of the questionnaire, people answered some evaluation questions:

1. How interesting did you find the questions?
2. How easy was it to answer the questions?
3. How clear did you find the wording of the questions?
4. What did you think of the layout?
5. How would you evaluate the duration?
6. What is your overall opinion of these questions?

These questions were asked on a ten-point scale ranging from 1 ('very bad'/'not at all) to 10 ('very good'/'very much').

For the evaluation of the layout, we found a significant difference between the four formats (F=17.13, p<0.001). The more items appeared on a single screen, the lower the evaluation (means were 7.67, 7.57, 7.37, and 7.06 for the one, four, ten, and all-items-per-screen format, respectively).

Here, we found a significant effect of the number of headers. For three formats, the 10-items-per-screen version with 10 headers, and the all-items-per-screen version with 1 and 40 headers, respondents had to scroll in order to see the complete screen[3]. A one-way between-groups analysis of variance showed there were no significant differences among these three, but these three were evaluated significantly worse than the other formats, suggesting that scrolling is considered unattractive.

2.5 Effects for different demographic subgroups

This section describes how the effects of grouping items on a screen differ across demographic groups, defined by education, age, and gender. We consider the education groups low (at most lower vocational training), intermediate (senior high or vocational community college), and high (vocational college or university). Age is divided into young and old with a cut-off point at 50 years of age (distinguishing groups of approximately equal size). In the discussion below we consider one demographic characteristic at the time. We also looked at demographic groups characterized by more than one characteristic (e.g. low educated women) but this did not lead to additional insights.

Effects on the outcome variables

We found an insignificant effect of questionnaire format on the mean arousal score and the within respondent variance of the 40 items for each of the demographic groups, in line with the insignificant results for the complete sample (cf. Section 2.4.1).

Effects on item non-response

The fraction of men with at least one missing item differed significantly between formats, as can be seen in Table 2.3. Women did not show significant differences between formats. The table also suggests that reduction in cognitive functioning due to aging leads to higher item non-response rates. More importantly for our study, the probability that at least one answer is missing increased significantly with the number of items per screen for older respondents but not for the younger group. Low educated respondents showed differences between formats for both the number of items not responded to as well as the probability that at least one answer was missing, while respondents with intermediate or high education levels did not show significant differences. These

[3] For the ten items per screen version with a single header, the need to scroll depended on the screen resolution.

Table 2.3: Effects of Grouping Items on a Screen per Demographic Group

	Gender		Age		Education level		
	Men	Women	<50 years	>49 years	Low	Inter-mediate	High
Number of items not responded to							
M1	.156	.129	.085	.217	.152	.138	.127
M4	.156	.184	.123	.22	.136	.163	.202
M10	.23	.207	.139	.333	.267	.214	.166
M40	.227	.265	.177	.334	.322	.239	.174
N	1319	1246	1442	1123	852	838	870
F-statistic	1.28	2.14	1.46	2.17	3.41	1.06	.42
p-value	.28	.09	.22	.09	.02	.37	.74
Fraction of respondents with at least one item missing							
M1	.266	.246	.21	.316	.248	.276	.236
M4	.253	.282	.246	.29	.318	.273	.221
M10	.293	.311	.245	.384	.318	.333	.251
M40	.36	.341	.288	.434	.39	.38	.285
N	1319	1246	1442	1123	852	838	870
F-statistic	3.83	2.01	1.59	5.69	2.67	2.54	.96
p-value	.01	.11	.19	<.001	.05	.06	.42
Duration of the interview (median time in seconds)							
M1	382	387	332	467	363	352	377
M4	320	313	272	380	307	303	312
M10	320	297	260	382	297	299	292
M40	312	288	241	399	285	291	272
N	1295	1222	1408	1109	836	822	855
F-statistic	1.31	3.97	2.51	1.66	1.15	1.35	2.17
p-value	.27	<.01	.06	.17	.33	.26	.09
Evaluation score of the layout							
M1	7.63	7.72	7.51	7.89	7.71	7.66	7.63
M4	7.56	7.58	7.42	7.72	7.6	7.6	7.5
M10	7.35	7.39	7.11	7.73	7.47	7.35	7.3
M40	6.99	7.14	6.72	7.51	7.18	7.08	6.95
N	1311	1237	1430	1118	846	831	866
F-statistic	1.46	6.83	16.07	2.75	4.56	5.76	6.84
p-value	<.001	<.001	<.001	.04	<.01	<.01	<.001

Note: M1=mean score for the 1-item-per screen format, M4=mean score for the 4-items-per screen format, M10=mean score for the 10-items-per screen format, M40=mean score for the all-items-per screen format.

analyzes all suggest that poorer cognitive functioning are associated with larger design effects.

Effects on the time needed to complete the interview

We found a significantly negative effect of the number of items per screen on interview length for women. Respondents aged 50 and older showed longer interview times than their younger counterparts, but there were no significant differences between formats for the two age groups we considered. The design effects for the three education groups were not significant either.

Effects on the evaluation of the questionnaire

With regard to the self-evaluation questions we found differences between demographic groups in the questions "How clear did you find the wording of the questions?" and "What did you think of the layout?". For the former, we only found significant design effects for men (not shown in the table) - men found the question wording in the 1-item-per-screen format clearer than in the other formats (F=2.92, p=.03; the means were 7.95, 7.65, 7.77, and 7.65 for the one, four, ten, and all-item-per-screen format, respectively). For the evaluation of layout, the number of items per screen was significant in each of the demographic groups, as can be seen in Table 2.3.

In all, we found evidence that the importance of the number of items per screen differs between demographic groups. Men seem to be more sensitive for it than women, and the effects on item non-response are associated with lack of cognitive skills.

2.6 Discussion and Conclusions

In this chapter we have investigated how the number of items placed on a single screen (one, four, ten, or forty) in a web survey influences answers, item non-response, interview length, and questionnaire evaluation. We extend previous research by (1) analyzing if any effects of grouping depend on personal characteristics, (2) varying the number of headers (one header per screen versus a header at each item), and (3) adding evaluation questions to find out how respondents experience each format.

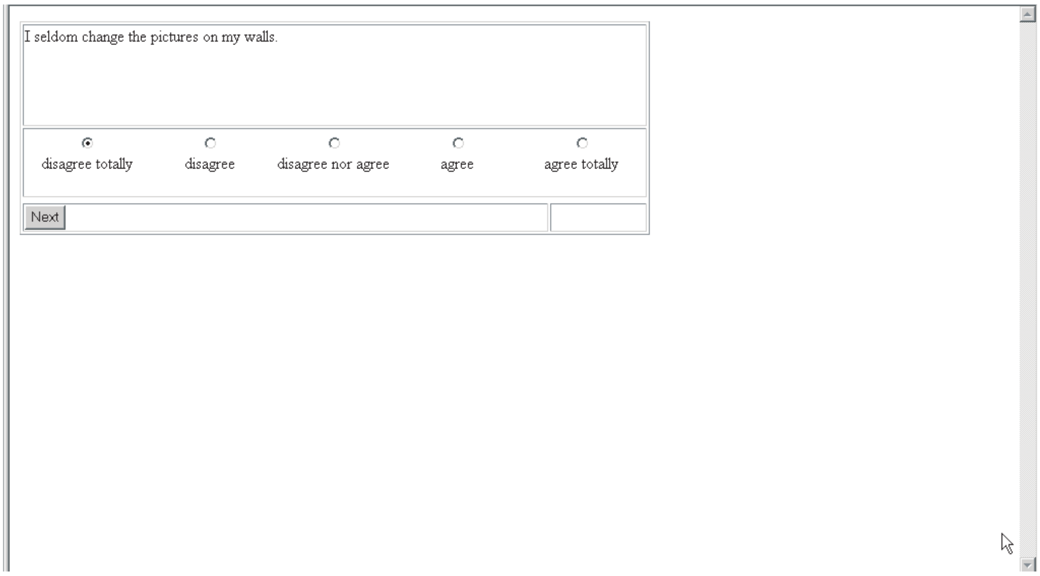
We found no positive effects of adding more headers. In fact, the only effect of adding headers at each item is negative: if adding headers forces the respondents to scroll, this negatively affects their evaluation of the survey. Apparently respondents remember the header on top of the screen and repeating it is not useful.

Reassuringly, we find no evidence that either the number of headers or the number of items per screen lead to different answers to questions or to differences in the arousal scales constructed from these answers. On the other hand, however, we find that putting more items on a screen increases item non-response, reduces the duration of the survey, and makes subjective assessments of the questionnaire less positive. These effects are generally similar for different demographic groups, though there are some differences in magnitude and significance level. For example, grouping affects item non-response for men, older respondents, and respondents with low education.

All in all, the optimal number of items per screen requires a trade- off: more items per screen shortens survey time but reduces data quality (item non-response) and respondent satisfaction (with potential consequences for motivation and cooperation in future surveys). Since the negative effects of more items per screen mainly arise when scrolling is required, in conclusion we are inclined to recommend placing four to ten items on a single screen, avoiding the necessity to scroll.

2.7 Appendix A: Screen shots

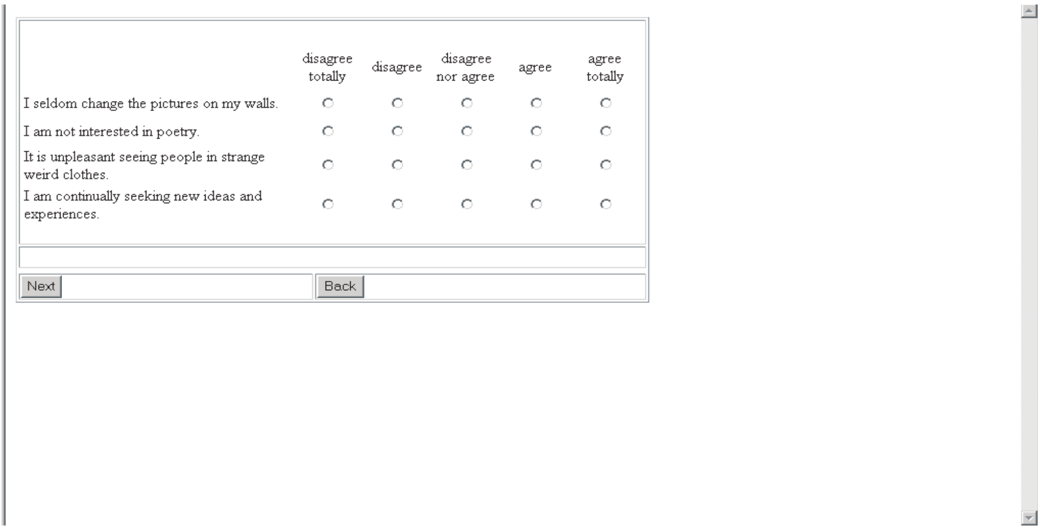
This appendix presents screen shots for the formats as were used in the experiment.



I seldom change the pictures on my walls.

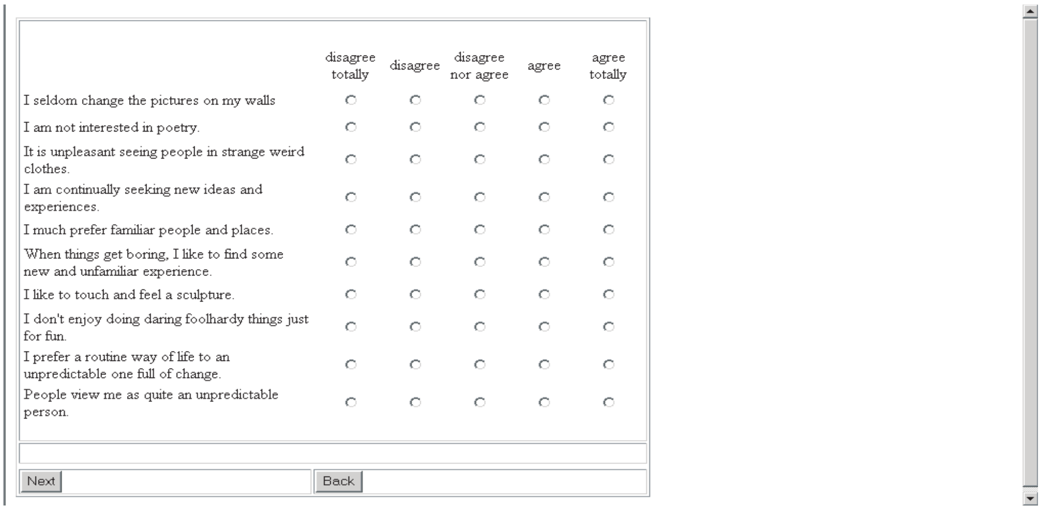
●	○	○	○	○
disagree totally	disagree	disagree nor agree	agree	agree totally

Figure 2.1: Format 1: 1-item-per-screen



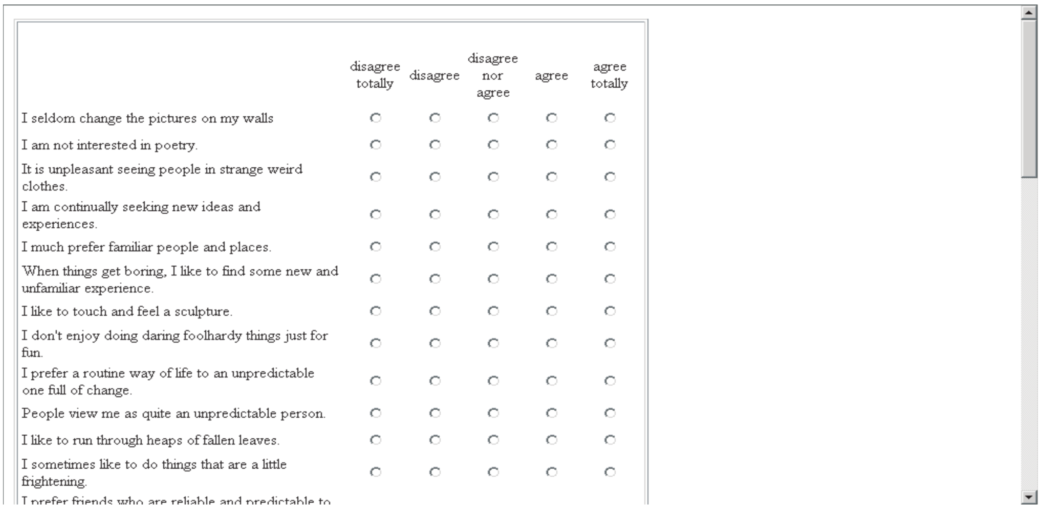
	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls.	○	○	○	○	○
I am not interested in poetry.	○	○	○	○	○
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
I am continually seeking new ideas and experiences.	○	○	○	○	○

Figure 2.2: Format 2: 4-items-per-screen



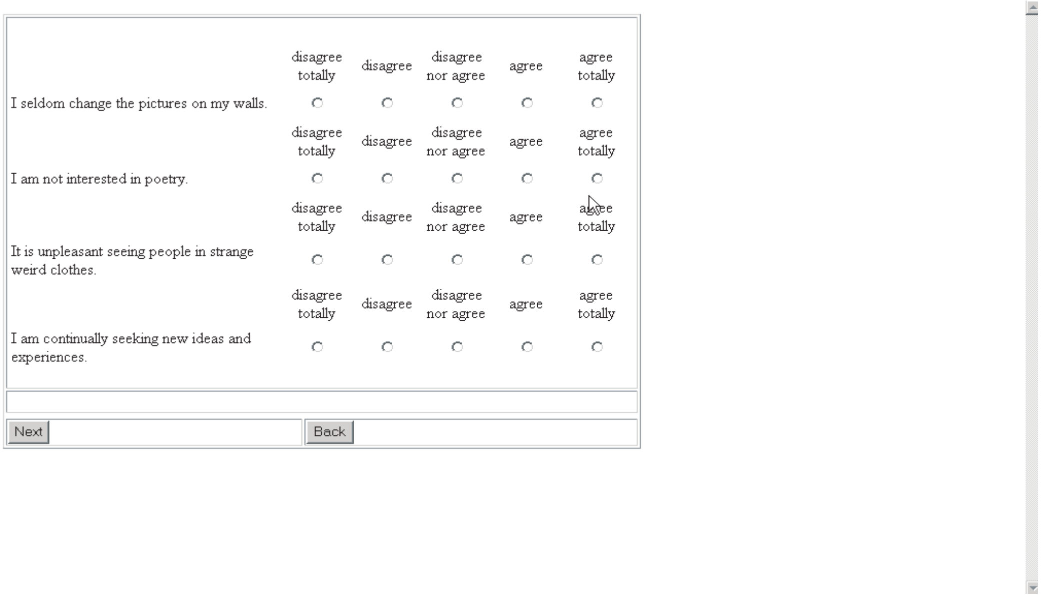
	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls	○	○	○	○	○
I am not interested in poetry.	○	○	○	○	○
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
I am continually seeking new ideas and experiences.	○	○	○	○	○
I much prefer familiar people and places.	○	○	○	○	○
When things get boring, I like to find some new and unfamiliar experience.	○	○	○	○	○
I like to touch and feel a sculpture.	○	○	○	○	○
I don't enjoy doing daring foolhardy things just for fun.	○	○	○	○	○
I prefer a routine way of life to an unpredictable one full of change.	○	○	○	○	○
People view me as quite an unpredictable person.	○	○	○	○	○

Figure 2.3: Format 3: 10-items-per-screen



	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls	○	○	○	○	○
I am not interested in poetry.	○	○	○	○	○
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
I am continually seeking new ideas and experiences.	○	○	○	○	○
I much prefer familiar people and places.	○	○	○	○	○
When things get boring, I like to find some new and unfamiliar experience.	○	○	○	○	○
I like to touch and feel a sculpture.	○	○	○	○	○
I don't enjoy doing daring foolhardy things just for fun.	○	○	○	○	○
I prefer a routine way of life to an unpredictable one full of change.	○	○	○	○	○
People view me as quite an unpredictable person.	○	○	○	○	○
I like to run through heaps of fallen leaves.	○	○	○	○	○
I sometimes like to do things that are a little frightening.	○	○	○	○	○
I prefer friends who are reliable and predictable to					

Figure 2.4: Format 4: all-items-per-screen (with the scroll bar on the right-hand side)



	disagree totally	disagree	disagree nor agree	agree	agree totally
I seldom change the pictures on my walls.	○	○	○	○	○
	disagree totally	disagree	disagree nor agree	agree	agree totally
I am not interested in poetry.	○	○	○	○	○
	disagree totally	disagree	disagree nor agree	agree	agree totally
It is unpleasant seeing people in strange weird clothes.	○	○	○	○	○
	disagree totally	disagree	disagree nor agree	agree	agree totally
I am continually seeking new ideas and experiences.	○	○	○	○	○

Figure 2.5: Multiple headers per screen (for format 2: 4-items-per-screen)

3 Design of Web Questionnaires: An Information-Processing Perspective for the Effect of Response Categories

ABSTRACT In this chapter an information-processing perspective to explore the impact of response categories on the answers respondents provide in web surveys is used. Response categories have a significant effect on response formulation in questions that are difficult to process, whereas in easier questions (where responses are based on direct recall) the response scales have a smaller effect. In general, people with less cognitive sophistication are more affected by contextual cues. The Need for Cognition and the Need to Evaluate indexes for motivation account for a significant part of the variance in survey responding. Interactions of ability to process information and motivation combine in regulating a response for questions that are more difficult to process. Our results hint at a substantial role of satisficing in web surveys.

MEDIO TUTTISSIMUS IBIS

3.1 Introduction

Response categories in survey questions are often chosen on the basis of the knowledge or intuition of the researcher. Studies about the cognitive and communicative processes underlying question answering in surveys suggest that the choice of response categories can have a significant effect on respondent answers (Krosnick and Alwin 1987; Rockwood et al. 1997; Schwarz et al. 1985; Schwarz and Hippler 1987; Strack and Martin 1987; Winter 2002a; Winter 2002b).

Based on a social information processing model proposed by Bodenhausen and Wijer (1987), Schwarz and Hippler (1987) argue that respondents use the response alternatives to determine the meaning of the question and use the frequency range suggested by the response alternatives as a frame of reference, extracting information about presumably common answers from the values stated in the scale. Respondents are more likely to be affected by response categories in frequency questions when they need to estimate because the behavior is not well presented in memory or when they resort to estimation strategies because they lack the motivation to recall.

This chapter replicates parts of previous studies on response categories by Schwarz et al. (1985) and Rockwood et al. (1997) and extends the existing literature in the following directions. First, an open-ended question format as a benchmark is added, as suggested by Rockwood et al. (1997). Second, our study uses a full range of question possibilities that vary in the difficulty of information processing. The literature suggests that response categories have a significant effect on response formulation in mundane and regular questions (questions for which estimation is likely to be used in recall and that refer to an event occurring regularly), whereas in salient and irregular questions (questions in which direct recall is used in response formatting and the event occurs episodically) the response categories do not have a significant effect. We use four questions that vary in the mundane-salient and regular-irregular dimensions to test if differences in difficulty of information processing influence response category effects. Our third contribution to the existing literature is that a heterogeneous sample is used, drawn from a broad population. Most of the previous studies on category effects used convenience samples (like a group of students). The heterogeneous nature of our sample makes it possible to measure the effect of differences in respondent's ability to process information on survey responses and category effects. To indicate a respondent's motivation for giving correct answers, indexes for the respondent's need to think and need to evaluate are included in the analysis. Fourth, previous studies used paper and telephone as modes of administration, while we consider response category effects in an online web survey. Despite the enormous use of web questionnaires, the knowledge of what people read and comprehend and why, is still in its infancy (Redline et al. 2003). Although a theory of web questionnaire design may draw from the principles for visual layout and design of paper questionnaires, it will also have new features (e.g use of screen/mouse) and require independent testing and evaluation (Dillman et al. 1998).

The remainder of this chapter is organized as follows. Section 3.2 discusses the background of the subject. Section 3.3 presents the design and implementation of the study. Section 3.4 shows the results, and Section 3.5 closes with concluding remarks.

Background 3.2

Response effects 3.2.1

To find out if response categories influence respondent behavior, we need to know how respondents answer questions. Trying to understand how respondents comprehend survey questions leads inevitably to a more basic search for cognitive processes involved in answering questions. Interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, formatting a response, and editing it are the main psychological components of a process that starts with respondent's exposure to a survey question and ends with their report (Sudman et al. 1996). Respondents may shortcut the cognitive processes necessary for generating the optimal answer, compromising one or more of these steps. These shortcuts are directed by cues in the questionnaire, such as words and visual stimuli. Since most of the answers that are recorded in surveys reflect judgments that respondents generate on the spot in the context of the specific interview, response effects are likely to emerge.

One of the most basic decisions a survey designer has to make is whether to use open or closed questions. From a cognitive perspective, open questions present a free-recall task to respondents whereas closed questions present a recognition task. Closed questions may fail to provide an appropriate set of meaningful alternatives in substance or wording. Furthermore, respondents are influenced by the specific closed alternatives given. One can expect an answer closer to the correct value if the respondent must produce an answer himself (Schuman and Presser 1981). Schwarz (1996) and Schwarz et al. (1985) recommend asking behavioral frequency questions in an open response format. However, open questions also have disadvantages. They may be affected by rounding (Tourangeau et al. 2000), and are therefore affected by estimation strategies as well. Furthermore, respondents find open questions more difficult to answer, so that item non-response tends to be higher compared to closed questions (Griffith et al. 1999; Hurd et al. 1998), more people abandon the survey (Crawford et al. 2001), and unit non-response is also higher if a survey consists of open-ended rather than closed questions with the same content (Blumberg et al. 1974).

In a closed question respondents are asked to give their opinion by checking the appropriate value from a given set of response alternatives. Schwarz (1996) argues that this given range may serve as a source of information to the respondent. A respondent assumes that the researcher constructed a meaningful scale that reflects knowledge about the distribution of the actual values. Values in the middle range of the scale are assumed to reflect 'average' values, whereas the extremes of the scale are assumed to correspond to the extremes of the distribution. Therefore, giving a response is the same as locating one's own position in the distribution. The more ambiguous the question is defined, the more pronounced is the impact of the response alternatives. But even when the behavior at target is well defined, the range of response alternatives may affect respondents' estimates. Watching television, for example, is not presented in memory as a distinct episode but the various episodes go together in a more generic presentation of the behavior that lacks temporal markers. When asked how often a respondent watches television, respondents therefore cannot recall the episodes to determine the frequency of the behavior. Instead, they rely on estimation strategies. Respondents may not even try to recall, but rather use their biographical knowledge to locate themselves in the distribution suggested by the response scale. For example, a respondent who considers himself an 'average TV viewer' may select a response category in the middle part of the response scale without reviewing his actual TV consumption. Or a respondent may be reluctant to select a response category that seems unusual in the range of responses. This results in higher estimates along scales that present high rather than low response alternatives.

Krosnick et al. (1996) give a cognitive explanation of response effects. Their theory assumes that most respondents answer survey questions by choosing the first satisfactory or acceptable response alternative rather than select the true answer. The tendency to satisfice depends on three things: (1) the difficulty of the question, (2) the respondent's ability to retrieve, process and integrate information from memory, and (3) the respondent's motivation. While the first depends on the question itself, the latter two depend on the respondent's personal characteristics. Whether or not a behavior is well represented in memory can depend on a host of variables like characteristics of the behavior (e.g. Menon et al. 1995) and the importance of the behavior for the respondent (e.g. Knauper et al. 2004). Motivation in answering survey questions can depend on a respondent's need for cognition and evaluation (Petty and Jarvis 1996). Krosnick et al. (1996) find interactions that support the notion that ability and motivation may combine in regulating a response. We will now discuss the three factors on which the tendency to satisfice depends in more detail.

Tendency to satisfice: difficulty of the question 3.2.2

Research shows that frequency judgments, which by necessity rely on a person's memories, contain difficulties that are not easy to correct. When respondents are asked to report how regularly they do something, they may use one of two strategies to arrive at an answer. If the question refers to activities that occur with a low frequency, such as buying a new car, they may try to recall all instances of that activity. In that case, the accuracy of their reports will depend on the accuracy of their memory. For more regular and mundane activities, such as watching TV, respondents have to provide an estimate of their activity, using whatever information is available to them at the time of judgment. In computing this estimate, they may use the range of the response alternatives as a frame of reference. Subjects tend to overestimate the frequency of irregular events and to underestimate the frequency of events that happen regularly (Schwarz and Hippler 1987; Strube 1987). Menon et al. (1995) find that the range of response alternatives affect frequency reports of moderately regular and irregular behaviors, but not of very regular behaviors. They suggest relevant frequency information was inaccessible for the less regular behaviors, causing respondents to rely on response alternatives as a cue in computing a frequency estimate. Rockwood et al. (1997) conclude that response categories have a significant effect on response formulation in mundane and regular questions (questions for which estimation is likely to be used in recall and the event occurs regularly), whereas in salient and irregular questions (questions in which direct recall is used in response formatting and the occurrence of the event is episodic) the response categories do not have a significant effect. Van der Vaart and Glasner (1999) also found that more difficult recall tasks, which involve less salient, less recent (and to some extent more frequent) behavior, coincide with greater recall error. Although these studies show different results in relation to question types and response effects, the argument that response effects are likely to emerge if information is more difficult to process holds for all studies.

Tendency to satisfice: ability to process information 3.2.3

In addition to behavioral factors, the respondent's ability to process information can influence response effects. In their analysis of order effects, Krosnick and Alwin (1987) find that respondents with less cognitive sophistication are more likely to be influenced by changes in response order. Respondents with less education and more limited vocabularies are influenced more by manipulation of answer categories. Lynch et al. (1991) show that context directly affects the people with less knowledge on the topic in question (novices) than experts. Knauper et al. (2004) find that older respondents are more influenced by frequency scales pertaining to mundane behavior, but are less influenced by frequency scales pertaining to salient behavior.

3.2.4 Tendency to satisfice: motivation

Reliance on the scale to avoid effortful attempts to recall relevant episodes increases with a lack of motivation in answering survey questions. A respondent's need for cognition and need for evaluation can be useful indicators for a respondent's motivation to give correct answers, since they are associated with two steps in the question answering process: retrieving information and generating an opinion.

Cacioppo and Petty (1982) developed a scale to measure the need for cognition. Need for cognition (NFC) represents the tendency for individuals to engage in and enjoy thinking. They reasoned that when respondents are motivated (such as when the topic is of high relevance to the respondent) respondents are more eager to think than when their motivation is low. In their view, not only situational factors determine how much thinking occurs. Individual differences in intrinsic motivation to engage in cognitive activity also affect the effort a respondent is willing to make. People with a high need for cognition (HNC) tend to seek more information and think more carefully before making an evaluation than people with a low need for cognition (LNC), who are more easily influenced by peripheral cues.

Not only a person's need to think can affect motivation (and thereby the effect of response categories) in answering a survey, a person's need to evaluate may also play a role. People who have a pre-consolidated answer are presumably better able to answer correctly than people who have not (yet) formed an opinion on the topic. Jarvis and Petty (1996) developed a measure to assess individual differences in the propensity to engage in evaluation, the Need to Evaluate Scale (NES). One could expect that those with a High Need to Evaluate (HNE) are more likely to have formed attitudes toward objects or situations, and are more motivated to formulate a response, than people with a low need to evaluate (LNE). Evaluation does not require effortful thought. The relation between the NES and the NFC was tested by Jarvis and Petty and was found to be moderate and positive (r=.35, p<.001).

Petty and Jarvis (1996)) suggest that LNCs and LNEs are expected to be more susceptible to various low effort biases than HNCs and HNEs, such as being influenced by cues in a survey that suggest one response over another. Whether or not a respondent formulates an answer based on retrieval (in memory) or construction (building an answer at the time of answering the survey) might also be influenced by the need for cognition and evaluation.

3.2.5 Mode of survey administration and sample

Mode effects exist when answers vary systematically with survey mode (due to differences in information transmitting). Previous work has shown that differences between the modes of survey administration influence respondent's answers (see, for example De Leeuw 2005; Kwak and Radler 2002; Lynn 1991; Tourangeau et al. 2000; Voogt and Saris 2005). Further exploration of differences across (self-report) response formats in different modes of administration is warranted (Rockwood et al. 1997; Thomas and Klein 2006). The influence of the data collection method on respondent's answers has been extensively studied for face-to-face, telephone, and paper- and pencil surveys, but the Internet is a relatively new medium. Although response effects in web surveys may draw from the principles for visual layout and design of paper questionnaires, it will also have new features (e.g., use of keyboard/mouse, screen) requiring independent testing and evaluation. De Leeuw (2005) argues that the Internet is the most dynamic of the modes of administration, allowing for multitasking and quickly skipping from one topic to the next. This may lead to more superficial cognitive processing and more satisficing in responding to survey questions.

Studies on broad samples with a large variation in demographic characteristics are needed to generalize conclusions on response effects (Kwak and Radler 2002). Those with lower educational levels and more limited cognitive skills are more likely to engage in satisficing (Chang and Krosnick 2003). Knauper et al. (2004) find that older respondents are more influenced by frequency scales than young respondents when they need to rely on estimation strategies. Therefore, studies using students as a sample may underestimate response effects. Furthermore, those with low educational levels, older people, and females are less likely to be online, so online access panels may underestimate response effects as well (Thomas and Klein 2006; Kwak and Radler 2002). In all, a web survey presented to a broad sample with a large variation in demographic characteristics gives the opportunity to evaluate response effects in this mode of administration, as well as generalize conclusions to the population.

Design and implementation 3.3

Our study was conducted in the CentERpanel, an online household panel consisting of more than 2,000 households. This panel is representative of the Dutch population and is administrated by CentERdata, Tilburg University (the Netherlands). CentERdata provides a Net.Box to people who do not have a computer to make it also possible for them to complete the questionnaires online. Of the 2924 panel members who were selected, 2393 (81.8%) participated in this particular survey.

Rockwood et al. (1997) conclude that further research on response effects should investigate a full range of question possibilities which vary in the use of memory: regular/mundane, regular/salient, irregular/salient and irregular/mundane. We investigate four questions covering these four cases. Rockwood et al. (1997) also advise not only to use low and high answer categories, but to add a third experimental condition with open-ended questions, since this would greatly improve the understanding of the issues involved with response effects in answer categories.

Our study uses four questions in which the response format was manipulated, with the following topics: hours per day watching television (as used by Rockwood et al. 1997 and Schwarz et al. 1985); number of attended birthday parties per year; number of visits per year to a hairdresser; and days per year on holiday (away from home). The existing literature suggests that response category effects are not the same for all question types. Hours watching TV is a question that is not presented in memory as a distinct episode but the various episodes go together in a more generic presentation of the behavior that lacks temporal markers. Respondents therefore cannot recall the episodes to determine the regularity of the behavior, and have to rely on estimation strategies. On the other hand, questions about a respondent's holiday are well defined and response formation can be based on direct recall. Information processing for the other question types is presumably somewhere in between these two question types. For most people, the frequency with which they visit a hairdresser is regular, e.g. one time per month. Furthermore, because visiting a hairdresser is distinct behavior on one's own initiative, this behavior is stored in memory relatively well. Therefore, we consider visiting a hairdresser as salient behavior. Visiting a birthday party on the other hand is irregular behavior, based on the initiative of another person. Therefore, information is more difficult to process. Of course, whether a behavior is considered regular/irregular or mundane/salient depends on individual frequencies. But since the time period between two consecutive haircuts is less variable than the time period between two consecutive birthday parties, these behaviors clearly differ in their regularity. Visiting a hairdresser is a more distinct behavior than visiting a birthday party (which is more ambiguous in terms of locations and temporal markers), making it easier to extract the requested information from memory. Questions about regular and mundane behavior are more likely to be affected by the choice of response format than questions about mundane and irregular, salient and regular, and salient and irregular activities. Respondents in our experiment were randomly assigned to format A (low response scale), format B (high response scale), or format C (open-ended question). Answer categories were based on the existing literature (Rockwood et al. 1997; Schwarz et al. 1985) for the TV question and on a pilot study for the other questions. See Table 3.1 for the response scales used. As mentioned before, apart from memory, motivation can play a role with regard to response effects. Therefore, we include respondent's need for cognition (NFC) and evaluation (NES) into the analysis. NFC and NES are measured with questions on 34 and 16 items, respectively (see Cacioppo and Petty 1982, and Jarvis and Petty 1996 for the items used). By counting the scores of the items, overall cognition and evaluation scores are derived. Using the mean scores, respondents were divided in a low and a high group for each construct.

Table 3.1: Questions and Response Scales Used in the experiment

Response Scales	Format A	Format B	Format C
How many hours do you typically watch TV?			
1	$\frac{1}{2}$ hour or less	2 $\frac{1}{2}$ hours or less	open-ended question
2	$\frac{1}{2}$-1 hour	$\frac{1}{2}$-3 hours	
3	$1-1\frac{1}{2}$ hours	$3-3\frac{1}{2}$ hours	
4	$1\frac{1}{2}-2$ hours	$3\frac{1}{2}-4$ hours	
5	$2-2\frac{1}{2}$ hours	$4-4\frac{1}{2}$ hours	
6	more than $2\frac{1}{2}$ hours	more than $4\frac{1}{2}$ hours	
How many birthday parties do you typically attend per year?			
1	9 or less	17 or less	open-ended question
2	9-11	17-19	
3	11-13	19-21	
4	13-15	21-23	
5	15-17	23-25	
6	more than 17	more than 25	
How many times did you go to the hairdresser last year?			
1	1 or less	9 or less	open-ended question
2	1-3	9-11	
3	3-5	11-13	
4	5-7	13-15	
5	7-9	15-17	
6	more than 9	more than 17	
How many days did you leave your home (have a holiday) last year?			
1	9 or less	17 or less	open-ended question
2	9-11	17-19	
3	11-13	19-21	
4	13-15	21-23	
5	15-17	23-25	
6	more than 17	more than 25	

Note: answer categories one to five in Format A match answer category one in Format B. Answer category six in Format A matches answer categories two to six in Format B.

Furthermore, based on Schwarz (1996) NFC and NES were combined into four quadrants (see Table 3.2). The first group consists of people who are low in their cognitive activity both in thinking and in evaluating. They are the most likely to be affected by the choice of response format, because they are more easily influenced by peripheral cues. The second group consists of persons who don't like to think but do like to evaluate. They do form opinions but don't think them through. The third group consists of people who do like to think but do not like to evaluate. Their answers are constructed at the time they complete the survey, but they do think about their answers. The last group consists of people with a high need for cognition and a high need to evaluate. These people are expected to be the least sensitive to the response format.

Table 3.2: Different Groups in the Experiment for Need for Cognition (NFC) and Need to Evaluate (NES) and Combination of NFC/NES into Four Quadrants

	Low	High
NFC	(NFC<112*)	(NFC>111*)
	N=688	N=638
NES	(NES<52*)	(NES>51*)
	N=663	N=633
	Low NFC	High NFC
Low NES	Group 1	Group 3
	N=490	N=173
High NES	Group 2	Group 4
	N=198	N=465

**Counting scores on the 34 NFC items and the 16 NES items yield the overall score per person. With a minimum of 53 and a maximum of 157, 111 is the mean score for NFC, and with a minimum of 27 and a maximum of 80, 51 is the mean score for NES.*

3.4 Results

The setup of the section is similar to the background section. When presenting the results of the response effects we will also focus on the three factors on which the tendency to satisfice depends.

3.4.1 Response effects

To assess the impact of the response scale on respondents' reports, the responses in the low response scale (see format A in Table 3.1) and the high response scale (see format B in Table 3.1) were summarized for hours watching TV as either (a) two and a half hours or less, or (b) more than two and a half hours (as in Rockwood et al. 1997 and Schwarz et al. 1985). Based on a pilot study, the low and high response scales for birthday parties and days on holiday were summarized as either (a) 17 or less, or (b) more than 17. For visiting a hairdresser the low and high response scales were summarized as either (a) nine or less, or (b) more than nine. We dichotomized the answer categories to remain consistent with previous research.

We used the open-ended question format as a benchmark, since it does not provide any anchors. We have analyzed the frequency distributions of the open-ended answers, and found some focal points (e.g. 6 and 8 for visits to a hairdresser) suggesting that rounding might occur, but this did not affect the comparison between low and high response scales.

Furthermore, we used information on survey experience (e.g. the number of weeks in the panel) to test for an interaction between survey experience and the effects of response categories on reported frequency of the four activities we consider, but we found no significant interaction. Still, almost none of the respondents are completely fresh to the panel, and it is possible that the effect of panel experience is nonlinear, with a noticeable effect of going from no to some experience but no effect of going from some experience to more experience. If this is the case, we expect that our findings for panel participants with some survey experience are a lower bound on the average effect of response categories for the complete (non-experienced) population.

As expected, the range of the response scale affected respondents' frequency reports, as can be seen in Table 3.3. Only 22.0% of the respondents who got the low response scale reported watching TV for more than two and a half hours, compared to 53.6% of the respondents who got the high response scale. In comparison, 52.1% of the respondents who got the open-ended question reported a TV consumption of more than two and a half hours. Comparing the different conditions, the high response scale apparently best matches the respondent's behavior; while the low response scale versus the high response scale and the low response scale versus an open-ended question show significantly different answers, the high response scale versus open-ended answers do not differ significantly (see Table 3.4). This can be due to the fact that the high response scale categories cover the central part of the distribution of open-ended answers, while the low response scale categories mainly cover unusually low hours for this question.

With regard to birthday parties, all three conditions show significant differences; 25.6% of the respondents in the low response scale attend more than 17 birthday parties a year, compared to 44.6% of the high response scale group and 39.4% of the respondents in the open-ended condition. With the open-ended question in the midst of the low and high response scale, the frequency ranges of the low and high scale both divert answers in relation to the free-recall task.

The question about visiting a hairdresser again shows statistically signifi-

Table 3.3: Overview of Frequencies of the Results from Different Response Formats

	Low response scale		High response scale		Open-ended	
	X* or less	more than X*	X* or less	more than X*	X* or less	more than X*
Mundane and Regular						
Hours Watching TV	78.0%	22.0%	46.4%	53.6%	47.9%	52.1%
Mundane and Irregular						
Hours Watching TV	74.4%	25.6%	55.4%	44.6%	60.6%	39.4%
Salient and Regular						
Hours Watching TV	84.7%	15.3%	72.1%	27.9%	81.5%	18.5%
Salient and Irregular						
Hours Watching TV	53.9%	46.1%	46.6%	53.4%	49.8%	50.2%

*X=two and a half for hours watching TV, nine for visiting a hairdresser, and 17 for birthday parties and days on holiday.

cant differences for the low and high scale conditions. But in this question, the answers on the low response scale are closer to the open-ended answers. For the question on the number of days a respondent spent on holiday, only the difference between high and low response scale respondents is significant.

In summary, the data provide strong evidence that the range of response categories affects respondent reports. Bonferroni corrected joint tests show that the hypothesis that high and low response scales do not lead to different answers for all four questions is rejected (all four p-values are smaller than 0.0125). Similarly, the hypothesis that low response scales and open-ended responses give the same answers is rejected (with two of the four p-values are smaller than 0.0125), as well as the hypothesis that high response scales and open-ended answers give the same answers (one p-value smaller than 0.0125). We find higher frequency estimates along scales that present high rather than low frequency response alternatives. This indicates an anchoring effect, as suggested by Schwarz (1996). All four questions show statistically significant differences in the high versus the low response scale. The open-ended condition is sometimes more similar to one response scale than to the other. How strongly the scale biases a respondent's answer, is influenced by how the scale relates to the population distribution. If the distribution of categories is closer to the distribution of open-ended answers, the influence of response categories is less pronounced.

Tendency to satisfice: difficulty of the question 3.4.2

The impact of response alternatives on behavioral frequency judgments is expected to depend on the regularity and the salience of the behavior. Questions about regular and mundane behavior are expected to be affected more by the choice of response format than questions on behavior that is mundane and irregular, salient and regular, or salient and irregular. Table 3.4 shows an overview of the correlation between answer score and response format for the different question types. A higher correlation coefficient (η) between the answer score and the scale that was used indicates a larger effect of the response scale. With the high versus low response scale, the largest correlation between the answer score and the scale is found in hours watching TV (mundane/regular), followed by the questions on birthday parties (mundane/irregular), visiting a hairdresser

Table 3.4: Overview of Correlations between Answer Score and Response Format per Question Type

	Low response scale versus high response scale		Low response scale versus open-ended		High response scale versus open-ended	
	η	p	η	p	η	p
Mundane and Regular						
Hours Watching TV	.325	.000	.311	.000	.558	.558
Mundane and Irregular						
Hours Watching TV	.199	.000	.148	.000	.052	.037
Salient and Regular						
Hours Watching TV	.152	.000	.042	.095	.112	.000
Salient and Irregular						
Hours Watching TV	.073	.000	.041	.106	.032	.193

Note: A higher correlation coefficient (η) between the answer score and the scale that was used indicates a greater difference between response scales. The p-value (p) relates to testing whether the difference between response scales is significant for a specific question type and response scales.

(salient/regular), and days on holiday (salient/irregular). As expected, the impact of response categories differs across question types. Comparison of the open-ended question with the different response scales shows similar results, although not all comparisons reach statistical significance. All in all, the effect of response scales depends on how well a behavior is presented in memory.

3.4.3 Tendency to satisfice: ability to process information

Table 3.5 presents the correlations between answer scores and response format for the four questions separately for sub samples with different individual characteristics. Men tend to be more affected by contextual cues than women - in three of the four questions they show a larger difference in answer score between the low and the high response scale.

In the mundane/regular question, the age group 15-24 is the least affected by the response scale offered, while the age group 25-34 show the highest correlation. With regard to the mundane/irregular question, the age pattern is different, with respondents in the age of 15-24 showing the highest difference. There is a U-shaped pattern, with a minimum response category effect at age 45. The same goes for the salient/regular question, but there the turning point is at age 35. Binary regression (a logit model with an age dummy, a format dummy, and an interaction of the age and format dummies) shows a significant interaction effect between format and age 35-44 (versus the other age groups). For the salient/irregular question no clear age affect is found.

Based on the literature, we would expect low educated respondents to be more susceptible to response effects. What we found is less clear-cut. Table 3.5 shows that for three out of four questions, the highest response scale effect is found for the intermediate vocational education group. Binary regression shows a significant interaction effect between format and intermediate vocational education (against the other education levels) in both the mundane/regular and mundane/irregular question. In these questions, the primary education level shows a relatively high correlation between answers and response scales. The same goes for the mundane/irregular question. The response scale influences the higher secondary education group the least (and not the highest education level, as we would have expected).

3.4.4 Tendency to satisfice: motivation

Because the existing literature suggests that need for cognition (NFC) and need to evaluate (NES) account for variance in survey responses, we include these in the analysis to indicate motivation effects. Table 3.5 shows the separate construct groups as well as the four quadrants in which we combine NFC and NES.

Table 3.5: Overview of Significance and Association between the Low and High Response Scale per Question Type for Different Personal Characteristics

	Mundane and regular Hours watching TV		**Mundane and irregular** Birthday parties		**Salient and regular** Visiting a hairdresser		**Salient and irregular** days on holiday	
	η	p	η	p	η	p	η	p
Gender								
Male	.331	.00	.191	.00	.165	.00	.099	.01
Female	.316	.00	.222	.00	.138	.00	.046	.19
Age								
15-24	.289	.00	.268	.00	.161	.05	.130	.15
25-34	.378	.00	.208	.00	.133	.02	.040	.47
35-44	.333	.00	.144	.02	.162	.01	.072	.21
45-54	.322	.00	.197	.00	.108	.05	.135	.01
55-64	.297	.00	.184	.00	.105	.10	.005	.94
>64	.313	.00	.225	.00	.241	.00	.066	.30
Education								
Primary	.341	.00	.336	.00	.072	.44	.079	.42
Lower secondary	.326	.00	.194	.00	.178	.00	.115	.02
Higher secondary	.285	.00	.159	.02	.071	.30	.047	.48
Intermediate vocational	.395	.00	.146	.01	.189	.00	.126	.02
Higher vocational	.344	.00	.264	.00	.141	.01	.015	.76
University	.294	.00	.171	.03	.171	.03	.096	.24
NFC								
low	.389	.00	.068	.20	.201	.00	.108	.02
high	.322	.00	.048	.19	.217	.00	.087	.07
NES								
low	.359	.00	.077	.09	.136	.00	.109	.02
high	.322	.00	.043	.49	.278	.00	.087	.07
NFC-NES								
Group 1 *	.409	.00	.136	.01	.151	.01	.102	.07
Group 2	.353	.00	.355	.00	.257	.01	.111	.20
Group 3	.267	.00	.133	.14	.122	.21	.102	.27
Group 4	.306	.00	.249	.00	.149	.01	.081	.15

**See Table 3.2 for the definition of groups.*

Note: A higher correlation coefficient (η) between the answer score and the scale that was used indicates a greater difference between response scales. The p-value (p) relates to testing whether the difference between response scales is significant for a specific demographic group.

In the mundane/regular question, the difference in frequency reports between respondents who were offered the low and high response scales is greater for respondents with a low need for cognition (NFC). Our hypothesis that respondents who score low on the NFC construct are more sensible for context effects is confirmed; binary regressions shows a significant interaction effect between format and NFC. In the mundane/irregular question type, however, we do not find evidence that NFC accounts for differences in response effects. In the salient/regular question type, respondents who do not like to think (LNC) show less response effects. For the salient/irregular question type, respondents with a high NFC are not sensitive to response category effects. The same results are found for the Need to Evaluate construct, although in the salient/regular question binary regression shows a significant interaction effect between format and NES in a different direction. Respondents with a high NES are more affected than respondents with a low NES in this question.

The bottom panel of Table 3.5 combines need for cognition and need to evaluate into four quadrants. For the regular/mundane question, similar results are found as for the separate constructs - people with a low NFC and a low NES show the largest deviation between the low and high response scale (η = .409). The first quadrant consists of people who are low in their cognitive activity both in thinking and in evaluating, who are the most likely to be affected by the choice of response format, because they are more easily influenced by peripheral cues. The second quadrant, consisting of persons who don't like to think but do like to evaluate, shows a lower correlation (η = .353). The third quadrant, with people who do like to think but do not like to evaluate, has the smallest correlation between the different response scales (η = .267). The people with a high NFC and a high NES are more affected by the response scale (η = .306) than people in the third quadrant. In the mundane/irregular question type the deviation scores in the quadrants increase drastically compared to the separate constructs, indicating that the combination of NFC and NES increases the differentiation in response effects. Especially in the quadrants with a high NES (groups 2 and 4) the deviation between the high and low response scale groups is high. Apparently they evaluate on the spot, influenced by peripheral cues. In this question type, people with a high NFC and a low NES (group three) have the most similar results in the different response scales alternatives: differences between the high and low scale now even do not reach statistical significance. Respondents who score low on both constructs, have low variances as well. Looking at the quadrants in the salient/regular question type, especially the people who do not think things through well and who do evaluate a lot show more differences in results between the low response scale and the high response scale. For the salient/irregular question, there are no significant differences between answer scores in the high versus the low response scale for respondents in the different quadrants.

Because the literature suggests that interactions of ability and motivation may at times combine in regulating a response, we added interactions of gender/age/education variables with NFC/NES variables and scale effects. We find interactions for the questions that are more difficult to process: watching TV (mundane/regular) and birthday parties (mundane/irregular). With regard to birthday parties, young people and respondents with low education levels (who report more birthday parties than older people and respondents with high education levels) show smaller differences between high NFC/NES and low NFC/NES with regard to scale effects. For watching TV (mundane/regular), women (who report watching TV more frequent than men) show smaller differences between high NFC/NES and low NFC/NES. In other words, for respondents who report higher frequencies, motivation has a smaller effect on survey responses. Motivation apparently helps when memory representation is bad; when memory representation is good motivation is not needed to correctly report the behavior.

Discussion and Conclusions 3.5

In this chapter an information-processing perspective to explore the impact of response categories on the answers respondents provide in web surveys is used. We replicate the findings in other modes of administration that response scales are perceived as informative. An extension of this study is that it also uses an open-ended format, avoiding the bias due to response scale anchors. How strongly the scale biases a respondent's answer, is influenced by how the scale relates to the population distribution. If the distribution of categories is closer to the distribution of open-ended answers, the influence of response categories is less pronounced.

Questions about regular and mundane behavior are more affected by the choice of response scale than irregular and mundane, regular and salient, and irregular and salient respectively. Response scales have a significant effect on response formulation in questions that are difficult to process, whereas in easier questions (where responses are based on direct recall) the response scales have a smaller effect. An open-ended format is preferable in questions in which estimation strategies have to be used. If this type of answer format is not desirable (e.g., because of higher item non-response on open-ended answers), categories in closed questions have to be chosen with care.

We have dichotomized the answer categories to remain consistent with previous research. Dichotomization may obscure effects in the data. New effects may be found in case one would analyze the original data without dichotomization. We leave this for further research.

The hypothesis that response category effects differ for respondents with different personal characteristics was confirmed. In general, men are more affected by contextual cues than women. Age and education effects are not as clear-cut as we would have expected. For example, there is no evidence that response effects fall monotonically with education level.

The Need for Cognition and the Need to Evaluate constructs to indicate motivation account for variance in survey responding. In most question types, the deviation in reports between respondents who were offered the low and high response scales is greater for respondents with a low need for cognition. The same goes for need to evaluate.

Interactions of ability to process information and motivation combine in regulating a response for questions that are more difficult to process. For respondents who report the requested behavior more frequent, motivation has a smaller effect on survey responses. Motivation apparently helps when memory representation is bad; when memory representation is good motivation is not needed to report the behavior. When designing questionnaires one should pay particularly attention to response categories in difficult questions (in which estimation strategies have to be used) when respondents are expected to have difficulty in reporting, or lack of motivation.

Our study shows that response category effects are present in web surveys as they are in other modes of administration. Unfortunately, we do not have a comparison condition that allows us to assess if the influence of scales is more or less pronounced in web surveys. The difference between the high and low response scale for hours watching TV is 32% in our survey, while Rockwood et al. (1997) find a difference of 15% and Schwarz et al. (1985) find a difference of 22% for the same question. Although Rockwood et al. (1997) did not find differences in a telephone mode compared to a mail mode (which are very different in information transmitting) these results could indicate a high tendency to satisfice in web surveys, as suggested by De Leeuw (2005). Or, because we used a heterogeneous sample (and previous studies a group of students) our results might hint at an effect of personal characteristics. Unfortunately, our study cannot point out which argument accounts for more satisficing. Therefore, future research is warranted.

4 Design of Web Questionnaires: The Effects of Layout in Rating Scales

ABSTRACT This article shows that respondents gain meaning from non-verbal cues in a web survey as well as from verbal cues (words). We manipulated the layout of a five point rating scale in two experiments. First, we compared linear and non-linear formats interrupting the graphical language of the scale. Second, we manipulated the linear layout using verbal, graphical, and numerical language. In addition it is analyzed in which way personal characteristics account for variance in survey responding. Our experiments show differences in responses when the visual language is altered. The elderly and the highly educated are the most sensitive to layout effects.

DEVO RARISSIMA NOSTRO SIMPLICITAS

4.1 Introduction

Ordinal scale questions are probably the most widely used measurement instrument in web surveys. These questions are presented in various ways: answer categories can be presented in (one or more) column(s), with labels for all categories or for the endpoint categories only, with radio buttons or an answer box, etc. It is well-known that differences in layout can lead to substantial differences in responses (Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Schwarz and Hippler 1987; Tourangeau et al. 2004; Tourangeau et al. 2007). Christian et al. (2005) suggest that writing effective questions for web surveys may depend at least as much on the presentation of the answer categories as on the question wording itself.

Researchers have developed a theoretical framework that draws on linguistics and Gestalt psychology to explain how visual design elements influence the question-answering process (Jenkins and Dillman 1997), and a growing body of empirical research now provides a foundation for visual design theory. Four languages for communication are distinguished: verbal, graphical, numerical, and symbolic (Dillman 2007). Despite the growing empirical support, the theory of visual design is virtually without any reference to respondent characteristics (Stern et al. 2007). In line with this, empirical tests have not analyzed how questionnaire format effects vary with demographic characteristics of the respondents.

The purpose of this chapter is to find out how visual language in a rating scale influences survey answers and how the effects of visual language vary with the respondent's characteristics, an issue which is at the frontiers of web survey methodology (Dillman 2007; Stern et al. 2007).

4.2 Background

In constructing ordinal scales for self-administered questionnaires, the visual layout of the scale is an important source of information that respondents use when selecting an answer (Christian 2003). Tourangeau et al. (2004); (2007) argue that respondents use several visual heuristics in interpreting a question. Each heuristic assigns a meaning to a visual cue. For example, respondents will see the middle option in a set of response options as the most typical. In addition, they will expect that the answering options are presented in some logical order. Another interpretive heuristic states that with a vertically oriented list, the top option will be seen as the most desirable. Also, visually similar options will be seen as closer conceptually. In addition to these visual heuristics, grouping principles from Gestalt Psychology can be used to understand visual design effects. For example, the Law of Pragnanz states that elements with simplicity, regularity, and symmetry are easier to perceive and remember (Dillman 2007). Presenting the response scale with a layout that is inconsistent with these heuristics and principles results in different responses (see, for example, Christian and Dillman 2004; Smith 1995; Smyth et al. 2006; Tourangeau et al. 2004; Tourangeau et al. 2007). Verbal and nonverbal cues can independently and jointly influence the survey answers. For example, Redline et al. (2003) provide evidence that the visual and verbal complexity of information in a questionnaire affects what respondents read, the order in which they read it, and ultimately, their comprehension of the information. Dillman and Christian (2002) find that manipulating several aspects of the visual languages simultaneously changes respondent behavior significantly.

Verbal language

By changing the verbal orientation of a scale (decremental versus incremental), responses may be altered because of different verbal labels at the first up to the last response option. A primacy effect occurs when options in the beginning of a response list are more often selected, while a recency effect occurs when options near the end of a response list are chosen more often (Krosnick and Alwin 1987). Satisficing occurs when respondents are more likely to choose items earlier in a list because they choose the first response option they consider satisfactory, rather than processing all of them (see Krosnick and Alwin 1987; Krosnick et al. 1996; and Tourangeau et al. 2000, for a detailed description of satisficing).

Orientation effects impute to the position of a response option by itself, but can also be due to a change in perceived intensity of the verbal label, resulting from another position of this label on the scale. The perceived intensity of a verbal label on position x may differ from that of the same verbal label placed on position y. Both the meaning of the verbal label and its position on the scale, can therefore influence the appraisal made by respondents (Hofmans et al. 2007).

Research on orientation effects in rating scales is inconclusive. While in some studies respondents altered their responses when the orientation of a scale is changed, in other studies responses remained unaffected (Weng and Cheng 2000).

Graphical language

Graphical language communicates visual features such as size, space, and location of information on a page. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales do not elicit the same responses. However, the direction of the difference varied across items. Their results are thus inconclusive and future research is warranted. A non-linear layout (where options are presented in multiple rows and columns) can also result in different responses compared to a linear layout because the graphical language conveying the scale is interrupted (Christian 2003; Christian and Dillman 2004).

Numerical language

Numbers in answer categories contain numerical language. Schwarz et al. (1985) have shown that respondents gain information about the researcher's expectations using the numerical labels on a scale as frames of reference. Schwarz et al. (1991) found that changing the numerical values attached to scales resulted in different answers. In particular, respondents hesitate to assign a negative score to themselves (see also Tourangeau et al. 2000, p.248, and Tourangeau et al. 2007). Although negative signs are treated as numerical language in literature, they can also be seen as symbols and therefore convey symbolic language to a scale, which may explain their effect on the responses.

Visual design theory is virtually without any reference to respondent characteristics (Dillman 2007). As a result, the empirical tests have not analyzed how the effects of questionnaire vary with respondent characteristics. Couper (2000) argues that design may interact with the type of web survey conducted and the population at which the survey is targeted. Of the few studies on personal characteristics, some suggest effects (mainly caused by working memory capacity), while others found no variation in response effects with personal characteristics. Tourangeau et al. (2007) observed no consistent variation in the impact of the layout of a response scale by gender, age, or education group. Stern et al. (2007) also showed that the layout of survey questions affects different demographics groups in similar ways. In addition, McFarland (2001) did not find evidence that gender and education level interact with the ordering of questions. Krosnick and Alwin (1987), on the other hand, found that respondents with less education and more limited vocabularies were influenced more than others by different answer categories. Knauper et al. (2004) and Borgers et al. (2004) also found differences due to working memory capacity. Fuchs (2005) found that the effects of response order, scale, and of labeling response categories with numerical values decreases with age when children and adolescents are compared, supporting the hypothesis that response effects decrease with the level of cognitive sophistication.

Literature suggests that additional research on the visual design of web questionnaires is needed to develop more general principles for how the visual layout influences the answers (Christian and Dillman 2004; Dillman and Christian 2002; Dillman and Christian 2005; Friedman and Leefer 1994; Jenkins and Dillman 1997; Schwarz et al. 1991). Deutskens et al. (2004), Dillman et al. (2000), Friedman and Leefer (1994), Hofmans et al. (2007), and Stern et al. (2007) conclude future research on visual design should be directed at confirming the effects of presentation on different questionnaires and populations. Such work is essential for effective survey construction and offers the possibility for methodological improvements of survey research. The experiments described in the following section analyze whether visual languages influence respondents' answers in a Dutch online panel and if any effects are tied to personal characteristics.

Design and Implementation 4.3

Studies on scalar questions have focused on the number of scale points, the use of verbal labels, the use of a midpoint, the use of numerical labels, the use of a 'don't know' filter, and the graphical layout of scales. See Christian (2003), Krosnick and Fabrigar (1997), and Schwarz (1996) for a discussion of these factors in relation to response scales of ordinal questions. We use five point scalar questions, since they give many possibilities for manipulations of visual cues.

Two experiments using eight different formats were carried out in the CentERpanel, an online household panel consisting of more than 2,000 households. This panel is administered by CentERdata (Tilburg University, The Netherlands). The panel is aimed to be representative of the Dutch-speaking population in the Netherlands. Although it is an Internet-based panel, there is no need to have a personal computer with an Internet connection. The households that do not have access to Internet when recruited are provided with a so-called Net.Box, with which a connection can be established via a telephone line and a television set. If the household does not have a television, CentERdata provides one too. The recruitment of new panel members is done in three stages. In the first stage, a random sample (landline numbers) of candidates is interviewed by telephone. The interview ends with the question whether the person would like to participate in survey research projects. If so, the household is included in a database of potential panel members. If a household drops out of the panel, a new household is selected from the database of potential panel members. This is done on the basis of demographic characteristics, such that the panel will remain representative of the Dutch-speaking population (see http://www.centerdata.nl/en/CentERpanel for more details). The time the respondents in our experiment already are participating in the CentERpanel varied from a few months to seventeen years. We used this information to test for an interaction between survey experience and the effects of visual language in the questions we consider. However, we did not find a significant interaction.

The experiment had two questions: one on the quality of education and one on the quality of life in the Netherlands. These questions were taken from an experiment conducted by Christian (2003), who measured the quality of education and the quality of student life at Washington State University.

Our study was conducted in week 37 (September) and week 41 (October) of 2005. The response percentage was 78.3% (2787 panel members were selected, 2182 responded) for the first experiment and 78.8% (2830 panel members were selected, 2229 responded) for the second. A first group of respondents in each experiment answered a rating scale with answer categories excellent, very good, good, fair, and poor in a linear vertical format from positive to negative. In the first experiment we varied the graphics: three non-linear manipulations were used. In the second experiment we manipulated graphical, numerical, and verbal languages in a linear format (see Appendix A for screenshots).

The first experiment is a replication of an experiment done by Christian (2003) on graphical language interrupting a scale, to find out if similar results occur using a representative sample in a different culture. We compared a linear vertical format (Appendix A: 1a) to two non-linear formats: a triple banked format with options running horizontally (Appendix A: 1b) and a triple banked format with options running vertically (Appendix A: 1c). To test whether numbers would help reading the triple vertical format, a fourth group answered the questions in a triple vertical format with numbers (Appendix A: 1d).

In the second experiment, the first group again answered on a rating scale in a linear vertical format from positive to negative (Appendix A: 2a). All other groups have individually different linear manipulations in relation to this format. The second group answered on the same scale, but from negative to positive (poor to excellent, Appendix A: 2b). For the third group the graphics were changed: a linear horizontal format was used (Appendix A: 2c). In the fourth group we added numbers 1 to 5 (Appendix A: 2d). For the fifth group the numbers 5 to 1 were added in the education question, while in the life question the numbers varied from 2 to minus 2 (Appendix A: 2e). The objective was to learn which respondents are more sensitive to verbal and to non-verbal cues. Therefore, scores for different gender, age and education groups were compared.

4.4 Results

In this section we discuss the results of the first and second experiment. We first consider the effects for the complete sample and then consider subsamples with specific demographic characteristics.

4.4.1 Experiment 1: Graphical language: Linear versus Non-Linear Layout

Based on Christian (2003), Christian and Dillman (2004), and Tourangeau et al. (2004), we expected that a non-linear layout results in different responses compared to the linear layout because the graphical language conveying the scale is interrupted. Some respondents might read the top line only. Therefore we hypothesized that in the non-linear format respondents more often choose options in the first line (particularly the response option right next to the first one). In addition, we expected that response times are different across formats because of visual heuristics and Gestalt Psychology, with the reference level (linear format) showing the shortest completion time because this layout is easier to perceive and remember.

Table 4.1 displays response distributions for the two questions. In addition, statistics from Chi square and t-tests are presented to see whether differences in the distribution of individual responses across formats and in mean responses exist. These tests are the same as those in previous research (Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Stern et al. 2007). Lower mean scores indicate more positive ratings (1 = 'excellent',..., 5 = 'poor').

The results in Table 4.1 show that graphical language influences the evaluations of the educational system. The overall Chi-square and differences of means tests reject the null hypothesis of no differences between the four versions (χ^2=33.86, p<.001; F=6.71, p<.01). Separate tests show in five out of six cases that the linear version has significantly different responses and mean scores compared to each of the triple versions, as hypothesized. We found no evidence that respondents are more eager to select an option from the top line in non-linear formats, however. The effect of visual language gets smaller if numbers are added to the vertical format. There may be a hierarchy of features that respondents pay attention to, with numerical labels dominating purely visual cues, as suggested by Tourangeau et al. (2007).

Comparing the triple horizontal and triple vertical format, the frequencies seem to confirm the conjecture that respondents tend to select the answer right next to the first option on the first line. For example, in the education question respondents selected the response option "very good" more often in the triple horizontal format than in the triple vertical format (12.9% and 10.8%, respectively), while the response option "good" was chosen significantly more often in the triple vertical format (52.1% versus 44.0%). Still, the null hypothesis that the complete distributions of the answers are the same for these two formats cannot be rejected at the 5% level.

Respondents chose the first options more often in the linear version than in all non-linear versions, indicating a stronger primacy effect in the linear format than in non-linear formats. In a linear format the visual heuristics of Tourangeau et al. (2004); (2007) are not interrupted, which may lead to less cognitive processing and therefore explain the increased primacy effect. Response times do not support this conjecture, however. We found no difference in response times between formats, and therefore no evidence of the conjecture that it takes longer to process a question if the graphical language is not in line with visual heuristics and principles.

In summary, the results of the first experiment based upon a Dutch hetero-

Table 4.1: Experiment 1. Frequencies (in %), Mean Scores, Correlations and Mean Differences in Linear and Non-Linear Formats

	1a. Linear	Nonlinear - Triple		
		1b. Horizontal	1c. Vertical	1d. Vertical with Numbers
Overall, how would you rate the quality of education in the Netherlands?				
1 Excellent	1.5	0.9	0.6	1.5
2 Very Good	17.8	12.9	10.8	14.7
3 Good	51.3	44.0	52.1	48.9
4 Fair	25.1	36.2	31.9	28.3
5 Poor	4.4	6.0	4.6	6.6
N	550	552	545	530
Mean	3.13	3.34	3.29	3.24
Overall, how would you rate the quality of life in the Netherlands?				
1 Excellent	2.9	2.0	1.5	4.4
2 Very Good	32.3	21.4	24.1	26.4
3 Good	49.9	51.6	56.3	47.3
4 Fair	13.9	23.4	17.0	20.7
5 Poor	0.9	1.7	1.1	1.2
N	545	543	536	518
Mean	2.78	3.01	2.92	2.88
	education question		*life question*	
	Chi Square Tests (χ^2)	Diff. Of means (t)	Chi Square Tests (χ^2)	Diff. Of means (t)
1a versus 1b	20.69**	-4.20**	27.32**	-5.12**
1a versus 1c	16.12**	-3.44**	12.84**	-3.26**
1a versus 1d	5.43	-2.14*	12.19*	-2.07*
1c versus 1b	7.66	-0.93	8.49	-2.02*
1c versus 1d	9.30*	1.12	14.43*	0.95
Overall-across all 4 formats	33.86**	F= 6.71**	43.96**	F= 8.96**

**=p<.05, **=p<.01*

Note: A high mean score indicates a negative judgment.

geneous sample are largely in line with the literature based on homogeneous samples (using students) in a different country and culture.

4.4.2 Experiment 2: Verbal, Graphical, and Numerical Manipulations of Layout

Table 4.2 shows the results for our second experiment. One important difference between the two questions is that a joint Chi square test and differences of means test for all non-verbal manipulations did not show differences in the education question (χ^2=15.97, p=.19; F=1.98, p=.12) while it did in the life question (χ^2=115.16, p<.001; F=32.01, p<.001). This difference is caused by the adding of different numbers in format 2e (5 to 1 in the education question and 2 to -2 in the life question). We looked at the duration of response times to find out if some formats take longer to process, but we found no significant differences between formats.

Verbal language

By changing the verbal orientation of a scale, visual heuristics like 'left and top means first' and 'up means good' (Tourangeau et al. 2004; Tourangeau et al. 2007) are violated. In addition, the theory of satisficing (Krosnick and Alwin 1987; Krosnick et al. 1996; and Tourangeau et al. 2000) states that respondents are more likely to choose items earlier in a list because they find the first position that they can reasonably agree with and consider it a satisfactory answer, rather than processing each response option separately. Therefore, we hypothesized that changing the verbal orientation of a scale would cause different responses, because respondents would select the first options more often.

Our two questions show statistically different answer distributions and mean scores between a decreasing and an increasing scale, indicating that respondents are affected by verbal language. The Chi square tests indicate significant differences in the responses across the two versions (χ^2=14.76, p=.01 in the education question, and χ^2=103.79, p<.001 in the life question). The mean score in the positive to negative scale is lower than the mean of the negative to positive scale in both questions (mean=2.91 for the decreasing scale and 3.28 for the incremental scale in the education question; 2.60 and 2.88, respectively, in the life question), providing evidence for a primacy effect. For example, in the education question the response option "very good" was selected by 24.0% when it was presented as the second alternative, and by 10.7% when it was presented as the fourth alternative. The option "fair" was chosen by 31.1% of the respondents when it was presented as the second alternative and by 16.5% of the respondents as the fourth alternative. Despite the label, the first response options were selected more often. Our results provide empirical support, in a different country and culture, of the theory of satisficing and primacy effects. In particular, in the Netherlands an incremental scale is much more commonly used in everyday life than a decremental scale (e.g., in school grades). Therefore our results suggest that the effect of satisficing leading to a primacy effect is larger than the effect of violating visual heuristics.

Table 4.2: Experiment 2. Frequencies (in %), Mean Scores, Correlations and Differences of Means in the Verbal, Graphical, and Numerical Manipulations

	2a.	2b.	2c.	2d.	2e.
	Reference:	**Verbal:**	**Graphical:**	**Numerical:**	**Numerical:**
	Linear	Linear	Linear	Linear	Linear
	Vertical	Vertical	Horizontal	Vertical	Vertical
	Positive	Negative		With	With
	to	to		Numbers	Numbers
	Negative	Positive		1 to 5	5 to 1#
Overall, how would you rate the quality of education in the Netherlands?					
1 Excellent	2.7	1.5	0.5	3.1	2.5
2 Very Good	24.0	10.7	23.4	22.8	25.4
3 Good	54.8	51.3	52.8	53.8	55.1
4 Fair	16.5	31.1	21.9	17.9	15.2
5 Poor	2.0	5.4	1.4	2.4	1.8
N	442	460	415	457	448
Mean	2.91	3.28	3.00	2.94	2.88
Overall, how would you rate the quality of life in the Netherlands?					
1 Excellent	5.7	3.7	2.7	4.2	8.1
2 Very Good	35.7	25.6	37.4	40.4	40.1
3 Good	52.3	51.1	49.0	43.3	41.3
4 Fair	5.7	18.5	10.1	11.3	9.4
5 Poor	0.7	1.1	0.7	0.9	0.9
N	440	454	414	453	446
Mean	2.60	2.88	2.69	2.64	2.54
	education question			*life question*	
	Chi Square	Diff. Of		Chi Square	Diff. of
	Tests (χ^2)	means (t)		Tests(χ^2)	means (t)
Verbal: 2a versus 2b	14.76**	-7.17**		103.79**	-5.50**
Graphical: 2a versus 2c	10.43*	-1.82		71.92*	-1.80
Numerical: 2a versus 2d	.68	.55		7.03	1.08
Numerical: 2a versus 2e	.58	1.07		13.29**	1.85
Overall across all non-verbal manipulations	15.97	F=1.98		115.16**	F=32.01**
Overall across all formats	47.68**	F= 8.74**		220.57**	F= 52.27**

**=p<.05, **=p<.01*

Note: A high mean score indicates a negative judgment.

#For the second question (life question) numbers 2 to -2 are added.

Graphical language

By changing the graphical orientation of the scale from vertical to horizontal, the graphical language is altered. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales do not elicit the same responses. However, the direction of the difference they found was not consistent. They found no evidence of a shift to the left due to the necessity of more hand/eye movement to select the last options in a horizontal format.

Chi square tests indicate significant differences in the responses across the vertical and horizontal versions (χ^2=10.43, p=.04 in the education question, and χ^2=71.92, p<.01 in the life question), but the mean scores do not statistically differ (t=-1.82, p=.07 in the education question and t=-1.80, p=.07 in the life question). Differences resulted in selecting the fourth option "fair" in the horizontal format more often. Thus, in the horizontal format a shift to the left (as suggested by Friedman and Friedman 1994) is not detected. Respondents may be more willing to read and process all options separately in a horizontal format, since in western countries people read from left to right and from top to bottom. Therefore, a primacy effect might be more likely to emerge in a vertical format. Lower mean scores in the vertical format support this conjecture, but these differences did not reach statistical significance.

Numerical language

Based on the literature (Fuchs 2005; Schwarz et al. 1985; Schwarz et al. 1991) we expected numerical language to influence respondents' answers, especially when negative numbers were added.

No evidence was found that adding the numbers 1 to 5 caused different responses. Chi square tests indicated no significant differences in the responses across the linear version and the linear versions with numbers 1 to 5 (χ^2=.58, p=.97 in the education question, and χ^2=13.29, p=.10 in the life question). In addition, no differences in mean scores were found (t =.55, p=.58 in the education question, and t =1.08, p=.28 in the life question). The numbers 1 to 5 were probably seen as answer category numbers, so respondents did not interpret the answer categories differently when numerical labels were added to the verbal labels.

When adding the numbers 5 to 1 in the first question, we did not find significant differences either. The mean score in the 5 to 1 version was lower than in the 1 to 5 version (respectively 2.88 and 2.91), indicating that respondents select an answer more easily when a higher number is added to the verbal labels. However, the mean scores did not differ significantly (t=1.07, p=.29). In the question about quality of life in the Netherlands the mean score in the 2 to -2 format (2.54) was lower than the mean score in the reference format (2.60) and this difference in means almost reached significance (t=1.85, p=.07). The Chi square test indicated significant differences in the response distribution when numbers 2 to -2 were added to the reference format (χ^2=13.29, p=.01). This confirms that negative numbers are interpreted as implying more extreme judgments than low positive numbers (scale label effect, see Tourangeau et al.

2000, p.248; Schwarz et al. 1991; Tourangeau et al. 2007).

We found little evidence that numbers influenced responses to the five point scale. We only found different response distributions where negative numbers are added to the verbal labels, which might indicate that signs (symbols) affect responses.

4.5 Effects for different demographic subgroups

Based on previous research concerning personal characteristics (Krosnick et al. 1996; Stern et al. 2007), we defined the following demographic subgroups: men and women; two educational categories (with and without a college degree), and two age categories (65 years old and older, and under the age of 65). We used the correlation ratio between the answer score and the scale that was used as a measure of association. A higher correlation ratio (η) between the answer score and the scale indicates a larger design effect. Based on Borgers et al. (2004), Fuchs (2005), Knauper et al. (2004), and Krosnick and Alwin (1987) we hypothesized that the design effects are larger for older respondents and respondents without a degree. Detailed results are presented in Table 4.3 . In the discussion below, we focus on the most salient findings.

Gender

In our first experiment, significant differences across the four formats were found for men in both questions (η=.102, p=.01 in the education question and η=.137, p<.001 in the life question), while women did not report statistically different answers across formats (η=.097, p=.11 in the education question and η=.099, p=.24 in the life question). Men selected the second response option more often in the linear format, and the third option less often in the triple horizontal version[1]. This suggests that men are more sensitive to satisficing than women and less often select options that require many eye/hand movements. This is in line with the results of Stern et al. (2007), who found that satisficing effects were greater for men.

A test across all formats in our second experiment showed significant differences for men and women in both questions: η=.128 p<.01 for men and η= .123 p<.01 for women in the education question and η=.275 p<.01 for men and η= .322, p<.01 for women in the life question. Overall, we found little evidence that men react differently to visual language compared to women.

[1] This option is presented at the right of the screen in the triple horizontal format (see Appendix A 1b).

Table 4.3: Overview of Significance (Chi Square) and Association (η) between Formats for Gender, Age, and Education

Exp. 1	1a versus 1b	1a versus 1c	1a versus 1d	1c versus 1b	1c versus 1d	Overall- across all 4 formats
Overall, how would you rate the quality of education in the Netherlands?						
gender						
men	.122 (.04)	.127 (<.01)	.080 (.27)	.001 (.10)	.041 (.11)	.102 (.01)
women	.135 (.02)	.076 (.25)	.046 (.65)	.063 (.49)	.027 (.25)	.097 (.11)
age						
<65 years	.118 (<.01)	.102 (.02)	.046 (.52)	.021 (.32)	.052 (.09)	.093 (.02)
>64 years	.179 (.07)	.173 (.08)	.180 (.08)	.011 (.26)	.014 (.68)	.169 (.06)
education						
<college	.120 (<.01)	.103 (.01)	.075 (.15)	.023 (.12)	.024 (.09)	.092 (<.01)
college	.191 (.07)	.105 (.35)	.008 (.61)	.098 (.49)	.098 (.37)	.155 (.27)
Overall, how would you rate the quality of life in the Netherlands?						
gender						
men	.171 (<.01)	.141 (.01)	.044 (.05)	.039 (.07)	.088 (<.01)	.137 (<.01)
women	.137 (.02)	.053 (.62)	.081 (.23)	.088 (.25)	.031 (.59)	.099 (.24)
age						
<65 years	.139 (<.01)	.094 (.09)	.061 (<.01)	.048 (.26)	.028 (.04)	.099 (<.01)
>64 years	.215 (.02)	.127 (<.01)	.080 (.27)	.116 (.06)	.041 (<.01)	.174 (<.01)
education						
<college	.131 (<.01)	.065 (.32)	.051 (.02)	.069 (.09)	.010 (.04)	.093 (<.01)
college	.337 (<.01)	.379 (<.01)	.174 (.24)	.019 (.18)	.199 (.06)	.304 (<.01)

Exp. 2	Verbal: 2a versus 2b	Graphical: 2a versus 2c	Numerical: 2a versus 2d	Numerical 2a versus 2e	Overall- 5 formats across all 5 formats
Overall, how would you rate the quality of education in the Netherlands?					
gender					
men	1.07 (.16)	.085 (.08)	.045 (.70)	.010 (.22)	.128 (.01)
women	1.35 (.06)	.045 (.05)	.005 (.74)	.044 (.23)	.123 (<.01)
age					
<65 years	.075 (.34)	.033 (.10)	.006 (1.00)	.054 (.62)	.081 (.19)
>64 years	.355 (<.01)	.176 (.22)	.080 (.26)	.140 (.32)	.365 (<.01)
education					
<college	.114 (.03)	.064 (.01)	.012 (.99)	.022 (.96)	.116 (<.01)
college	.206(.14)	.044 (.88)	.072 (.69)	.017 (.42)	.212 (.19)
Overall, how would you rate the quality of life in the Netherlands?					
gender					
men	.308 (.01)	.262 (<.01)	.056 (.01)	.049 (.10)	.275(.01)
women	.353(<.01)	.284 (<.01)	.003 (.31)	.021 (.14)	.322(<.01)
age					
<65 years	.351 (<.01)	.224 (<.01)	.001(.03)	.068(.01)	.312 (.01)
>64 years	.225(.02)	.444 (<.01)	.169 (.04)	.111 (.32)	.319 (<.01)
education					
<college	.50 (<.01)	.41 (<.01)	.195 (.28)	.047 (.96)	.277 (<.01)
college	.303(<.01)	.253 (.<.01)	.009(<.01)	.045(.01)	.427 (<.01)

Age

Adding numbers apparently influenced respondents under the age of 65 in reading the non-linear vertical format: we found differences between the linear and non-linear vertical layout (without numbers) (η=.102, p=.02) in the education question while the differences diminished when numbers were added to the non-linear vertical format (η=.046, p=.52). Respondents aged 65 and older showed no differences between the reference format and the non-linear vertical format with numbers (η=.180, p=.08 in the education question and η=.080, p=.27 in the life question). They may rely more on numerical language in helping them to process the scale when graphical language is altered.

In the education question, respondents in the age of 65 and older showed significantly different results when the verbal orientation was changed (decremental/incremental; η=.355, p<.01 in the education question), while younger respondents did not. Older respondents showed a primacy effect when the visual heuristic 'up means good' was violated.

When the graphical orientation was changed from vertical to horizontal, older respondents showed a larger recency effect than younger respondents; for example in the education question the fourth response option 'fair' was chosen by 5% in the vertical format, while in the horizontal format 19% of the respondents younger than age 65 chose this option and 33% of the respondents age 65 and older.

Respondents younger than 65 years old chose more extreme judgments when negative signs were added to the verbal labels (η=.068, p=.01); the middle option 'good' was chosen by 54% in the reference format, while 40% chose this option when numbers 2 to -2 were added.

Younger respondents showed a primacy effect when numbers 1 to 5 were added (for example, in the education question 39% chose one of the first two options in the reference format while 49% chose one of the first two options when numbers 1 to 5 were added to the verbal labels), while older respondents showed a recency effect (5% chose one of the last two options in the reference format while 21% chose one of the last two options when numbers 1 to 5 were added to the verbal labels).

We found evidence that reduction in cognitive functioning due to the aging process causes larger response effects due to visual language in 19 out of 22[2] comparisons; respondents age 65 and older showed stronger response effects due to visual language than their younger counterparts. Older respondents were more sensitive to the verbal and graphical orientation of a scale in a linear format. On the other hand, older respondents were less sensitive to negative signs in numerical labels.

[2] four formats in experiment 1, five formats in experiment 2, and two overall test in two questions.

Education

Respondents with a college degree were more sensitive to verbal language than their counterparts without a degree. For example, 24% selected the response option 'very good' in the reference level while 44% selected this option when it was presented as a fourth alternative in the education question. Therefore, we found a recency effect for respondents with a college degree when the visual heuristic 'up means good' was violated. This is in line with the results of Stern et al. (2007), who found that orientation effects were the greatest among those with a college degree. Our data shows that respondents with a college degree are more sensitive to visual language than the respondents without a college degree; response effects due to visual language were larger in 16 out of 22 comparisons.

Discussion and Conclusions 4.6

This article shows that respondents gain meaning from non-verbal cues in a web survey as well as from verbal cues. We manipulated the layout of a five point scalar question in two experiments using two questions. In the first experiment, a linear layout was compared with three non-linear layouts (graphical manipulation). In the second experiment we manipulated verbal, graphical, and numerical language individually, to learn how these verbal and non-verbal cues influence answers in rating scales. This chapter extends previous research as both linear and non-linear and verbal, graphical, and numerical languages are individually manipulated on the same rating scale and it is analyzed in which way personal characteristics account for variance in survey responding.

In the comparison between linear and non-linear formats we found differences across all versions. Triple horizontal and triple vertical format show significant different means compared to the linear format. In a triple visualization, respondents are more eager to select the second answer on the top line. Our results support a primacy effect in answering scalar questions. Options that require less movement of the mouse might be more easily chosen than answers requiring more hand/eye movements. The effect of visual language gets smaller if numbers are added to the vertical format. This seems to point at a hierarchy of features that respondents pay attention to, with numerical labels taking precedence over purely visual cues, as suggested by Tourangeau et al. (2007). Future research can make this effect more clear.

In experiment 2, again different responses due to visual language were found. The verbal manipulation ('excellent'-'poor' versus 'poor'-'excellent') shows significantly different responses compared to the other manipulations. This indicates satisficing and also that a negative tone of the first option changes reports in a negative manner (anchoring effect, as suggested by Schwarz 1996). Despite the label, respondents select the second option more often.

Statistically significant differences were also found when comparing the nonverbal manipulations with each other, caused by graphical manipulation. Changing the answer categories to a horizontal format changed the answers. An interpretation is that respondents may be more willing to read all options in the horizontal format (because they first read horizontally and then vertically). Adding the numbers 1 to 5 or 5 to 1 to the vertical format did not influence the answers. Adding the numbers 2 to -2 resulted in respondents being less eager to assign negative scores.

Looking at the mean scores in the different formats, the mean score on the horizontal scale is closest to the overall mean. Because all other formats have a vertical format, this is remarkable. While we already have seen that the horizontal format is the least sensitive for primacy effects, it could be that presenting a five point scale horizontally makes sure that respondents read the answer categories more accurately, decreasing the influence of layout. Further research in web surveys on a horizontal layout of scalar questions in different contexts is warranted.

The effect of format varies with personal characteristics of the respondents. The elderly and the highly educated are in general more sensitive to layout effects than others. Deriving conclusions on a student-based sample might show more differences between different formats than a heterogeneous sample of the population. Future research should be conducted comparing student based and representative samples to find out if studies using students as respondents show more significant results.

This chapter shows that the visual presentation of answer categories must be taken into consideration in order to reduce measurement error. This goes especially for researchers who want to compare results across surveys. Similarly worded questions may be presented to respondents in visually dissimilar ways. Do different results then come from a different time of measurement or from a different visualization? This is a challenge for further research.

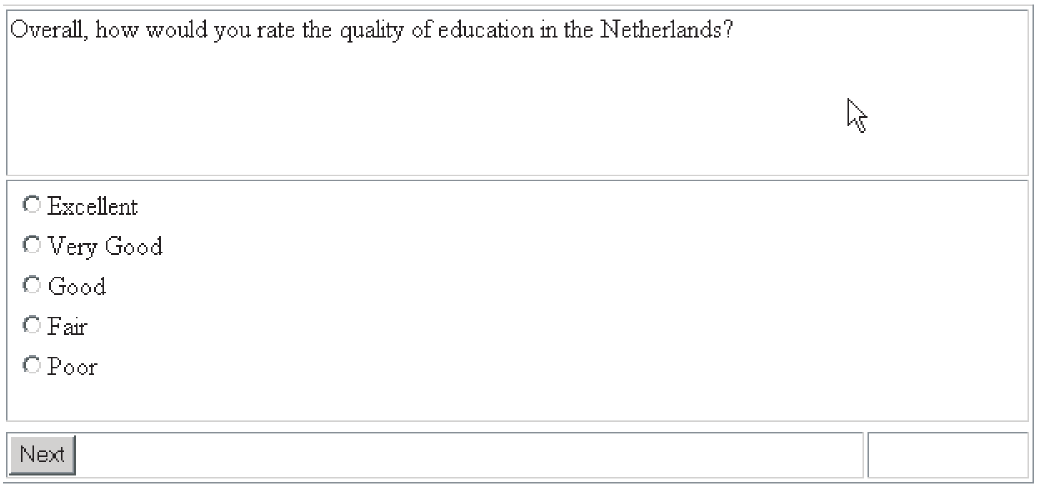
Appendix A: Screen shots 4.7

Experiment 1

Four different layouts were used, using a linear and a non-linear format, in two questions, namely

1. Overall, how would you rate the quality of education in the Netherlands?
2. How would you rate the quality of life in the Netherlands?

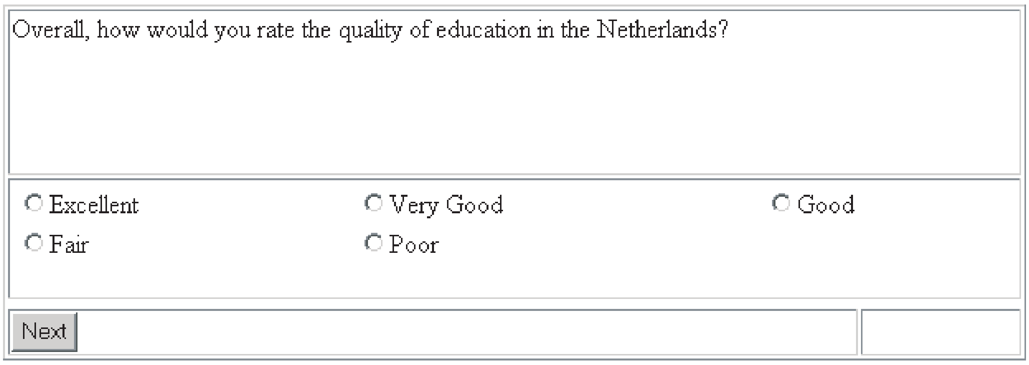
The screen shots below show the different layout formats for the education question. The layout formats used in the life question are exactly the same.



Overall, how would you rate the quality of education in the Netherlands?

○ Excellent
○ Very Good
○ Good
○ Fair
○ Poor

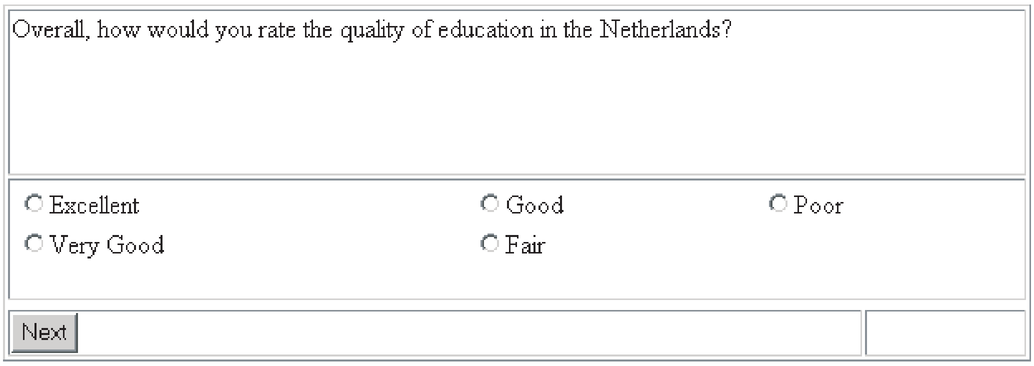
Figure 4.1: format 1a. Linear



Overall, how would you rate the quality of education in the Netherlands?

○ Excellent ○ Very Good ○ Good
○ Fair ○ Poor

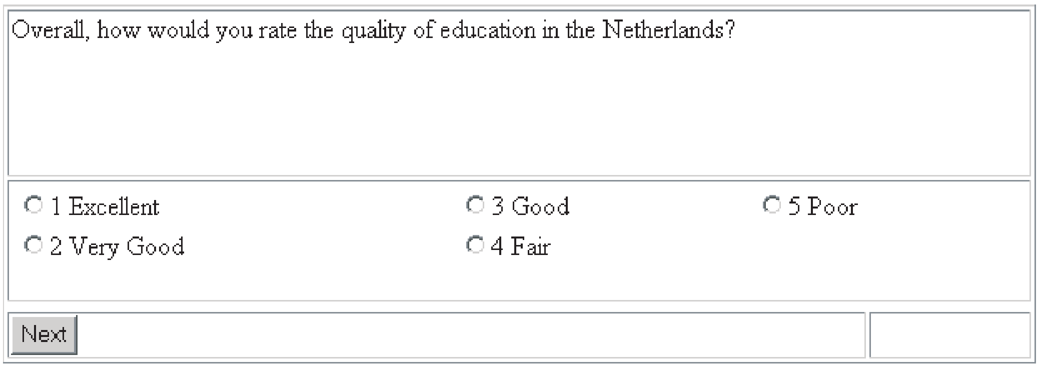
Figure 4.2: format 1b. Nonlinear - triple horizontal



Overall, how would you rate the quality of education in the Netherlands?

○ Excellent ○ Good ○ Poor
○ Very Good ○ Fair

Figure 4.3: format 1c. Nonlinear - triple vertical



Overall, how would you rate the quality of education in the Netherlands?

○ 1 Excellent ○ 3 Good ○ 5 Poor
○ 2 Very Good ○ 4 Fair

Figure 4.4: format 1d. Nonlinear-triple vertical with numbers

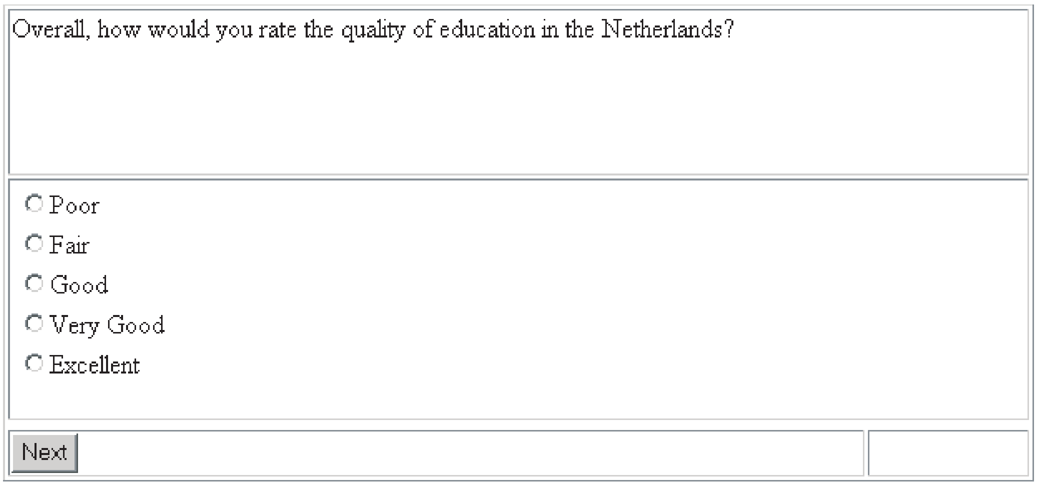
Experiment 2

Five different layouts were used in the same two questions (as in experiment 1):

1. Format a: reference format (see 1a);
2. Format b: verbal manipulation: response scale is in this format from negative to positive;
3. Format c: graphical manipulation: response scale is in this format from vertical to horizontal;
4. Format d: numerical manipulation: numbers 1 to 5 are added in this format;
5. Format e: numerical manipulation: numbers 5 to 1 are added in education question, while numbers 2 to -2 are added in the life question).

The screen shots below show the different layout formats for the education question, the layout formats used in the life question are the same except for format e (see above).

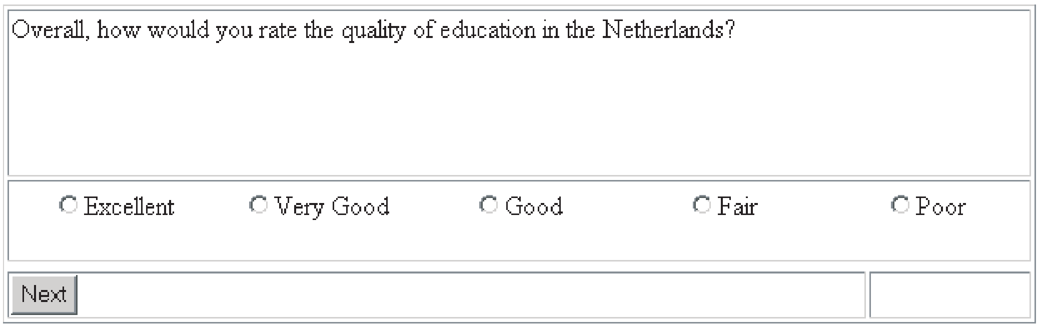
Format 2a. Linear positive to negative: See screen dump 1a.



Overall, how would you rate the quality of education in the Netherlands?

○ Poor
○ Fair
○ Good
○ Very Good
○ Excellent

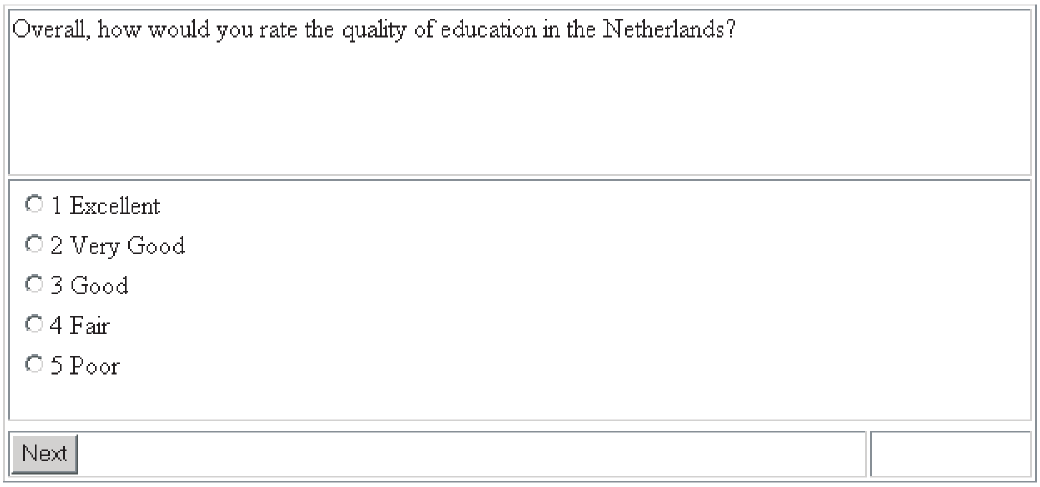
Figure 4.5: format 2b. Linear negative to positive (verbal)



Overall, how would you rate the quality of education in the Netherlands?

○ Excellent ○ Very Good ○ Good ○ Fair ○ Poor

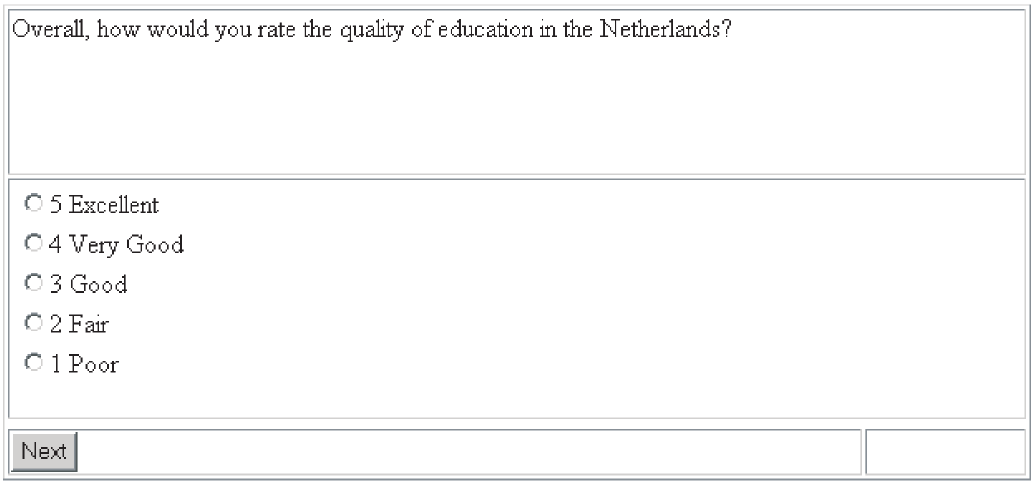
Figure 4.6: format 2c. Linear horizontal (graphical)



Overall, how would you rate the quality of education in the Netherlands?

○ 1 Excellent
○ 2 Very Good
○ 3 Good
○ 4 Fair
○ 5 Poor

Figure 4.7: format 2d. Linear with numbers 1 to 5, 1=positive (numerical)



Overall, how would you rate the quality of education in the Netherlands?

○ 5 Excellent
○ 4 Very Good
○ 3 Good
○ 2 Fair
○ 1 Poor

Figure 4.8: format 2e. with numbers 1 to 5, 5=positive in education question (numerical) Note: Format 2e for the life question ranges from 2 (positive) to -2 (negative).

5 Can I Use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys

ABSTRACT Over the past decades there has been an increasing use of panel surveys at the household or individual level, instead of using independent cross-sections. Panel data have important advantages, but there are also two potential drawbacks: attrition bias and panel conditioning effects. Attrition bias can arise if respondents drop out of the panel non-randomly. Panel conditioning arises if responses in one wave are influenced by participation in the previous wave(s). The literature has mainly focused on estimating attrition bias; less is known on panel conditioning effects.

In this study we discuss how to disentangle the total bias in panel surveys due to attrition and panel conditioning into a panel conditioning and an attrition effect, and develop a test for panel conditioning allowing for non-random attrition. First, we consider a fully nonparametric approach without any assumptions other than those on the sample design, leading to interval identification of the measures for the attrition and panel conditioning effect. Second, we analyze the proposed measures under additional assumptions concerning the attrition process, making it possible to obtain point estimates and standard errors for both the attrition bias and the panel conditioning effect.

We illustrate our method on a variety of questions from two-wave surveys conducted in a Dutch household panel. We found a significant bias due to panel conditioning in knowledge questions, but not in other types of questions. The examples show that the bounds can be informative if the attrition rate is not too high. Point estimates of the panel conditioning effect do not vary a lot with the different assumptions on the attrition process.

BIS REPETITA PLACEAT

5.1 Introduction

One of the most important developments in the social sciences over the past decades has been the increasing use of panel surveys at the household or individual level. Panel data have important advantages for research, such as creating the possibility to analyze changes at the micro-level, without making additional assumptions, to disentangle permanent from transitory characteristics, to distinguish between causal effects and individual heterogeneity, etc. (see, e.g., Baltagi 2001 or Lee 2002). Two potential drawbacks compared to, e.g., independent cross-sections are attrition bias and panel conditioning effects (see, e.g., Sharot 1991 or Trivellato 1999).

Attrition bias can arise if respondents drop out of the panel non-randomly, i.e., when attrition is correlated to a variable of interest. Panel attrition has been studied extensively, usually without discussing the possibility of panel conditioning effects. See, e.g., Fitzgerald et al. (1998), Vella (1998), and Nicoletti (2006). Hirano et al. (2001) show how a refreshment sample can be used to relax the assumptions under which attrition can be identified. Their first model makes the assumption that the observations in the second period are missing at random (MAR, Rubin 1976). Their second model is closely related to the model of Hausman and Wise (1979), allowing the probability of attrition to depend on second period variables, but not on first period variables. With a refreshment sample, the distinction between these two models can be non-parametrically identified.

Panel conditioning arises if responses in one wave are influenced by having participated in the previous wave(s). The experience of the previous interview(s) may affect the answers of respondents in a next interview on the same topic, such that their answers differ systematically from the answers of individuals who are interviewed for the first time. This may be a good thing and reduce measurement error, if respondents learn how to interpret questions and make fewer errors. On the other hand, experienced respondents may become strategic and learn, e.g., that answering “no" reduces the burden of their task, avoiding follow up questions (see, e.g., Meurs et al. 1989 and Duan et al. 2007). Sturgis et al. (2007) expand on the main theory behind panel conditioning: the cognitive stimulus hypothesis. Questions asked about certain topics may induce respondents to reflect more closely on them after the interview has ended, and possibly to talk about them with friends and relatives or to acquire additional information through the media. This should particularly lead to a difference between knowledge or attitudes reported at the first and second interview. They find some empirical evidence in favor of this, but have to ignore attrition effects as well as time trends. Brannen (1993) asked explicit questions on the effects of survey participation and also found that respondents became more aware of and interested in the research issues (child behavior and parental roles).

Panel conditioning has been studied in many social sciences, with mixed findings. While Williams (1970), Williams and Mallows (1970), and Meurs et al. (1989) showed that systematic biases occur in panel data sets, due to attrition as well as panel conditioning, Coombs (1973) found differences in knowledge due to re-interviewing, i.e., panel conditioning, but little impact on behavior or attitudes. Waterton and Lievesley (1989) found some evidence that respondents are influenced by re-interviewing, especially respondents with low knowledge scores. On the other hand, Dennis (2001) and Clinton (2001) found little evidence for attrition or panel conditioning in the Knowledge Networks' panel (an online panel that is representative of the entire US population) and Pennell and Lepkowski (1992) found hardly any evidence of panel conditioning or attrition bias in income sources reported in the Survey of Income and Program Participation. Mathiowetz and Lair (1994) found evidence of panel conditioning in the measurement of functional health limitations, which can be explained by strategic behavior: by not reporting limitations, follow-up questions can be avoided. Similar results for the use of various types of health care services were found by Duan et al. (2007), who concluded that there was underreporting in the later items reported in the same survey. Van der Zouwen and Van Tilburg (2001) showed that most of their evidence of panel conditioning for measurement of personal network size in repeated personal interviews could be attributed to behavior of the interviewers. Sharpe and Gilbert (1998) find that repeated testing (interrupted by a 1 week interval) increases the scores on the Beck depression scale and attribute this to socially desirable responding, mood-congruent associative processing, or self-monitoring, triggered by the first interview. Similar effects, called “testing effects" in this context, were found within the same experimental session by Chan and McDermott (2007).

In practice, it is difficult to separate the effects of panel conditioning from those of other changes between waves (Kalton et al. 1989). Many studies on panel effects do not explicitly distinguish between attrition and panel conditioning and only look at the total bias induced by both, see, e.g., Pennell and Lepkowski (1992) on income sources, Bartels (1999) on campaign interest and turnout at national elections, Lohse et al. (2000) on consumer buyer behavior, Wang et al. (2000) who found some significant panel effects in a set of 32 variables on use of medical care and social security, or Golob (1990) who found panel effects on reported travel time expenditures.

In this chapter we aim at disentangling panel conditioning from attrition bias, with the goal of testing for panel conditioning while controlling for attrition bias. We extend the framework of Hirano et al. (2001) incorporating the possibility of panel conditioning effects, emphasizing the usefulness of a refreshment sample. The setup, with an initial sample interviewed once (in case of attrition) or twice (non-attrition) and a refreshment sample interviewed once, is described in Section 5.2. Section 5.3 proposes two measures for the attrition bias and the panel conditioning effect. Without further assumptions these measures are not point-identified. We then consider two approaches. First, we follow Manski (1989; 1995) and derive bounds on the panel conditioning and attrition effects, without making further assumptions. Second, we discuss several sets of additional assumptions on the attrition process under which we can obtain point estimates and standard errors for the attrition and panel conditioning effects. In Section 5.4 we illustrate our method for several repeated measurements conducted in the CentERpanel, a representative panel of Dutch households. We find evidence of panel conditioning in knowledge questions, but not in questions on behavior or attitudes. Section 5.5 concludes.

5.2 Setup

We consider the case of two interview times, time 1 and time 2, with the same population (assumed to be the same at both points in time). For notational convenience we work with questions that can only have two answers, coded as 0 and 1. Our approach can in principle be extended to other questions, since the distribution of the outcome of interest can be fully characterized by binary events. For example, if we are interested in panel conditioning on a continuous variable Z, we can study panel conditioning on the binary variables $I[Z > t]$ for each t, where I is the indicator function. See Manski (1995). Similarly, we only look at marginal distributions, but the approach also applies to conditional distributions given an always observed set of covariates X. In practice, this means estimation by subsample with given values of X if X is discrete, while some smoothing technique needs to be applied if X has continuous components.

We are interested in the following (population) variables. The variable $Z_1 \in \{0,1\}$ denotes the answer to the question of interest at time 1. $Z_2(1) \in \{0,1\}$ is the answer to the same question given at time 2 that the respondent (would) give(s) if the interview at time 2 is her first interview. The variable $Z_2(2) \in \{0,1\}$ denotes the time-2-answer that the respondent (would) give(s) if the interview at time 2 is her second interview. Finally, the variable W takes value 1 if the respondent, if interviewed at time 1, also responds at time 2 (“panel observation"), and takes value 0 otherwise (“attrition"). Compared to the setup of Hirano et al. (2001) we incorporate panel conditioning, i.e., we allow for the possibility that the answer to the question at time 2 can be affected by being interviewed at time 1, i.e. $Z_2(1) \neq Z_2(2)$. The parameters of interest that we consider in this chapter are all functions of the population distribution of $(Z_1, Z_2(1), Z_2(2), W)$, described by 16 parameters $\Pr(Z_1 = a, Z_2(1) = b, Z_2(2) = c, W = d)$, $a, b, c, d \in \{0,1\}$.

The sample design is as follows. At time 1 a random sample of size n_1 is drawn from the population of interest, Sample 1. We assume throughout the chapter that there is no initial (unit or item) non-response (or that initial non-response is MAR). The respondents in Sample 1 answer the question of interest and their answers are denoted by $Z_{i,1}$, $i = 1,\ldots,n_1$. At time 2, all Sample 1 individuals are approached for a second interview. If respondent i responds, then $W_i = 1$ and $Z_{i,2}(2)$ is observed. If respondent i does not respond, we only observe $W_i = 0$. Hence, $n_P = \sum_{i=1}^{n_1} W_i$ is the number of respondents in Sample 1 that stay in the panel (“panel members") and $n_A = n_1 - n_P$ is the number of respondents that attrite.

At time 2, a refreshment sample is available. This is a (new) random sample ("Sample 2") of size n_R from the population of interest (to be precise: the population excluding the respondents in Sample 1, but we assume the population is infinitely large). We assume there is no non-response in this sample (or that non-response is MAR). Since the respondents are interviewed for the first time, this sample yields observations $Z_{i},_2(1)$, $i = 1,\ldots,n_R$.

In summary, at time 1, we only have respondents interviewed for the first time (attrition and panel sample, the union of them is a simple random sample, Sample 1). At time 2, we have respondents interviewed for the second time (panel part of Sample 1), respondents who are interviewed for the first time (refreshment sample Sample 2, again a simple random sample), and respondents who do not respond at time 2 (attrition part of Sample 1).

Parameters identified without further assumptions

The sample design implies that eight functions of the sixteen population parameters are identified and can be estimated consistently without further assumptions. From Sample 1 we can consistently estimate six probabilities using corresponding sample analogues: the two probabilities $\Pr(Z_1 = z_1, W = 0)$, $z_1, \in \{0,1\}$, and the four probabilities $\Pr(Z_1 = z_1, Z_2(2) = z_2, W = 1)$, $z_1, z_2 \in \{0,1\}$.

Similarly, the refreshment sample can be used to consistently estimate the two probabilities $\Pr(Z_2(1) = z_2)$, $z_2 \in \{0,1\}$ using their sample analogues.

This is obviously not enough to estimate the complete joint distribution of the four variables $Z_1, Z_2(1), Z_2(2)$ and W. For example, we only know the marginal distribution of $Z_2(1)$, and nothing about how $Z_2(1)$ relates to the other three variables, since $Z_2(1)$ is never observed jointly with any of the other three. Similarly, we know nothing of the distribution of $Z_2(2)$ when $W = 0$. The latter is the familiar problem of identification under selective attrition, as in Hirano et al. (2001). The difference with Hirano et al. (2001) is that we want to allow for arbitrary panel conditioning effects, implying that we do not impose any restrictions on the relation between $Z_2(1)$ and $Z_2(2)$. The refreshment sample is informative about the distribution of $Z_2(1)$ but not about the distribution of $Z_2(2)$.

5.3 Measures for attrition and panel conditioning bias

This section introduces several parameters of interest that are functions of the 16 population parameters describing the distribution of $(Z_1, Z_2(1), Z_2(2), W)$. The (true) *trend effect* (taking outcome 1 as the reference level) is given by $TE = \Pr(Z_2(1) = 1) - \Pr(Z_1 = 1)$. The second term can be estimated consistently from Sample 1. Typically, ignoring possible effects of attrition and panel conditioning and not using a refreshment sample, one would estimate the first term by

$$\frac{1}{n_P} \sum_{i=1}^{n_1} Z_{i,2}(2) W_i.$$

This is a consistent estimator of $\Pr(Z_2(2) = 1|W = 1)$, which, in general, differs from $\Pr(Z_2(1) = 1)$. Using it to estimate TE would thus induce the asymptotic “total bias" TB given by:

$$TB = \Pr(Z_2(2) = 1|W = 1) - \Pr(Z_2(1) = 1).$$

With the refreshment sample, $\Pr(Z_2(1) = 1)$ can be estimated consistently in a straightforward way. Thus TB is identified (without additional assumptions) and can be estimated consistently by

$$\begin{aligned} T\hat{B} &= \hat{\Pr}(Z_2(2) = 1|W = 1) - \hat{\Pr}(Z_2(1) = 1) \\ &= \frac{1}{n_P} \sum_{i=1}^{n_1} Z_{i,2}(2) W_i - \frac{1}{n_R} \sum_{i=1}^{n_R} Z_{i,2}(1). \end{aligned}$$

Inference on TB is straightforward, because samples 1 and 2 are independent of each other. Thus, for example, a test for the null hypothesis $H_0 : TB = 0$ (versus the alternative $H_1 : TB \neq 0$) can be based upon the difference between two independent sample fractions.

5.3.1 Decompositions

The main point of our chapter is to decompose the total bias into two components that give an attrition bias (AB) and a panel conditioning effect (PC). This can be done in two ways, depending on the order.

Decomposition 1

In decomposition 1 the total bias is decomposed in the following way:

$$\begin{aligned} TB &= PC_1 + AB_1 \\ &= [\Pr(Z_2(2) = 1|W = 1) - \Pr(Z_2(1) = 1|W = 1)] \\ &\quad + [\Pr(Z_2(1) = 1|W = 1) - \Pr(Z_2(1) = 1)]. \end{aligned}$$

Without additional assumptions, we cannot identify AB_1 and PC_1, because $\Pr(Z_2(1) = 1|W = 1)$ is not identified. However, we can derive bounds on this probability, following Manski (1989, 1995). First, note that this probability equals

$$\Pr(Z_2(1) = 1|W = 1) = \frac{\Pr(Z_2(1) = 1, W = 1)}{\Pr(W = 1)}.$$

The denominator is identified. The numerator is not - we can identify the marginal probabilities $\Pr(Z_2(1) = 1)$ and $\Pr(W = 1)$ but other than that, the data are not informative about the joint probability. Thus it is straightforward to show that sharp lower and upper bounds on $\Pr(Z_2(1) = 1, W = 1)$ are given by: $\ell \leq \Pr(Z_2(1) = 1, W = 1) \leq r$, with

$$\begin{aligned} \ell &= \max(0, 1 - \Pr(Z_2(1) = 0) - \Pr(W = 0)), \\ r &= \min(\Pr(Z_2(1) = 1), \Pr(W = 1)). \end{aligned}$$

This immediately implies the following bounds on PC_1 and AB_1: $\ell \leq PC_1 \leq r$, with

$$\begin{aligned} \ell &= \Pr(Z_2(2) = 1|W = 1) - \min\left(\frac{\Pr(Z_2(1) = 1)}{\Pr(W = 1)}, 1\right), \\ r &= \Pr(Z_2(2) = 1|W = 1) - \max\left(0, 1 - \frac{\Pr(Z_2(1) = 0)}{\Pr(W = 1)}\right), \end{aligned}$$

and $\ell \leq AB_1 \leq r$ with,

$$\begin{aligned} \ell &= \max\left(0, 1 - \frac{\Pr(Z_2(1) = 0)}{\Pr(W = 1)}\right) - \Pr(Z_2(1) = 1), \\ r &= \min\left(\frac{\Pr(Z_2(1) = 1)}{\Pr(W = 1)}, 1\right) - \Pr(Z_2(1) = 1). \end{aligned}$$

All expressions in these bounds can be estimated straightforwardly, replacing probabilities by their sample analogues. Note that the distance between upper and lower bound is bounded by $\Pr(W = 0)/\Pr(W = 1)$ for both effects. Thus the bounds are informative if attrition is low, i.e., if $\Pr(W = 0)$ is small.

The panel conditioning effect in Decomposition 1 is the panel condition effect for the non-attritors. Conceptually, it might be more interesting to consider the (potential) panel conditioning effect in the whole population. This is achieved in Decomposition 2.

Decomposition 2

In decomposition 2 the total bias is decomposed as follows:

$$TB = AB_2 + PC_2$$

$$= [\Pr(Z_2(2)=1|W=1) - \Pr(Z_2(2)=1)] + [\Pr(Z_2(2)=1) - \Pr(Z_2(1)=1)].$$

Without additional assumptions, we cannot identify AB_2 or PC_2, because $\Pr(Z_2(2) = 1)$ is not identified (since we have no observations on $Z_2(2)$ if $W = 0$). Decomposing $\Pr(Z_2(2)=1) = \Pr(Z_2(2)=1, W=1) + \Pr(Z_2(2)=1|W=0)\Pr(W=0)$, the following sharp bounds can be derived straightforwardly:

$$\begin{aligned} PC_2 \in & [\Pr(Z_2(2)=1, W=1) - \Pr(Z_2(1)=1), \\ & \Pr(Z_2(2)=1, W=1) - \Pr(Z_2(1)=1) + \Pr(W=0)]; \\ AB_2 \in & [\Pr(Z_2(2)=1|W=1) - \Pr(Z_2(2)=1, W=1) - \Pr(W=0), \\ & \Pr(Z_2(2)=1|W=1) - \Pr(Z_2(2)=1, W=1)]. \end{aligned}$$

The bounds can be estimated consistently by replacing probabilities by their sample analogues. Again, the distance between the bounds depends on the attrition probability – it is given by $\Pr(W = 0)$.

5.3.2 Additional assumptions

The previous section shows that further assumptions are needed to obtain point identification of the panel conditioning effect and the attrition bias. Particularly if the attrition rate is substantial, the bounds are too wide to be informative, and additional assumptions are needed to make useful inferences. In this section we discuss several possible additional assumptions concerning the attrition process. Which of them is most plausible will depend on the application of interest.

1. Attrition is not associated with time 2 answers

Assumption 1a (before panel conditioning): for $a \in \{0,1\}$:

$$\Pr(W=1|Z_2(1)=a) = \Pr(W=1).$$

Assumption 1b (after panel conditioning): for $a \in \{0,1\}$:

$$\Pr(W=1|Z_2(2)=a) = \Pr(W=1).$$

Both assumptions are similar to the assumption that wave 2 non-response is missing completely at random (CMAR, cf. Little and Rubin 2002). They are rather strong, since they do not condition on the wave 1 answer. So in most applications it seems better to introduce a third and a fourth version, replacing CMAR by MAR, missing at random, conditional on observables, in this case the time 1 answer Z_1:[1]

[1] Fitzgerald et al. (1998) and others refer to this as no selection on unobservables.

Assumption 1c (before panel conditioning):

$$\Pr(W=1|Z_1=z_1, Z_2(1)=z_2) = \Pr(W=1|Z_1=z_1)\ (z_1, z_2 \in \{0,1\}).$$

Assumption 1d (after panel conditioning):

$$\Pr(W=1|Z_1=z_1, Z_2(2)=z_2) = \Pr(W=1|Z_1=z_1)\ (z_1, z_2 \in \{0,1\}).$$

Assumption 1c does not help in identifying the components in decomposition 1, since $Z_2(1)$ and Z_1 are never observed jointly. In the remainder, we therefore do not consider CMAR Assumption 1c.

2. Attrition has the same effect at times 1 and 2

Assumption 2a (before panel conditioning): for $a \in \{0,1\}$:

$$\Pr(Z_1=a|W=1) - \Pr(Z_1=a) = \Pr(Z_2(1)=a|W=1) - \Pr(Z_2(1)=a).$$

Assumption 2b (after panel conditioning): for $a \in \{0,1\}$:

$$\Pr(Z_1=a|W=1) - \Pr(Z_1=a) = \Pr(Z_2(2)=a|W=1) - \Pr(Z_2(2)=a).$$

Both of these assume, in different senses, stationarity of the attrition bias.

Point estimation under additional assumptions

How can the additional assumptions discussed above be used to obtain point estimates? All our point estimates are based on sample analogues of unconditional or conditional probabilities.

Assumption 1a
Under Assumption 1a, W and $Z_2(1)$ are independent and hence $\Pr(Z_2(1)=1|W=1) = \Pr(Z_2(1)=1)$. Thus under this assumption $AB_1 = 0$ and $PC_1 = TB$, and AB_1 and PC_1 are identified since TB is identified.

Assumption 1b
Under Assumption 1b, W and $Z_2(2)$ are independent and hence $\Pr(Z_2(2)=1|W=1) = \Pr(Z_2(2)=1)$. Thus under this assumption $AB_2 = 0$ and $PC_2 = TB$, and AB_2 and PC_2 are identified.

Assumption 1d
Under Assumption 1d we have

$$\Pr(Z_1=z_1, Z_2(2)=1) = \frac{\Pr(Z_1=z_1, Z_2(2)=1, W=1)}{\Pr(W=1|Z_1=z_1)}, z_1 \in \{0,1\}$$

and hence

$$\Pr(Z_2(2)=1) = \frac{\Pr(Z_1=0, Z_2(2)=1, W=1)}{\Pr(W=1|Z_1=0)} + \frac{\Pr(Z_1=1, Z_2(2)=1, W=1)}{\Pr(W=1|Z_1=1)}.$$

The four probabilities on the right hand side can all directly be estimated with their sample analogues, so under Assumption 1d, AB_2 and PC_2 are identified.

Assumption 2a
Under Assumption 2a we have

$$\Pr(Z_2(1)=1|W=1) = \Pr(Z_2(1)=1) - \Pr(Z_1=1) + \Pr(Z_1=1|W=1),$$

and all three probabilities on the right hand side can be directly estimated with their sample analogues. Thus AB_1 and PC_1 are identified.

Assumption 2b
Under Assumption 2b we have

$$\Pr(Z_2(2)=1) = \Pr(Z_2(2)=1|W=1) + \Pr(Z_1=1) - \Pr(Z_1=1|W=1),$$

and the probabilities on the right hand side can be estimated directly by their sample analogues, so that AB_2 and PC_2 are identified.

It is straightforward to check that Assumptions 2a and 2b give the same expression for AB_1 and AB_2 (and thus also for PC_1 and PC_2). The estimators based upon sample analogues will therefore also be the same.

5.4 Empirical illustrations

In this section we use the estimated bounds and point estimates of the previous section to compute estimates of panel conditioning effects and attrition bias (for the two decompositions) in several examples. We make use of the CentERpanel, an Internet panel representative of the Dutch population ages 16 and over, administered by CentERdata, Tilburg University. Because not everyone owns a personal computer or has access to Internet, CentERdata provides a set-top box for people who do not have a computer, enabling them to complete the questionnaires online. The setup is similar to the one chosen by Knowledge Networks in the US.

Respondents of the CentERpanel are asked to fill out a questionnaire every week. We selected various binary variables in several two-wave research projects. Details of the questions and the results are presented in Appendix A. Standard errors for the estimates (point estimates or lower and upper bounds of the interval estimates) were calculated using the Central Limit Theorem and the Delta-method.

The hypothesis that the *total bias* is equal to zero is rejected for only a few of the variables we analyzed. In particular, this only happened if the question referred to knowledge. For other question types, referring to actual behavior or actual circumstances, attitudes and opinions, or future expectations, no significant total bias was found. The fact that knowledge questions are the most sensitive to panel conditioning is consistent with the literature (cf. Section 1).

Table 5.1 summarizes the results for the three knowledge questions for which we find a significant *total bias*: “Do you know what campylobacter is?", “Do you know what cross-infection is?", and “Have you ever heard of a foundation named “Stichting Pensioenkijker?". The first two stem from a survey module on hygiene knowledge, fielded in November 2003 and November 2005. The third question is from a survey module on pensions and pension knowledge, held in February 2004 and February 2005. Stichting Pensioenkijker is a Dutch non-profit organization that aims at increasing the Dutch population's knowledge about pensions and to help them prepare financially for retirement. Their main instrument is a web site (http://www.pensioenkijker.nl).

Consider the first example - knowledge of campylobacter. At time 1, 19.3% report they know what this is. Among panel observations, this increases to 28.1% at time 2, whereas in the refreshment sample, it increases much less – to 21.9%. The difference is the estimate of the total bias, 6.17%-points, due to panel conditioning, attrition, or both. Without making further assumptions, the estimates on the lower and upper bound of the panel conditioning component of the total bias are 1.10 and 24.27 %-points according to decomposition 1 and 0.90 and 19.70%-points for decomposition 2. Neither the 1.10 nor the 0.90 are significantly different from 0 (standard errors are 2.16 and 1.76, respectively). Thus in this example, without making further assumptions, we cannot reject the hypothesis that there is no panel conditioning. But of course it is possible that this is due to the width of the bounds - they may be not informative enough to give the test enough power.

This changes if additional assumptions are made on the nature of attrition so that point identification is obtained. Under all additional assumptions we consider, Ass. 1a (or 1b, which gives the same as 1a – $PC = TB$), Ass. 1d, or Ass. 2a (or 2b, which gives the same as 2a) we find that all or almost all of the total bias can be attributed to panel conditioning, with estimates of the panel conditioning effect that are 6.17%-points, 5.85%-points and 5.96%-points, respectively, all significantly different from zero. Which of the assumptions is most plausible is hard to judge without further analysis and is beyond the scope of our empirical illustration, but it seems reassuring that the result is insensitive to the choice of assumption or the choice of decomposition.

In the second example, on knowing the meaning of cross-infection, the results are similar. The total bias is estimated to be 6.71%-points, and under the additional assumptions that allows for point estimation, most or all of this is panel conditioning (6.71, 5.88 or 6.31%-points, all significantly different from zero). The only difference with the first example is that the point estimates of the lower bound of the panel conditioning effect are negative so that the estimated interval contains zero, making a test whether the lower bound is significantly different from zero unnecessary – the fact that the lower bound is negative and the upper bound is positive already implies that without additional assumptions on attrition, the zero hypothesis of no panel conditioning cannot be rejected. Again, the lack of information reflected in the width of the bounds may be driving this result. If additional assumptions guaranteeing point identification are made, panel conditioning becomes significant, and reassuringly, this result is robust for the choice of additional assumptions or the choice of decomposition.

The third example, on knowing “Stichting Pensioenkijker", gives the strongest evidence of panel conditioning. At time 1, 7.55% of respondents have heard of this organization. For panel respondents, this rises to 16.47% one year later. In the refreshment sample drawn at the same time, 11.27% report they know “Stichting Pensioenkijker." The difference of 5.20%-points is statistically significant. Without further assumptions, the implied lower bound on the panel conditioning effect is 3.85 or 3.44%-points (for decompositions 1 and 2, respectively), and both are significantly positive (standard errors are 1.64 and 1.47, respectively). Thus even without making further assumptions, we find significant evidence of panel conditioning. The main reason why we find this here and not in the example on campylobacter is the lower attrition rate – 10.7% versus 18.8%. Under additional assumptions 1a, 1d, or 2a, the point estimates of the panel conditioning effect are always 5.2%-points (and, as expected, significantly larger than zero). The reason that the point estimates are all virtually identical is that in this example, the sample analogues of $\Pr(W = 1|Z_1 = 1)$ and $\Pr(W = 1|Z_1 = 0)$ are virtually identical, which implies that the attrition bias is zero under any of the additional assumptions.

Table 5.1: Panel Conditioning in Three Knowledge Questions

	Campylobacter	Cross-infection	Stichting Pensioenkijker
Size Sample 1	1510	1510	1734
Attrition rate (%)	18.8	18.8	10.7
Size Sample 2	891	891	701
Total Bias (%-points)	6.17*	6.71*	5.20*
Panel Conditioning Effect			
Interval estimates			
Decomposition 1	[1.10; 24.27*]	[-7.92; 15.24*]	[3.85*; 15.87*]
Decomposition 2	[0.90; 19.70*]	[-6.43; 12.38*]	[3.44*; 14.16*]
Panel Conditioning Effect			
Point estimates			
Ass. 1d, Decomp. 2	5.96*	6.31*	5.20*
Ass. 2a/b, Decomp. 1/2	5.85*	5.88*	5.20*

*=significant at 5% level

Conclusion 5.5

In this chapter we have analyzed panel conditioning effects on the estimates of binary outcome probabilities in two-wave panel surveys, using a refreshment sample and allowing for selective attrition. We introduced two definitions of a panel conditioning effect, based upon different decompositions of the total bias induced by estimating the time 2 distribution of the variable of interest into a panel conditioning effect and an attrition bias. We have shown that without additional assumptions, point identification of the panel conditioning effect (or the attrition bias) is not possible, but the panel conditioning effect is identified up to a bounding interval. How informative this bounding interval is will depend on its width, which is driven by the attrition rate. In many cases, the attrition rate will be so large that meaningful inferences are not possible without making further assumptions. We also introduced several additional assumptions on the attrition process, and showed how they guarantee point identification of the panel conditioning effect. Which of these assumptions are most plausible has to be studied on a case to case basis. Our empirical illustrations give results that are the same for each of the assumptions, which gives some confidence in the robustness of our procedure.

Applying our method to various empirical examples, we found that the problem of panel conditioning plays a role in knowledge questions, and not in questions on attitudes, actual behavior, or expectations concerning the future. For three out of four knowledge questions we studied, we found a significant panel conditioning effect under either of the additional assumptions guaranteeing point identification. In one case, the bounding interval analysis showed that the effect remained significant even without making such an assumption. In all cases the panel conditioning effect was positive, suggesting that some people who have had the question once, are triggered to increase their knowledge about the phenomenon in the question before taking part in the next survey.

Although what we have presented only concerns the marginal distribution of a binary outcome, extending the approach to non-binary outcomes is straightforward. This also applies to extending it to conditional distributions given time invariant covariates X like race, birth year or gender. Such extensions may also be useful because they change the additional assumptions and may make them more plausible - assuming that attrition is independent of health knowledge, for example, seems less plausible than assuming it is independent of health knowledge conditional on a given age and education level. Future work is needed if we want to consider a population that changes over time (such as a specific age group, with entry and exit) or on the lagged value of the variable of interest. Particularly the latter seems a limitation of our study, since it prevents us from analyzing individual changes.

The conclusion that for most types of questions no evidence of panel conditioning is found seems reassuring. One reason may be that interviewer effects are excluded, since our panel is an Internet panel. This is in line with the finding of Van der Zouwen and Van Tilburg (2001) who find that panel conditioning is mainly caused by interviewer behavior. Of course this needs to checked further, with more examples than the ones we have analyzed here, before a general conclusion can be drawn. For questions concerning knowledge, panel conditioning seems an issue that researchers need to be aware of. Refreshment samples are a useful tool to do this. Even without concerns about panel conditioning, refreshment samples were already shown to be useful tools to analyze selective attrition (Hirano et al., 2001). Thus this chapter supports the conclusion that for survey designers, a solid and sizable refreshment sample may be as important as reducing attrition by another fraction of a percentage point.

Appendix A: Examples 5.6

This appendix presents five numerical examples in which we demonstrate the use of interval and point estimates for measuring panel conditioning and attrition bias in two-wave data sets. Results for assumption 1a and 1b are not presented, since for these assumptions the attrition bias is zero (by definition) and the panel conditioning effect is equal to the total bias (for decomposition 1 and 2). Results for assumption 2b (decomposition 2) can be found in the row for assumption 2a (decomposition 1) since these are identical. Standard errors for the estimates were calculated using the Central Limit Theorem and the Delta-method.

Example A

-Fieldwork: November 2003 and November 2005
-n_1=1510 (Sample 1), n_P=1226 (Panel Sample), n_A=284 (Attrition Sample), n_R=891 (Refreshment Sample)
-Variable 1: Do you know what “Campylobacter" is (Yes/No); 19.3% answered ’Yes’ at time 1 (n_1) and 25.5% at time 2 (n_P+n_R).
-Variable 2: Do you know what “Salmonella" is (Yes/No); 96.8% answered ’yes’ at time 1 and 95.1% at time 2.
-Variable 3: Do you know what “Cross-infection" is (Yes/No); 55.7% answered ’yes’ at time 1 and 67.1% at time 2.

Table 5.2: Total Bias, Panel Conditioning and Attrition Bias for Three Knowledge Questions in a Questionnaire about Hygiene (in %)

	Knowledge Campylobacter		Knowledge Salmonella		Knowledge Cross-infection	
Total Bias						
Estimate	6.17^*		-1.61		6.71^*	
Decomposition 1						
	PC_1	AB_1	PC_1	AB_1	PC_1	AB_1
Interval Estimate	$[1.10, 24.27]$	$[-18.10, 5.07]$	$[-5.55, -0.71]$	$[-0.91, 3.93]$	$[-7.92, 15.24]$	$[-8.53, 14.64]$
Ass. 2a	5.85^* (1.86)	0.32 (0.48)	-1.76 (1.93)	0.15 (0.23)	5.88^* (2.08)	0.83 (0.62)
Decomposition 2						
	PC_2	AB_2	PC_2	AB_2	PC_2	AB_2
Interval Estimate	$[0.90, 19.70]$	$[-13.53, 5.28]$	$[-19.38, -0.58]$	$[-1.04, 17.77]$	$[-6.43, 12.38]$	$[-5.66.15, 13]$
Ass. 1d	5.96^* (1.86)	0.21 (0.32)	-1.69 (0.92)	0.07 (0.11)	6.31^* (2.06)	0.40 (0.30)

*=null hypothesis bias=0 is rejected at 5%-level, standard errors are reported between parentheses

Example B

-Fieldwork: November 2005 and January 2006
-n_1=1954, n_P=1888, n_A=66, n_R=481
-Variable 1: How much meat do you eat in a regular week (less than 5 times/5 or more); 44.6% answered 'less than 5 times' at time 1 (n_1) and 43.4% at time 2 (n_P+n_R).
-Variable 2: How much bird do you eat in a regular week (less than 1 time/1 or more); 12.6% answered 'less than 1 time' at time 1 and 13.1% at time 2.

Table 5.3: Total Bias, Panel Conditioning and Attrition Bias for Two Behavior Questions in a Questionnaire about the Bird Flue (in %)

	Number of Meals with Meat		Number of Meals with Poultry	
Total Bias				
Estimate	-1.69		1.86	
Decomposition 1				
	PC_1	AB_1	PC_1	AB_1
Interval Estimate	$[-3.25, 0.24]$	$[-1.93, 1.56]$	$[1.46, 4.95]$	$[-3.09, 0.41]$
Ass. 2a	-1.40 (2.54)	-0.29 (0.21)	1.79 (1.66)	0.07 (0.13)
Decomposition 2				
	PC_2	AB_2	PC_2	AB_2
Interval Estimate	$[-3.14, 0.23]$	$[-1.93, 1.45]$	$[1.41, 4.79]$	$[-2.92, 0.46]$
Ass. 1d	-1.51 (2.53)	-0.18 (0.13)	1.81 (1.66)	0.50 (0.09)

null hypothesis bias=0 is never rejected at 5%-level, standard errors are reported between parentheses

Example C

-Fieldwork: February 2004 and February 2005.
-n_1=1322, n_P=1170, n_A=152, n_R=598 for variable 1 and 2
-n_1=1734, n_P=1548, n_A=186, n_R=701 for variable 3 (due to routing sample sizes are different for variable 3)
-Variable 1: Have you thought about your pension last year (Yes/No); 40.6% answered 'yes' at time 1 (n_1) and 35.0% at time 2 (n_P+n_R).
-Variable 2: Have you received a working disability pension (Yes/No); 9.8% answered 'yes' at time 1 and 9.6% at time 2.
-Variable 3: Have you ever heard of a foundation named "Stichting Pensioenkijker (a foundation about pensions)" (Yes/No); 7.6% answered 'yes' at time 1 and 14.9% at time 2.

Table 5.4: Total Bias, Panel Conditioning and Attrition Bias for a Behavior, Fact, and Knowledge Question in a Questionnaire about Pensions (in %)

	Think about Pension		Receive a Disability Pension		Heard of StPensioenkijker	
Total Bias						
Estimate	3.88		-0.42		5.20*	
Decomposition 1						
	PC_1	AB_1	PC_1	AB_1	PC_1	AB_1
Interval Estimate	$[0.31, 12.68]$	$[-8.80, 4.19]$	$[-1.71, 9.44]$	$[-9.87, 1.29]$	$[3.85^{**}, 15.87]$	$[-10.66, 1.35]$
Ass. 2a	3.33 (2.38)	0.58 (0.48)	-0.34 (1.47)	-0.86 (0.30)	5.20* (1.53)	0.00 (0.22)
Decomposition 2						
	PC_2	AB_2	PC_2	AB_2	PC_2	AB_2
Interval Estimate	$[0.28, 11.22]$	$[-7.34, 4.16]$	$[-1.51, 10.03]$	$[-10.45, 1.09]$	$[3.44^{**}, 14.16]$	$[-8.96, 1.77]$
Ass. 1d	3.63 (2.36)	0.25 (0.21)	-0.35 (1.47)	0.07 (0.26)	5.20* (1.52)	0.00 (0.06)

*=null hypothesis bias=0 is rejected at 5%-level, standard errors are reported between parentheses
**=null hypothesis 'left bound *PC* interval'=0 is rejected at 5%-level

Example D

-Fieldwork: May 2006 and June 2006
-n_1=1033, n_P=938, n_A=95, n_R=449 for variable 1
-n_1=1040, n_P=943, n_A=97, n_R=451 for variable 2
-n_1=468, n_P=433, n_A=35, n_R=244 for variable 3 (due to item non-response sample sizes are different for each variable)
-Variable 1: Do you expect that pensions will be less in the future (Yes/No); 62.0% answered 'yes' at time 1 (n_1) and 60.6% at time 2 (n_P+n_R).
-Variable 2: Are you satisfied with your (future) pension (Yes/No); 28.9% answered 'yes' at time 1 and 30.4% at time 2.
-Variable 3: Do you have the possibility of a part-time pension (Yes/No); 47.0% answered 'yes' at time 1 and 43.8% at time 2.

Table 5.5: Total Bias, Panel Conditioning and Attrition Bias for an Expectation, Attitude, and Fact Question in a Questionnaire about Pensions (in %)

	Pensions will be less		Satisfaction Pension		Possibility Part-Time Pension	
Total Bias						
Estimate	-3.76		-2.24		-1.72	
Decomposition 1						
	PC_1	AB_1	PC_1	AB_1	PC_1	AB_1
Interval Estimate	$[-10.17, 0.04]$	$[-3.72, 6.41]$	$[-5.52, 4.77]$	$[-7.00, 3.28]$	$[-5.33, 2.76]$	$[-4.47, 3.61]$
Ass. 2a	-3.00 (2.79)	-0.76 (0.46)	-2.56 (2.65)	0.33 (0.44)	-1.82 (3.98)	0.10 (0.66)
Decomposition 2						
	PC_2	AB_2	PC_2	AB_2	PC_2	AB_2
Interval Estimate	$[-9.23, 0.04]$	$[-3.73, 5.47]$	$[-5.01, 4.32]$	$[-6.56, 2.77]$	$[-4.93, 2.55]$	$[-4.27, 3.21]$
Ass. 1d	-3.47 (2.78)	-0.29 (0.18)	-2.41 (2.64)	0.18 (0.24)	-1.76 (3.96)	0.04 (0.28)

null hypothesis bias=0 is never rejected at 5%-level, standard errors are reported between parentheses

Example E

-Fieldwork: November 2004 and December 2004.
-n_1=1435, n_P=1400, n_A=35, n_R=688
-Variable 1: What is your attitude towards Turkey joining the EU (Positive/Negative); 58.5% answered 'positive' at time 1 (n_1) and 63.7% at time 2 (n_P+n_R).
-Variable 2: When do you think Turkey will join the EU (Less than 10 years/10 years or more); 44.9% answered 'less than 10 years' at time 1 and 38.1% at time 2.
-Variable 3: Do you think immigration ia an important issue associated with Turkey joining the EU (Yes/No); 45.8% answered 'yes' at time 1 and 44.1% at time 2.

Table 5.6: Total Bias, Panel Conditioning and Attrition Bias for an Expectation and Two Attitude Questions in a Questionnaire about Turkey joining the EU (in %)

	Attitude Turkey joins EU		Period Turkey will join EU		Importance Immigration	
Total Bias						
Estimate	1.79		-4.35		1.21	
Decomposition 1						
	PC_1	AB_1	PC_1	AB_1	PC_1	AB_1
Interval Estimate	$[0.22, 2.72]$	$[-0.94, 1.56]$	$[-5.37, -2.87]$	$[-1.48, 1.02]$	$[-2.33, 0.17]$	$[-1.38, 1.12]$
Ass. 2a	1.82 (2.25)	0.04 (0.20)	-4.18 (2.27)	-0.16 (0.21)	-1.14 (2.32)	-0.07 (0.21)
Decomposition 2						
	PC_2	AB_2	PC_2	AB_2	PC_2	AB_2
Interval Estimate	$[0.22, 2.66]$	$[-0.87, 1.57]$	$[-5.24, -2.80]$	$[-1.55, 0.89]$	$[-2.28, 0.17]$	$[-1.38, 1.07]$
Ass. 1d	1.81 (2.24)	0.02 (0.11)	-4.26 (2.27)	0.08 (0.11)	-1.18 (2.31)	0.02 (0.07)

null hypothesis bias=0 is never rejected at 5%-level, standard errors are reported between parentheses

6 Design Effects in Web Surveys: Comparing Trained and Fresh Respondents

ABSTRACT In this chapter we investigate whether there are differences in design effects between trained and fresh respondents. In three experiments, we varied the number of items on a screen, the choice of response categories, and the layout of a five point rating scale. We find that trained respondents are more sensitive to satisficing and select the first acceptable response option more often than fresh respondents. Fresh respondents show stronger effects with regard to verbal and non-verbal cues than trained respondents, suggesting that fresh respondents find it more difficult to answer questions and pay more attention to the details of the response scale in interpreting the question.

SCRIBERE SCRIBENDO, DICENDO DICERE DISCES

6.1 Introduction

Socio-economic panel surveys, where the same households or individuals are interviewed repeatedly at various points in time, have important advantages over independent cross-sections, such as efficiency gains in recruiting, reduced sampling variation in the measurement of change, and the possibility to analyze behavior at the individual respondent level (see, e.g. Baltagi 2001). However, the fact that experienced panelists may respond differently than panelists without experience ("panel conditioning"), raises concern over survey quality. In particular, many researchers fear that online survey panels, where respondents are interviewed at a high frequency such as once a month or more, create trained respondents. Brannen (1993) suggests that the issue of the effects of surveying on respondents has been more a matter of speculation than of empirical investigation. Suggestions on how to treat trained respondents are increasing rapidly on the Internet (as shown by, e.g., searching the web for 'professional respondents' or 'data quality'; see also, for example, www.comscore.com, www.quirks.com, www.hisbonline.com). Although commercial companies address the issue of trained respondents in web surveys, there appears to be little empirical research to date on the effect of prior survey participation on survey answers.

Trained respondents may answer questions differently than those with little or no experience as panelist. This can be due to changes in behavior or knowledge induced by previous surveys (e.g. because respondents acquire knowledge on topics addressed in a previous survey), as well as to changes in the question-answering process. Panel members may learn from taking surveys. They may prepare for future surveys and increase their knowledge on the topics addressed, or develop attitudes towards certain topics. In addition, they may become familiar with the question-answering process, learn how to interpret questions, and make fewer errors than new respondents. Or, conversely: experienced respondents may answer strategically to avoid follow-up questions and to reduce the burden of their task or accelerate the completion of the survey, thereby making more errors than fresh respondents.

This chapter addresses the issue of procedural learning from taking surveys: the question-answering process. Trained respondents may react differently to web survey design choices than inexperienced respondents. First, they may be able to process more information on a screen and, for example, make fewer errors when multiple items are presented on a single screen. Second, they may be less or more susceptible to social desirability bias and less or more reluctant to select a response category that seems unusual in the range of responses. Third, they may react differently to (changes in) question layout. The goal of this study is to explore differences in web design effects between trained and fresh respondents in these three aspects.

The remainder of this chapter is organized as follows. Section 6.2 addresses the background of design effects and panel conditioning, while Section 6.3 discusses the design and implementation of our experiments. Section 6.4 presents the results. This section is divided into three subsections to separately discuss each of the three experiments (items per screen, answer categories, and layout). In each subsection we discuss whether a design effect is found, to subsequently compare trained and fresh respondents with regard to this effect.

Background 6.2

Survey experience may influence responses to survey questions. In ongoing household panels, one could in principle test whether the time since respondents entered the panel (the duration) or the number of surveys in which they have participated affects responses. However, in most panels almost none of the respondents are completely fresh, while the effect of panel experience may possibly be non-linear, with a noticeable difference between no and some experience, but much less or no effect when going from some to more experience. Bartels (1999) argues that panel surveys should routinely include parallel fresh cross-sections, to provide a solid basis to assess (and adjust for) biases arising from re-interviewing. In most panel surveys, comparable data from a fresh cross-section are not available.

Literature shows that answers to questions in (web) surveys are affected by design choices, such as the ordering of questions (see e.g. Couper et al. 2000; Krosnick and Alwin 1987; Toepoel et al. 2009a), the categorical answers that the respondent can choose from (see e.g. Rockwood et al. 1997; Schwarz et al. 1985), or the layout of the questions (see e.g. Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Toepoel et al. 2006; Winter 2002a, Winter 2002b). Some studies have also analyzed whether such design effects vary with respondent characteristics such as age, gender, or education level (see e.g. Fuchs 2005; Knauper et al. 2004; Krosnick and Alwin 1987; Stern et al. 2007; Tourangeau et al. 2007), or attitudes such as a *need for cognition* or *need to evaluate* (see e.g. Toepoel et al. 2009b). Despite the growing empirical support for (web) design effects, there exists virtually no reference to respondents' experience in answering surveys. As a result, empirical tests have not taken into account how experience may affect the question-answering process in web surveys. In this study, we analyze the differences in web design effects between experienced and fresh panel respondents.

6.2.1 Experience and the response process

Van der Zouwen and Van Tilburg (2001) find that panel conditioning effects sometimes arise and sometimes not, without a clear indication of the situations in which these effects occur. Trivellato (1999) concludes that panel participation mainly affects the way in which behavior is reported (response process), while it does not have pervasive effects on behavior itself. Coombs (1973) and Das et al. (2007) find that panel conditioning arises for knowledge questions, but not for other types of questions. Sturgis et al. (2007) formulate a theory for panel conditioning: the cognitive stimulus hypothesis. Questions about certain topics may induce respondents to reflect on them after the survey has ended, to talk about them with friends and relatives, and to acquire additional information. Golob (1990) concludes that no panel conditioning effects exist in questions that require simple reporting tasks, suggesting instead that panel conditioning relates to the cognitive difficulty in answering questions. He finds no panel conditioning on car ownership variables that are measured using simple reporting requirements, but he does find panel conditioning effects for more cognitively demanding questions such as travel times for different modes of transport. Van der Zouwen and Van Tilburg (2001), on the other hand, conclude that panel conditioning does not take place through cognitive processes within the respondent's mind but through the task-related behavior of the interviewer.

Mathiowetz and Lair (1994) find evidence that respondents become familiar with the question-answering process and adjust their responses accordingly. They hypothesize that an improvement in daily life activities noted in a subsequent survey wave was a function of panel conditioning. Respondents learned in wave 1 that for every reported difficulty there was a series of follow-up questions, and they therefore altered their responses in the subsequent wave to avoid the follow-up questions. Meurs et al. (1989) also find that experienced respondents respond strategically, for instance after learning that answering "no" means evading follow-up questions, thereby reducing the burden of their task.

Trained respondents may be more sensitive to social desirability bias than fresh respondents. Sharpe and Gilbert (1998) find that repeated testing increases the scores on the Beck depression scale and attribute this to socially desirable response behavior, triggered by the first interview. Chan and McDermott (2007) and Wang et al. (2000) find similar effects.

Coen et al. (2005) compare frequent and non-frequent respondents. They find evidence that responses of frequent responders are more in line with actual consumer behavior than responses of less frequent responders. This finding is in contrast to the conventional view that past experience is not desirable with regard to measurement errors (Bartels 1999; Brannen 1993; Golob 1990; Mathiowetz and Lair 1994; Meurs et al. 1989; Sharpe and Gilbert 1998; Sturgis et al.

2007; Williams 1970; Williams and Mallows 1970). Coen et al. (2005) find no evidence that frequent responders try to speed through the survey. In fact they find a relatively high number of marks on check-all-that-apply questions. Inexperienced panelists more often choose socially desirable answers. This is in line with the results of Dennis (2001). Coen et al. (2005) also demonstrate that experience (number of surveys completed) is more associated with response behavior than duration (the length of time on the panel).

Experience and web survey design 6.2.2

There is a growing literature that suggests that the design of a web survey has a significant impact on measurement error (see e.g. Christian and Dillman 2004; Couper et al. 2000; Dillman 2007; Dillman and Christian 2002; Tourangeau et al. 2004, Tourangeau et al. 2007). Design may be more important in web surveys than in other modes of administration, because there are many tools available and because of potential variation in how the survey appears on a screen. Couper (2000) concludes that more work is needed to determine the optimal designs for different groups of people, emphasizing the need for research on panel conditioning and web page design.

Despite the widespread use of online panels, there appears to be no empirical research to date on the difference in response effects between trained and fresh respondents. There are some papers offering suggestions on questionnaire design in relation to prior survey experience in general. Trivellato (1999), for example, offers a number of strategies with regard to initial and follow-up sampling, panel length and number of waves, and to tracking and tracing techniques to locate respondents to maintain high participation rates. Moreover, he outlines questionnaire design strategies such as the sequence of questions, probing, skip patterns, and consistency checks to limit response errors. He also recommends a low-frequency measuring of variables that are reasonably stable over time, preferably in the first interview. Web surveys are particularly suited to implementing Trivellato's suggestions thereby improving the longitudinal consistency of the data. This chapter addresses three design issues in which trained and fresh respondents may differ.

Items per screen 6.2.3

For web questionnaires, interface design varies in terms of the distribution of questions on the screen and the navigation methods used. At one end of the design continuum are form-based designs that present questionnaires as one long form in a scrollable window, at the other end are screen-by-screen questionnaires that present only a single item at a time (Norman et al. 2001). Presenting questions in a matrix is somewhere in between, reducing the number of screens without the need for scrolling.

The grouping of related items on a single screen is likely to lead respondents to view the items as related entities, thus increasing the correlation among them (Dillman 2007; Schwarz 1996; Strack et al. 1991; Sudman et al. 1996; Tourangeau et al. 2004, Tourangeau et al. 2007). Couper et al. (2001) conclude that correlations are consistently higher among items appearing together on a screen than among items separated across several screens. However, the overall effect is not large, and none of the differences between pairs of correlations reach statistical significance. Tourangeau et al. (2004) replicate the above findings. Respondents seem to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. Peytchev et al. (2006) find few differences between paging and scrolling designs.

Non-response and time to complete the interview can also be indicators for the optimal number of items per screen. Lozar Manfreda et al. (2002b) find that a one-page design results in higher item non-response. Couper et al. (2000), Lozar Manfreda et al. (2002b), and Tourangeau et al. (2004) find that a multiple-item-per-screen design takes less time to complete than a one-item-per-screen design. Evaluation questions can show whether respondents are comfortable with a particular survey design. Toepoel et al. (2009a) find that placing more items on a screen negatively influences the respondent's evaluation of the layout.

We are not aware of any studies on the optimal number of items on a screen in relation to survey experience. Our conjecture is that trained respondents can process more information on a screen, thus showing less item non-response when more items are placed on a single screen than fresh respondents. We expect them to complete the survey faster than fresh respondents, especially if many items are placed on a screen. We also expect them to better evaluate a large number of items on a screen than fresh respondents.

6.2.4 Response categories

Studies on the cognitive and communicative processes involved in answering survey questions suggest that the choice of response categories can have a significant effect on the answers. Toepoel et al. (2009b), Winter (2002a); and Winter (2002b) find response category effects in web surveys, while Krosnick and Alwin (1987), Rockwood et al. (1997), Schwarz et al. (1985), Schwarz and Hippler (1987), and Strack and Martin (1987) find effects in other modes of administration. Schwarz and Hippler (1987) argue that respondents use the response alternatives to determine the meaning of the question and use the frequency range as a frame of reference, presuming the values stated in the scale to be commonly held values. In other words, a respondent may be reluctant to select a response category that seems unusual in the range of responses. This results in higher estimates along scales that present high rather than low ranges. The literature suggests that response categories have a significant effect on responses to questions for which estimation is likely to be used in recall, whereas in questions in which direct recall is used in response formatting the response categories do not have a significant effect.

Choquette and Hesselbrock (1987) suggest that respondents attempt to present themselves more favorably in later waves. This would lead to the conjecture that trained respondents are more prone to social desirability bias and more reluctant to select a response category that seems unusual in the range of responses. On the other hand, Coen et al. (2005) and Dennis (2001) find that inexperienced panelists more often choose socially desirable answers. Survey experience may also make the respondents less uncertain and thus less susceptible to social desirability bias. The second experiment in this chapter assesses the impact of a response scale on both trained and fresh respondents.

Layout 6.2.5

Differences in question layout can lead to detectable differences in responses to survey questions (see, e.g. Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Schwarz and Hippler 1987; Toepoel et al. 2006; Tourangeau et al. 2004). A question format contains verbal and nonverbal cues that influence respondent behavior. Nonverbal cues include graphical, numerical and symbolic languages that convey meaning in addition to the verbal language (Dillman and Christian 2002). Jenkins and Dillman (1997) have developed a conceptual framework to explain how visual languages may influence respondent behavior.

Redline et al. (2003) confirm that the visual and verbal complexity of information in a questionnaire affects what respondents read, the order in which they read it, and ultimately, their comprehension of the information. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales (graphical manipulation) in paper questionnaires do not elicit the same responses. Schwarz et al. (1985) show that respondents gain information about the researcher's expectations using numerical labels as frames of reference. Schwarz et al. (1991) find that changing the numerical values attached to scales changes the answers, and that respondents hesitate to assign a negative score to themselves in a face-to-face interview: a scale with numbers 0-10 results in lower scores than a -5 to 5 format.

We expect trained panelists to be more sensitive to layout choices than fresh panelists. They may be used to a particular question format so that changing that format (e.g. from disagree-agree to agree-disagree) may not be noticed. In addition, we expect them to be more sensitive to added numerical labels and signs than fresh respondents.

6.3 Design and implementation

To study design effects on trained and fresh respondents, we used two online household panels administered by CentERdata. The first, the CentERpanel (see also http://www.centerdata.nl/en/CentERpanel), has existed for 17 years. Panel members fill out questionnaires every week. Panel duration of respondents ranges from seventeen years to a few months. The second panel is the LISS-panel (see http://www.centerdata.nl/en/LISSpanel). Our experiments were the very first questionnaire for this panel. Both panels are designed to be representative for the Dutch population. Thus, the CentERpanel consists of trained respondents (varying in panel duration, with a mean duration of 6 years and 8 months, standard deviation equals 4 years), while the LISS-panel consists of completely fresh respondents.

We fielded the questionnaire in June 2007. In the CentERpanel, 1356 panel members were selected to fill out the questionnaire; 981 respondents (72.3%) responded. In the LISS-panel, 4530 panel members were selected; 2809 respondents (62.0%) filled out the questionnaire. To correct for differences due to nonresponse, we used weights based on gender, age and education.

The questionnaire consisted of three different experiments. In the first, we used the Marlowe-Crowne Social Desirability Scale (the 10-item version of Strahan and Gerbrasi (1972) and varied the number of items per screen. We used three groups, with 1, 5, and 10 items per screen. We added some questions to determine whether respondents react differently to the number of items displayed per screen.

In the second experiment we varied the answer scale in four questions. We used the same questions as Toepoel et al. (2009b), varying in cognitive difficulty. We used a low response scale, a high response scale, and an open-ended format.

In the third experiment we varied the layout of a five-point rating scale. The first group was presented answer categories in a linear vertical format from positive to negative (excellent, very good, good, fair, and poor). Five other groups were presented with different manipulations. The second group answered from negative to positive, the third in a horizontal format, for the fourth group we added numbers 1 to 5 to the response categories, for the fifth group numbers 5 to 1, and for the sixth group numbers 2 to -2.

6.4 Results

In this section we discuss the results of the three experiments. For each experiment, we first discuss the response effect and then compare the answers of trained and fresh respondents.

Response effect: items per screen 6.4.1

We found differences in inter-item correlations when the items were presented (1) one-item-per-screen (Cronbach's alpha of .473 for the trained panel and .528 for the fresh panel), (2) 5-items-per-screen (alpha of .602 for the trained panel and .516 for the fresh panel), and (3) 10-items-per-screen (alpha of .515 for the trained panel and .498 for the fresh panel).

In principle, the web survey software can force the respondent to give a response. If a respondent fails to give an answer, he/she would then be presented with an error message indicating a need to choose an answer. We deliberately did not program this feature, so that respondents could proceed without filling in answers. We found no significant differences in item non-response when more items were placed on a single screen in the trained panel. In the fresh panel, the more items were placed on a single screen, the lower the item non-response (F=3.795, p=.023). This is contrary to the findings of Lozar Manfreda et al. (2002b).

If more items are placed together on one screen, fewer physical actions (keystrokes or mouse clicks) are required than when items are presented separately. Therefore, we expected that placing more items on a single screen would reduce the time needed to complete the questionnaire. However, we found no significant differences in mean duration between formats (1, 5, and 10 items per screen) in both panels.

Respondents answered some evaluation questions about the social desirability questions:

1. How interesting did you find the questions?

2. How would you evaluate the duration?

3. How clear did you find the wording of the questions?

4. How easy was it to answer the questions?

5. What did you think of the layout?

6. What is your overall opinion of these questions?

These questions were asked on a ten-point scale ranging from 1 ('very poor'/'not at all') to 10 ('very good'/'very much'). In the trained panel we found a significant effect of format in question 4, with the 5-items-per-screen format receiving the highest rating (F=3.32, p=.037). This suggests that respondents found that the 10-items-per-screen format contained too much information, while the 1-item-per-screen format contained too many screens. The fresh panel also preferred the layout of the 5-items-per-screen format to other formats (F=3.816, p=.022).

The counting of all ten social desirability items resulted in an overall score of social desirability. Neither the trained nor the fresh panel showed differences in social desirability scores between the 1, 5, and 10-item-per-screen format.

6.4.2 Comparison of trained and fresh respondents: items per screen

Trained respondents had higher inter-item correlations for multiple-items-per-screen formats, while fresh respondents showed the highest inter-item correlation in the one-item-per-screen version. Trained panelists seem to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. Fresh panelists may be triggered by the new experience of participating in a survey and therefore read each item more carefully.

We found no significant difference in item non-response between trained and fresh respondents; 1.2% (12 out of 981 respondents) had one or more items missing in the trained panel, compared to 1.5% (42 out of 2809 respondents) in the fresh panel. Linear regression of item non-response on the number of items per screen, a dummy for panel (trained versus fresh), and the interaction between these two showed no significant interaction effect.

There was a difference in mean duration of the entire survey[1] between panels (t=-2.4, p=.016): 436 seconds for the trained panel and 576 seconds for the fresh panel. The mean duration to complete just the ten social desirability items did not differ significantly between panels. Linear regression of the duration of the survey on the number of items per screen, a dummy for panel, and the interaction between these two showed no significant interaction effect either.

Although this chapter discusses design effects, we also looked at the mean score of the Social Desirability Scale used for the items-per-screen experiment. In contrast to Choquette and Hesselbrock (1987), we found no evidence for social desirability bias for trained respondents. The mean scores of the Social Desirability Scale in the two panels were not significantly different (F=2.16, p=.642).

6.4.3 Response effect: response categories

To assess the impact of a response scale on respondents' answers, we asked four questions on the frequency of various activities with a randomized answering format: a low response scale, a high response scale, and an open-ended format. See Appendix A for the questions and response scales used. We dichotomized answers to compare the results.

We found a scale range effect (see Tourangeau et al. 2000): the range of the response scale affected respondents' frequency reports. Table 6.1 shows that

[1] The questionnaire consisted of all experiments discussed in this chapter.

20% of the trained respondents who were presented the low response scale reported watching TV for more than two and a half hours, compared to 51% of the trained respondents who were presented the high response scale. In comparison, 46% of the trained respondents who were presented the open-ended question reported watching TV for more than two and a half hours. Similar results were found for the fresh panel. Table 6.2 shows an overview of the correlations between answer score (1 if more than the reference level, 0 otherwise) and response format for the different question types. A higher correlation coefficient (η) between the answer score and the scale used indicates a larger effect of the response scale. With the high versus low response scale, the largest correlation between the answer score and the scale is found in hours watching TV (difficult to process), the lowest for days on holiday (easy to process). As expected, the effect of response scales depends on how well a behavior is presented in memory. More details of this experiment on response category effects can be found in Toepoel et al. (2009b).

Table 6.1: Overview of Frequencies (in %) from Different Response Formats for the Trained and Fresh Panel

	Low Response Scale		High Response Scale		Open-ended	
	Trained panel	Fresh panel	Trained panel	Fresh panel	Trained panel	Fresh panel
	more than X*	more than X*	more than X*	more than X*	more than X*	more than X*
Hours watching TV	20	18	51	49	46	44
Birthday parties	24	28	40	41	42	44
Visiting a hairdresser	14	17	28	33	25	21
Days on holiday	35	41	44	45	45	43

*X=two and a half for hours watching TV, nine for visiting a hairdresser, and 17 for birthday parties and days on holiday

6.4.4 Comparison of trained and fresh respondents: response categories

We found an effect of response categories on answers, but this effect is not significantly different for trained and fresh respondents. For none of the questions we found a significant interaction effect between format and panel. Our conjecture that trained respondents are more prone to social desirability bias and

Table 6.2: Overview of Correlations between Answer Score and Response Format

	High Response Scale versus Low Response Scale		Low Response Scale versus Open-ended		High Response Scale versus Open-ended	
	Trained panel	Fresh panel	Trained panel	Fresh panel	Trained panel	Fresh panel
	η	η	η	η	η	η
Hours watching TV	.329 (p<.0001)	.325 (p<.0001)	.267 (p<.0001)	.243 (p<.0001)	.062 (p=.137)	.067 (p<.0001)
Birthday parties	.168 (p<.0001)	.137 (p<.0001)	.505 (p<.0001)	.482 (p<.0001)	.352 (p<.0001)	.358 (p<.0001)
Visiting a hairdresser	.182 (p<.0001)	.180 (p<.0001)	.136 (p=.002)	.044 (p=.001)	.045 (p=.225)	.133 (p<.0001)
Days on holiday	.089 (p=.056)	.050 (p=.102)	.097 (p=.036)	.019 (p=.548)	.008 (p=.830)	.031 (p=.348)

Note: A higher correlation coefficient (η) between the answer score and the scale that was used indicates greater differences between response scales.

more reluctant to select a response category that seems unusual in the range of responses was not confirmed. The conjecture that survey experience may make the respondents less uncertain and thus less susceptible to social desirability bias was not confirmed either.

6.4.5 Response effect: layout

In our third experiment, we manipulated the layout of a five-point rating scale using verbal and non-verbal manipulations. Appendix B presents the question that was asked and shows the answer distributions for all formats for both panels. Table 6.3 shows that the distributions of the answers in a negative-positive format differ significantly from those in a positive-negative format (verbal manipulation: 1 versus 2). Respondents selected the response option 'very good' less often when it was presented as a fourth alternative. No significant differences were found for the graphical manipulation (changing the layout from vertical to horizontal), i.e., comparing format 1 versus 3. Adding numbers 1 to 5 to the scale did not lead to significant differences in answer scores either, suggesting that respondents take a numbering beginning with 1 as a kind of default labeling that does not convey much information about the meaning of the scale points. Comparing adding the numbers 5 to 1 to 1 to 5 (formats 4 and 5) did produce significant differences, indicating that respondents react to numbers as well as words in a numerical ordering not beginning with 1. The strongest effect was found when numbers 2 to -2 were added. This manipulation showed significantly different answer scores compared to all other manipulations. Respondents are apparently reluctant to assign negative scores. Negative numbers might be interpreted as implying more extreme judgments than low positive numbers (scale label effect, see Tourangeau et al. 2000; see also Tourangeau et al. 2007, who make a similar argument and provide additional evidence for the added attention that negative signs receive).

A Chi Square test and a difference of means test showed significant differences for all non-verbal manipulations (all formats except format 2), indicating that the layout of the answer categories influences the answers. Also, the overall test comparing all six formats showed significant differences between formats.

Comparison of trained and fresh respondents: layout 6.4.6

Although the third response option ’good’ has the same number (3) in formats 4 and 5, fresh respondents selected this answer significantly more often in format 4 (numbers 1 to 5: 53.4%) than in format 5 (numbers 5 to 1: 44.3%). The effect for trained respondents was much smaller. Apparently, fresh respondents extract information not only from the number itself but also from the ordering of numbers added to the verbal labels.

Although changing the layout from vertical to horizontal did not change the answer distributions significantly (see Table 6.3: 1 versus 3), trained respondents selected the second response ’very good’ more frequently than fresh respondents. The fresh respondents selected the response option ’fair’ more often in the horizontal format. This indicates a primacy effect for trained respondents and a recency effect for fresh respondents.

Combining all six formats and looking at the distribution of all answers, independent of the layout manipulations, we found a similar result: trained respondents more easily selected one of the first options, while fresh respondents more often selected one of the last options (χ^2=14.93, p=.01). A possible interpretation of this difference is that trained respondents are more sensitive to satisficing and therefore select the first satisfying response category more often (cf. Krosnick and Alwin 1987; and Tourangeau et al. 2000).

Linear regression explaining the answer to the question by dummies for the five format manipulations (with format 1 as reference level), a panel dummy, and interaction terms between the panel dummy and the five formats showed no significant interaction effect between panel experience and the five formats. However, the interaction effect between the panel dummy and the graphical manipulation (horizontal format) almost reached significance (t=1.83, p=.07).

Table 6.3: Chi square Tests and Differences of Means in the Different Manipulations

	Trained panel		Fresh panel	
	Chi Square Tests χ	Diff. of Means t	Chi Square Tests χ	Diff. of Means t
Verbal: 1 versus 2	13.901 (p=.016)	1.311 (p=.253)	23.430 (p<.0001)	14.834 (p<.0001)
Graphical: 1 versus 3	2.557 (p=.634)	1.829 (p=.177)	3.492 (p=.625)	1.594 (p=.207)
Numerical: 1 versus 4	4.477 (p=.483)	1.757 (p=.186)	5.743 (p=.332)	.310 (p=.578)
Numerical 4 versus 5	9.082 (p=.059)	7.081 (p=.008)	13.424 (p=.020)	9.509 (p=.002)
Numerical: 5 versus 6	16.337 (p=.006)	17.361 (p<.0001)	30.988 (p<.0001)	27.091 (p<.0001)
Overall across all non-verbal manipulations (except 2)	37.727 (p=.010)	F=5.399 (p<.0001)	67.840 (p<.0001)	F=8.871 (p<.0001)
Overall across all 6 formats	55.618 (p<.0001)	F=5.944 (p<.0001)	102.906 (p<.0001)	F=11.943 (p<.0001)

Note:
1 Reference: Linear Vertical Positive to Negative
2 Verbal: Linear Vertical Negative to Positive
3 Graphical: Linear Horizontal
4 Numerical: Linear Vertical with Numbers 1 to 5
5 Numerical: Linear Vertical with Numbers 5 to 1
6 Numerical: Linear Vertical with Numbers 2 to -2

6.5 Discussion and conclusions

Despite the growing empirical support for (web) design effects, there exists virtually no reference to respondents' experience in completing surveys. This means that empirical tests have not taken into account how experience may affect the question-answering process in web surveys. We have tried to gain more insight into the response processes of trained and fresh respondents. We did so by conducting three experiments on web survey design issues with two different panels: a new panel of fresh respondents, and a panel that has been in place for seventeen years now, thus consisting of respondents that have extensive experience. The web survey design issues we considered were the number of items per screen, response category effects, and layout effects.

First of all, the social desirability scale used to assess the impact of a 1, 5, and 10-item-per-screen format showed no difference in social desirability scores between the trained and fresh panel. A small effect with respect to inter-item correlations for multiple-items-per-screen formats was found, indicating that trained panelists use the proximity of the items as a cue to their meaning more than fresh panelists do. We did not find evidence that trained respondents are able to process more information on a screen, that is, that they show less item non-response when more items are placed on a single screen. They did complete the survey in less time than fresh respondents. Our analysis showed no interaction effect between the number of items per screen and panel experience on item non-response, time to complete the survey, and evaluation questions. We did not find evidence that the number of items per screen influences the answers respondents provide, but it does have an influence on respondents' evaluation of the questionnaire. Both the trained and the fresh panelists appreciated the 5-items-per-screen format the most. Keeping the respondent satisfied is important for panel maintenance, and therefore it is important to place more than one item of a battery on a screen, but not too many.

With regard to response category effects, we found no significant interaction effect between web survey design and panel experience either; our conjecture that trained respondents are more prone to social desirability bias and more reluctant to select a response category that seems unusual in the range of responses is not confirmed, but neither is the conjecture that survey experience may make the respondents less uncertain and thus less susceptible to social desirability bias.

Fresh panelists showed stronger effects than trained respondents with regard to the verbal and non-verbal cues in a five-point scale. We found no significant interactions between panel experience and layout manipulations. Our results show a primacy effect for trained respondents and a recency effect for fresh respondents, suggesting that trained respondents more often select the first acceptable response option than fresh respondents.

In summary, we found some evidence that survey experience influences the question-answering process. Trained respondents seem to be more sensitive to satisficing. The advantage of using trained respondents is that they are less sensitive to visual cues. Fresh respondents show stronger effects for details of the response scales than trained respondents, even though some features may simply be a matter of style rather than adding any meaning to the scale. They may be more uncertain which answer to select and therefore base their answers more often on cues in a questionnaire (see also Tourangeau et al., 2007, who make a similar argument for the greater impact of non-verbal cues for ambiguous questions). Survey researchers should pay attention to these differences between trained and fresh respondents, and additional research is needed to determine whether these conclusions hold in different settings.

Appendix A: Questions in experiment categories 6.6

Table 6.4: Questions and Response Scales Used in the Experiment

Response Scales	Format A	Format B	Format C
How many hours do you typically watch TV?			
1	$\frac{1}{2}$ hour or less	2 $\frac{1}{2}$ hours or less	open-ended
2	$\frac{1}{2}$-1 hour	$\frac{1}{2}$-3 hours	question
3	$1-1\frac{1}{2}$ hours	$3-3\frac{1}{2}$ hours	
4	$1\frac{1}{2}-2$ hours	$3\frac{1}{2}-4$ hours	
5	$2-2\frac{1}{2}$ hours	$4-4\frac{1}{2}$ hours	
6	more than $2\frac{1}{2}$ hours	more than $4\frac{1}{2}$ hours	
How many birthday parties do you typically attend per year?			
1	9 or less	17 or less	open-ended
2	9-11	17-19	question
3	11-13	19-21	
4	13-15	21-23	
5	15-17	23-25	
6	more than 17	more than 25	
How many times did you go to the hairdresser last year?			
1	1 or less	9 or less	open-ended
2	1-3	9-11	question
3	3-5	11-13	
4	5-7	13-15	
5	7-9	15-17	
6	more than 9	more than 17	
How many days did you leave your home (have a holiday) last year?			
1	9 or less	17 or less	open-ended
2	9-11	17-19	question
3	11-13	19-21	
4	13-15	21-23	
5	15-17	23-25	
6	more than 17	more than 25	

Note: answer categories one to five in Format A match answer category one in Format B. Answer category six in Format A matches answer categories two to six in Format B.

6.7 Appendix B: Results experiment layout

Table 6.5: Frequencies (in %), Number of Observations, and Mean Scores in Experiment 3: Layout Effects. Fresh panel between Parentheses.

%	1 Reference: Linear Vertical Positive to Negative	2 Verbal: Linear Vertical Negative to Positive	3 Graphical: Linear Horizontal	4 Numerical: Linear Vertical With Numbers 1 to 5	5 Numerical: Linear Vertical With Numbers 5 to 1	6 Numerical: Linear Vertical With Numbers 2 to -2
Excellent	.0 (.2)	1.6 (.0)	.0 (.5)	.8 (.0)	.6 (.2)	3.9 (.9)
Very Good	14.2 (11.1)	5.2 (4.5)	19.9 (8.6)	15.4 (9.3)	9.0 (7.6)	19.9 (13.3)
Good	42.1 (46.5)	49.4 (40.5)	40.7 (43.8)	46.5 (53.4)	42.2 (44.3)	45.8 (50.2)
Fair	36.8 (37.0)	33.8 (48.2)	34.9 (41.0)	33.8 (31.9)	38.1 (39.8)	25.1 (29.2)
Poor	6.9 (5.3)	10.1 (6.8)	4.5 (6.1)	3.5 (5.5)	10.1 (8.1)	5.3 (6.5)
N	162 (453)	181 (460)	159 (460)	172 (474)	138 (466)	162 (483)
Mean	3.36 (3.36)	3.46 (3.57)	3.24 (3.44)	3.24 (3.34)	3.48 (3.48)	3.08 (3.27)

Note: Scores for all versions are transformed back to the reference layout. Thus, a high mean score indicates a negative judgment.

7 Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents

ABSTRACT Panel conditioning arises if respondents are influenced by participation in previous surveys, such that their answers differ from the answers of individuals who are interviewed for the first time. Having two panels-a trained one and a completely fresh one-created a unique opportunity for analyzing panel conditioning effects. To determine which type of question is sensitive to panel conditioning, 981 trained respondents and 2809 fresh respondents answered nine questions with different question types. The results in this chapter show that panel conditioning only arise in knowledge questions. Answers to questions on attitudes, actual behavior, or facts were not sensitive to panel conditioning. The effect of panel conditioning in knowledge questions was bigger for questions where fewer respondents knew the correct answer.

USUS MAGISTER EST OPTIMUS

7.1 Introduction

Trained respondents may give different answers to survey questions than those with little or no experience in a panel. This can be due to behavior or knowledge changes induced by previous surveys (e.g. because respondents acquire knowledge on topics addressed in a previous survey) as well as to changes in the question-answering process. Panel members may learn from taking surveys. They may prepare for future surveys (increase their knowledge), or develop attitudes towards certain topics. In addition, they may become familiar with the question-answering process, learn how to interpret questions, and make fewer errors than new respondents. Or the opposite: experienced respondents may also make more errors than fresh respondents - they may more often speed through the survey or answer strategically to avoid follow-up questions. This chapter investigates which type of question is sensitive to panel conditioning, comparing the answers of fresh and experienced panel respondents to nine questions with different question types.

The remainder of this chapter is organized as follows. Section 7.2 discusses the background of the subject. Section 7.3 presents the design and implementation of the study. Section 7.4 shows the results, and Section 7.5 closes with concluding remarks.

7.2 Background

One of the basic decisions in survey design is whether to use trained respondents (using a panel) or fresh respondents (e.g. a repeated cross section). Sharot (1991) discusses advantages and disadvantages of panels. There are two important methodological issues associated with the use of panel surveys: panel attrition and panel conditioning. Panel conditioning arises if having been interviewed before causes differences in knowledge, behavior or attitude, affecting the answers in re-interviews.

Panel conditioning has been studied in many social sciences, with mixed findings. Duan et al. (2007), Meurs et al. (1989), Waterton and Lievesley (1989), Williams (1970), and Williams and Mallows (1970) found evidence for panel conditioning. On the other hand, Dennis (2001) and Clinton (2001) found little evidence for panel conditioning in the 'Knowledge Networks' panel (an online panel that is representative of the entire US population) and Pennell and Lepkowski (1992) found hardly any evidence of panel conditioning or attrition bias in income sources reported in the Survey of Income and Program Participation.

According to Van der Zouwen and Van Tilburg (2001), panel conditioning effects sometimes do and sometimes do not appear, without a clear indication of the conditions under which these effects occur. Sturgis et al. (2007) discuss a potential theory behind panel conditioning: the cognitive stimulus hypothesis. Questions about certain topics may induce respondents to reflect on them after the interview has ended, to talk about them with friends and relatives, or to acquire additional information. According to Trivellato (1999), panel participation mainly affects the way in which behavior is reported (response process), while it does not have pervasive effects on behavior itself. Coombs (1973) and Das et al. (2007) found that panel conditioning only arises for knowledge questions but not in other types of questions. Golob (1990) concluded that no panel conditioning effects exist in questions that require simple reporting tasks, implying that panel conditioning relates to the cognitive difficulty in answering questions. Van der Zouwen and Van Tilburg (2001), on the other hand, concluded that panel conditioning does not take place via cognitive processes within the respondent's mind but via the task-related behavior of the interviewer.

Design and implementation 7.3

To study the relation between panel conditioning and question type, we used two online household panels administered by CentERdata (see www.centerdata.nl and Appendix A in Chapter 1 for more details about the panels). The first, the CentERpanel, exists since 1991. Panel members fill out questionnaires every week. At the time of our survey, panel duration of respondents varied between seventeen years and a few months (the mean duration is 6 years and 8 months; the standard deviation is 4 years). The second panel is the new LISS-panel. Our questions were included in the first questionnaire presented to respondents in this panel. We fielded the questionnaire in June 2007. See Appendix A for the response numbers. To correct for differences due to unit non-response, we used weights based upon gender, age, and education (see Appendix A for the response distribution after weighting). We used nine questions on two different topics: food infection and old-age pensions[1]. These topics had already been asked to the trained panel several times (and not to the fresh panel, because this was their first questionnaire). The answers in the trained panel may therefore be affected by panel conditioning, either because they have already seen the same questions, or because their panel experience in general (not specifically the questions we discuss here) has affected their re-

[1] These questions were embedded in a questionnaire with three experiments on design issues. The questionnaires in both panels were exactly the same, both in content and in appearance. There was a difference in mean duration of the entire interview between panels (t=-2.4, p=.02): 436 seconds for the trained panel and 576 seconds for the fresh panel (where means were calculated after deleting outliers with more than twice the standard deviation (28 respondents in the fresh panel and 4 in the trained panel)).

sponse behavior. (Disentangling these two possibilities is beyond the purpose of the chapter; we only analyze whether panel conditioning occurs and if so, for which questions.)

7.4 Results

Table 7.1 presents the nine questions and the distribution of the answers in the two panels. All questions can be answered with yes or no only. The trained and fresh respondents answer the knowledge question about campylobacter[2] significantly different: 25.2% of the trained panelists know what campylobacter is compared to 13.9% of the fresh panelists. The question whether respondents know what salmonella is does not give significant differences between the two panels. The fact that salmonella is well-known (more than 98% of both panels say they know what it is) could explain why there is hardly any difference between the trained and the fresh panel. For cross-infection, the two panels significantly differ: 80.9% of the trained panelists know what cross-infection is compared to 76.4% of the fresh panelists. The difference between panels is not as large as the difference for the question about campylobacter, which is a less well-known concept. We also found differences in the question about "Stichting Pensioenkijker", an association to promote pension awareness of the Dutch population. Almost twice as many trained respondents compared to fresh respondents heard, saw, or read something about this association (39.7% of the trained panelists versus 22.0% of the fresh panelists).

The answers to the other types of questions (attitude, fact, and behavior) in Table 7.1 were not sensitive to repeated interviewing. Our results show that only knowledge questions are sensitive to panel conditioning. The difference between trained and fresh respondents gets larger the fewer respondents know the concept the question refers to.

To find out if the differences between the trained and fresh panel relate to respondent characteristics we conducted some probit analyzes. Table 7.2 and Table 7.3 show the results. Table 7.2 presents the estimation results for the questions with significantly different frequencies of 'Yes'-answers in the trained and fresh panel.

In the probit models the answer to each question is explained by a panel dummy (0 for the trained panel, 1 for the fresh panel), education, age, and gender of the respondent, and interaction terms of the panel dummy with these personal characteristics. The personal characteristics are included as deviations from their (overall) means, implying that the coefficient on the panel dummy can be interpreted as the panel conditioning effect for the average respondent. The results in Table 7.2 show that the panel conditioning effect remains signif-

[2] Campylobacter is a bacterium found in the intestines of many types of animals and is the most common bacterial cause of diarrheal illness.

Table 7.1: Comparison of Answers of Trained and Fresh Respondents to various Yes/No Questions

	Type of Question	%Yes Trained Panel	%Yes Fresh Panel
1. Do you know what Campylobacter is?	Knowledge	25.2	13.9*
2. Do you know what Salmonella is?	Knowledge	98.3	98.4
3. Do you know what Cross infection is?	Knowledge	80.9	76.4*
4. Did you think about your age of retirement the last year?	Behavior	60.5	59.1
5. Did you ever hear, see, or read something about "Stichting Pensioenkijker"?	Knowledge	39.7	22.0*
6. Do you think pensions will be higher about ten years from now?	Attitude	24.1	26.8
7. Do you think people will be more satisfied with their pensions about ten years from now?	Attitude	10.2	9.6
8. Do you think many people will retire partially in the future?	Attitude	64.0	62.8
9. Are you retired?	Fact	21.8	20.9

*Difference between trained and fresh panel is significant (p<.01).

icant if we correct for personal characteristics. We found no significant interaction terms, except for question 5 ("Did you ever hear, see, or read something about 'Stichting Pensioenkijker'?"). In particular, the panel conditioning effect declines with age for this question, suggesting that the younger people tend to seek more information about their pension as a result of having been interviewed. Since pension knowledge increases with age (cf. the positive age coefficient in Table 7.2), this is in line with the earlier finding that the panel conditioning effect in knowledge questions falls with the fraction of respondents who know the concept.

Table 7.3 presents the estimation results for the questions on which the trained and fresh panel showed no significant different frequencies of 'Yes'-answers. The panel conditioning effect for the average respondent remains in-

Table 7.2: Probit Estimation Results for (Knowledge) Questions with Significantly Different Frequencies in Trained and Fresh Panel

Question	1 Campylobacter	3 Cross Infection	5 StPensioenkijker
Panel	-.797**	-.271**	-.867**
Edu	.231**	.121**	.055
Age	.015	-.076	.129**
gender	.109	.184	-.295**
Panel*Edu	.045	.093	.001
Panel*Age	.104	.049	.165**
Panel*gender	-.072	-.013	.095
Constant	-1.089**	1.500**	-.382**

*p<0.05 **p<0.01

Note: Exact questions are defined in Table 7.1.
Panel is coded as 0=trained panel, 1=fresh panel.
Other explanatory variables are defined in Appendix A
and are included in the model as deviations from their (overall) means.

Table 7.3: Probit Estimation Results for Questions on which the Trained and Fresh Panel showed No Different Frequencies of 'Yes'-Answers

Question	2(K)	4(B)	6(A)	7(A)	8(A)	9(F)
Panel	-.445	-.004	.096	-.184	-.118	-.141
Edu	.652**	.426**	-.036	-.081	.124**	-.095
Age***	.403*	.298**	.113*	.115	.111*	6.781**
gender	1.502	-.081	-.568**	.021	-.006	-.039
Panel*Edu	-.375	-.186**	-.040	-.106	-.004	-.032
Panel*Age***	.190	.069	-.062	-.072	-.090	-.720
Panel*gender	-1.307	-.288	.082	-.230	-.008	-.607
Constant	5.034**	.564**	-1.207**	-2.202**	.648**	-2.647**

*p<0.05 **p<0.01

Note: Exact questions are defined in Table7.1.
K=Knowledge, B=Behavior, A=Attitude, and F=Fact
Panel is coded as 0=trained panel, 1=fresh panel.
Other explanatory variables are defined in Appendix A
and are included in the model as deviations from their (overall) means.
****For question 9 the variable age is replaced by a dummy variable instead of categorical variable; age=1 if the respondent is 65 years or older, 0 otherwise*

significant if we control for respondent characteristics. We did, however, find a significant effect of the interaction of panel experience with education level in question 4 ("Did you think about your age of retirement last year?"). Respondents with higher education tend to think more about their age of retirement than low educated respondents (keeping age and gender constant), but the difference is much larger in the experienced panel than in the fresh panel. This would suggest that an interview about pensions triggers respondents with higher education to think about their retirement age, but would have the opposite effect on the lower educated. This result does not seem plausible and deserves further investigation.

For question 9 ("Are you retired?") we changed the definition of the variable age due to the rather discontinuous relation between the fact whether the respondent is retired or not and age. Because in the Netherlands the benchmark age of retirement is 65, we replaced age by a dummy variable which equals 1 if the respondent is 65 years or older and zero otherwise. The estimation results in Table 7.3 show that the answer to this factual question is entirely explained by this dummy variable with no panel conditioning effect present.

We also conducted some probit estimations with panel duration (the number of weeks the respondent participates in the panel; zero for the respondents in the fresh panel) as an additional explanatory variable. Interaction terms between panel duration and personal characteristics were included as well. Neither the interaction term nor the duration variable itself contributed significantly to the model.

In short, the results show that the difference (or absence of the difference) between the trained and fresh panel is hardly associated with education, age, gender, and panel duration.

Concluding remarks 7.5

It is important to understand issues related to panel conditioning and their potential impact on the quality of research. Panel research gives big advantages, but the fact that the panel is the foundation on which research projects are built, and trained respondents may respond differently than fresh respondents, causes concerns with regard to survey quality. This chapter shows that knowledge questions, especially on less-known subjects, are very much affected by panel conditioning. When asking these kind of questions, a researcher has to be particular careful about the kind of sample used. We found that other types of questions are not sensitive to repeated interviewing. The results show that panel conditioning is not associated with education, age, gender, and panel duration.

7.6 Appendix A: Response rates

Table 7.4: Response Rates (Before and After Weighting)

	Pop. Distr.*	Trained Panel			Fresh Panel		
		Selection Panel Members	Response	Response after Weighting	Selection Panel Members	Response	Response after Weighting
Number of respondents		1369	981 (71.7%)		4149	2809 (67.7%)	
gender							
0.Male	49.5%	50.5%	55.1%	49.4%	46.2%	46.1%	49.5%
1. Female	50.5%	49.5%	44.9%	50.6%	53.8%	53.9%	50.5%
Age							
1. 15-24	13.3%	6.7%	5.8%	12.9%	12.0%	13.0%	14.9%
2. 25-34	15.6%	20.3%	13.9%	15.5%	16.4%	17.0%	16.4%
3. 35-44	19.7%	19.4%	20.6%	19.9%	22.0%	22.5%	19.9%
4. 45-54	18.0%	20.2%	21.2%	18.1%	21.2%	21.9%	17.5%
5. 55-64	15.4%	17.2%	19.5%	15.5%	17.5%	17.4%	15.6%
6. 65 and older	18.0%	16.2%	19.1%	18.0%	10.9%	8.2%	15.7%
Education							
1. Primary	9.5%	6.9%	5.2%	9.2%	11.2%	11.0%	9.5%
2. Lower Secondary	24.8%	26.7%	26.5%	24.8%	28.0%	27.4%	24.8%
3. Higher Secondary	10.8%	12.4%	12.2%	10.9%	9.5%	9.5%	10.8%
4. Inter-mediate Vocational	29.4%	20.6%	19.6%	29.5%	23.7%	24.4%	29.4%
5. Higher Vocational	16.3%	22.8%	25.3%	16.4%	20.3%	20.6%	16.3%
6. University	9.2%	10.6%	11.2%	9.2%	7.3%	7.0%	9.2%

*Population Distribution, Source=Statistics Netherlands

8 Conclusion

ALPHA ET OMEGA

While the importance of question wording in influencing respondents' answers is well-recognized, there is a growing literature that suggests that the design of the survey instrument (visual cues, sample characteristics, etc.) also plays an important role (Couper 2000). According to Dillman (2007) new ideas and research in the area of visual design and layout are changing the way that surveys must be done. Both the design of web and mixed-mode surveys are affected by new knowledge on the likelihood that different visual layouts for questions lead to different answers. The act of responding to a question (the question-answering process) contains four steps: interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, and reporting it. Mistakes can be made because of problems at any one of these steps, resulting in measurement error. Therefore, it is valuable to know which factors cause these 'mistakes'.

There are two reasons why design in web surveys may be more important compared to other modes. First, because a researcher has so many tools available (picture, colors, sound, navigation), there are many ways in which respondents' answers can be influenced. Second, because of differences in screen resolutions and browser settings, a researcher never knows exactly how a web questionnaire is going to appear.

Survey respondents use information to decide which answer they are going to report. They use information available in the questionnaire, question, and answer format. They also use information they got in prior surveys, as well as information they have obtained in ordinary life. All this information is used to fill out a questionnaire. To correctly interpret the answers and to optimally formulate the survey questions, it is important for researchers to know which information is used by respondents. Design choices may leave cues in a questionnaire. Respondents use these cues to decide which answer to select. The most important design aspects influencing the question-answering process are the sample, mode of administration, questionnaire characteristics, and respondents' personal characteristics.

The studies discussed in this dissertation address the following main research question:

Which design factors affect the answers provided by respondents in web surveys?

This main research question was analyzed by conducting six interrelated studies. The dissertation ends with a summary of the results of the different studies, the implications, and limitations and suggestions for future research.

8.1 Summary of the results

Chapter 1 described the question-answering process, together with factors influencing this process, and an overview of the dissertation. Chapter 2 analyzed whether the questionnaire type affected answers from respondents. At one end of the design continuum are form-based designs that present questionnaires as one long form in a scrollable window, at the other end are screen-by-screen questionnaires that present only a single item at a time. Presenting several questions per screen fits somewhere in the middle of the design continuum. In our study four different formats were used, varying the number of items and headers on a screen (1, 4, 10, and all items/headers (scrolling format)). Evaluation questions were added to find out how respondents experience the formats used. No effect of questionnaire format on measurement was found. In relation to item non-response, we found that the more items appear on a single screen, the higher the number of people with one or more missing values. We found evidence that placing more items on a single screen shortened the duration of the interview, but negatively influenced the respondent's evaluation of the layout. The results showed that grouping items on a screen affect people of different gender, age, and education groups. The effects were (mostly) of the same kind, but differed in degree.

Chapter 3 focused on the impact of answer type, e.g. closed versus open answers. In this chapter an information-processing perspective to explore the impact of response categories on the answers respondents provide in web surveys was used. Response categories had a significant effect on response formulation in questions that were difficult to process, whereas in easier questions (where responses are based on direct recall) the response scales had a smaller effect. How strongly the scale biases a respondent's answer, was also influenced by how the scale relates to the population distribution. If the distribution of categories was closer to the distribution of open-ended answers, the influence of response categories was less pronounced. In general, people with less cognitive sophistication were more affected by contextual cues. The Need for Cognition and the Need to Evaluate personality indexes for motivation accounted for a significant part of the variance in survey responding. We found significant interactions of ability to process information and motivation for questions that were more difficult to process. Our results hint at a substantial role of satisficing in web surveys; our results show larger differences between answer scales than experiments in other modes of administration (e.g. Schwarz et al. 1985; Rockwood et al. 1997).

Chapter 4 investigated if respondents gain meaning from visual and verbal cues in a web survey. We manipulated the layout of a five point rating scale in two experiments. First, we compared linear and non-linear formats. Second, we manipulated the linear layout using verbal, graphical, and numerical language. In a non-linear visualization, respondents were more eager to select the second answer on the top line. Our results supported a primacy effect in answering scalar questions. Options that require less movement of the mouse might be more easily chosen than answers requiring more hand/eye movements. Our experiments showed differences due to verbal, graphical, and numerical language. Elderly, and highly educated respondents were the most sensitive to layout.

Chapter 5 focused on the effect of panel as sample type on respondents' answers. Panel data have important advantages, but there are also two potential drawbacks: attrition bias and panel conditioning effects. Attrition bias can arise if respondents drop out of the panel non-randomly. Panel conditioning arises if responses in one wave are influenced by participation in the previous wave(s). In this chapter we discussed how to disentangle the total bias in panel surveys due to attrition and panel conditioning into a panel conditioning and an attrition effect. We developed a test for panel conditioning allowing for non-random attrition. The results showed a significant bias due to panel conditioning in knowledge questions, but not in other types of questions. In all cases the panel conditioning effect was positive, suggesting that people who have had a question once, often increase their knowledge about the phenomenon in the question before taking part in the next survey.

In Chapter 6 we investigated whether there are differences in design effects between trained and fresh respondents using a questionnaire consisting of three experiments. In the experiments we varied the number of items on a screen (based upon Chapter 2), the choice of response categories (based upon Chapter 3), and the layout of a five point rating scale (based upon Chapter 4). We found that trained respondents were more sensitive to satisficing and selected the first acceptable response option more often than fresh respondents. Fresh respondents showed stronger effects with regard to verbal and non-verbal cues than trained respondents. This suggests that fresh respondents find it more difficult to answer questions and pay more attention to details of the response scale in interpreting the question.

Chapter 7 investigated which type of question is sensitive to panel conditioning. The results in this chapter showed that panel conditioning only arises in knowledge questions. Answers to questions on attitudes, actual behavior, or facts were not sensitive to panel conditioning. The effect of panel conditioning in knowledge questions was larger for questions where fewer respondents knew the correct answer.

8.2 Implications

Survey respondents use information to decide which answer they are going to report. The aim of this study was to learn more about the consequences of design choices (as source of information) in web surveys in order to develop a better understanding of the factors influencing the question-answering process. We examined the effects of visual language in a web survey on data quality, a panel as type of sample (with its effects of re-interviewing), the questionnaire characteristics (interface, question, and answer type) and respondents' personal characteristics (demographics as well as personality traits). This dissertation aimed at gaining deeper insights into which factors matter and how they influence the quality of the survey data. Implications found for the four most important design aspects for web surveys influencing the question-answering process are discussed below.

The sample

We investigated the effect of using a panel as type of sample and its effect on data quality. The effect of attrition bias on respondents' answers was found to be small. In most question types we did not find evidence for panel conditioning either. Only in knowledge questions we found a significant bias due to panel conditioning. The conclusion that for most types of questions no evidence of panel conditioning was found seems reassuring. For questions concerning knowledge, panel conditioning seems an issue that researchers need to be aware of. Using a refreshment sample to control for panel conditioning (or attrition bias) is a way to avoid measurement error due to re-interviewing. The results showed that panel conditioning was not associated with education, age, gender, or panel duration. Therefore, the results are expected to apply to other samples, e.g. student samples, volunteer opt-in panels, and access-panels.

We found some evidence that experience relates to design effects in web surveys. Trained respondents seem to be more sensitive to satisficing; they are more likely to choose items earlier in a list because they find the first position that they can reasonably agree with and consider it a satisfactory answer, rather than processing each response option separately. One way to correct for this is to present answer categories in a random order. The advantage of using trained respondents is that they are less sensitive to visual language. Fresh respondents appeared more sensitive for visual cues in the response scales than trained respondents. Therefore, visual cues have to be chosen with care if a fresh sample is used. The number of items per screen and the impact of response categories did not show differences between trained and fresh respondents.

The mode of administration

It is investigated whether visual cues in web surveys influence respondents' answers. We found that respondents gain meaning from non-verbal cues in a web survey as well as from verbal cues (words). In a non-linear visualization, respondents were more eager to select the second answer on the top line. Therefore, scales should appear in linear layouts rather than multiple columns or rows of categories.

Our experiments showed differences if answer categories were reversed in a five point scale. This is in line with other results on visual languages. Literature shows that people expect more positive categories, or ones that express a greater degree of satisfaction or some other opinion to have higher labels (Tourangeau et al. 2004). In addition, people may expect categories to appear from positive to negative (Tourangeau et al. 2004). Our results indicated that a negative tone of the first option changed reports in a negative manner (anchoring effect, as suggested by Schwarz 1996), which should be taken into account when designing a questionnaire. Randomizing response categories is a way to control for this effect.

Changing the layout of a five point scale from vertical to horizontal led to some changes in the answer distributions, but the effects were not large. Adding the numbers 1 to 5 to a vertical format did not influence respondent answers. Adding the numbers 5 to 1 and 2 to -2 to the answer categories showed some significant differences. Although some manipulations did not show strong effects, tests comparing all manipulations simultaneously always showed significant differences between answer distributions and mean scores, indicating that visual cues influence respondents' answers. These cues should be kept to a minimum and be built in a web questionnaire with careful considerations for data quality.

Questionnaire characteristics

The use of a screen-by-screen design, scrollable format, or multiple-items-per screen format did not influence respondents' answers, but it did have an influence on respondents' evaluation of the questionnaire. Keeping the respondent satisfied is important e.g. for panel maintenance, and therefore it is important not to place too many items on a screen.

We found strong support for the hypothesis that the range of response categories affects respondent reports, especially in questions that are difficult to process. An open-ended format is preferable in questions in which estimation strategies have to be used. If this type of answer format is not desirable (e.g., because of higher item non-response on open-ended answers), categories in closed questions have to be chosen with care. How strongly the scale biases a respondent's answer, was influenced by how the scale relates to the population distribution. A pre-test can help determine the answer categories and decrease the impact of response categories.

Personal characteristics

The fact that we made use of heterogeneous samples, made it possible to test for effects of personal characteristics on answers provided by survey respondents. The effects we found were limited. Age and education effects were not as clear-cut as we would have expected. For example, there was no evidence that response effects fall monotonically with education level as suggested by literature.

The Need for Cognition and the Need to Evaluate as personality factors accounted for variance in survey responding. We found significant interactions of ability to process information and motivation for questions that were more difficult to process. Motivation helped when memory representation was bad; when memory representation was good motivation was not needed to report the behavior. When designing questionnaires one should pay particularly attention to design effects when respondents are expected to have difficulty in reporting, or lack of motivation.

This dissertation aimed at gaining deeper insights into which factors influence the quality of the survey data. Our results imply the following suggestions on web questionnaire design:

1. Present several items per screen, but not too many.

2. Use open-ended questions for questions that are difficult to process; if this is necessary use a pre-test to develop the answer scale.

3. Constrain visual cues to a minimum; use no numbers in addition to the verbal labels and present the labels in the same way (e.g. negative to positive) across surveys.

4. Use a linear layout for scalar questions, preferably horizontal.

5. If possible, randomize response categories to diminish the effect of satisficing.

6. Do not use a panel for knowledge questions, if this is inevitable correct for panel conditioning with fresh respondents.

Limitations and suggestions for future research 8.3

This section discusses limitations and provides suggestions for future research raised in this dissertation. This dissertation aimed at gaining deeper insights into which design factors affect the question-answering process in web surveys. The main components of the process are interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, and reporting it. Our studies cannot point out at which step in the process respondents' answers are affected. In that sense, the question-answering process still remains a black box.

The sample

Using a probability sample for our studies without the need for Internet access makes it possible to generalize conclusions to the population. But using a representative sample may also lead to more measurement error. Saris and Gallhofer (2007) suggest that lower educated and older people may make more reporting errors.

Another possible limitation is the fact that our use of a panel as type of sample might interact with the design results. We tried to analyze the effect of panel experience by repeating some of the studies in a fresh panel (with respondents filling out a questionnaire in this panel for the first time).

Panel conditioning seems to be related to cognitive demand. Knowledge questions where fewer people knew the answer were more sensitive to panel conditioning. More cognitively demanding questions might be more sensitive to panel conditioning. Future research can make the relation between panel conditioning and cognition more clear.

The mode of administration

Our studies show that design aspects influence the answers from respondents in web surveys. Unfortunately, we do not have a comparison condition that allows us to assess if the influence of design choices is more or less pronounced in web surveys. For example, Rockwood et al. (1997) did not find differences in a telephone mode compared to a mail mode (which are very different in information transmitting) with regard to answer category effects. Our results show larger differences between response scales than their results in telephone and mail mode. This could indicate a high tendency to satisfice in web surveys, as suggested by De Leeuw (2005). Because we have no comparison condition of another mode, this dissertation cannot assess if design choices analyzed here are more or less pronounced in web surveys.

With regard to the visual cues in answer categories, our results hint at the conjecture that presenting a five point scale horizontally makes that respondents read the answer categories more accurate, therefore decreasing the influence of layout. Further research in web surveys on a horizontal layout of scalar questions in different contexts can make this effect more clear.

Questionnaire characteristics

There are many ways to investigate the effect of questionnaire characteristics such as interface design, question type, and answer type on respondents' answers. Our studies show that these three factors interact with each other in influencing the response process. For example, our data point at a relation between question type and answer type. The influence of response categories is more pronounced in questions in which estimation strategies have to be used in formatting a report. Future research is needed if we want to develop a better understanding of the exact way in which these questionnaire characteristics interact with each other and influence the response process.

We found interaction effects between motivation and memory for questions that were more difficult to process. Motivation apparently helped when memory representation is bad; when memory representation was good motivation was not needed to report the requested behavior. This could be due to the fact that we used relatively simple questions. For more difficult questions the conjecture that motivation helps when memory representation is good might be valid. This is certainly worth additional investigation.

Personal characteristics

Although we did several analyzes with respondents' personal characteristics as explanatory variables, the effect on the quality of the data is not entirely clear. This is in line with the literature, which also shows mixed findings. Future research on different types of questions may make the effect of personal characteristics more clear. It can be possible that personal characteristics affect respondents' answers in some type of questions, but not in others. To know which types of questions are affected by personal characteristics is a valuable asset to the theory and practice of (web) questionnaire design. In addition, other personality scales than the Need for Cognition and Need to Evaluate might be important factors influencing respondents' answers.

We showed that the highly educated seemed to be more sensitive for layout effects. Deriving conclusions on a university students based sample might show more differences between different formats than a heterogeneous sample of the population. Future research should be conducted comparing students based and representative samples to find out if studies using students show more or less significant results with regard to (web) questionnaire design.

Each factor influencing the question-answering process has far more elements than discussed in this dissertation. For example, the effect of mode of administration is limited to visual cues in web surveys, although the effect of the presence or absence of an interviewer and the use of computer-assisted surveys or not can also be important determinants of the question-answering process. In addition, differences between probability-based sampling and non-probability-based sampling, probability-based sampling of the full population and Internet users etc. is beyond the scope of this dissertation. The influence of questionnaire, question, and answer types can be analyzed in different ways as well as the influence of personal characteristics. So the conclusions in this dissertation are mere a beginning of understanding the effects of factors influencing the question-answering process. Our studies can be used to contribute towards a theory and practice of web questionnaire design. The understanding of the quality of respondents' answers depends on it.

Bibliography

Baltagi, B. (2001): *Econometric Analysis of Panel Data.* Wiley: Chischester.

Bartels, L. (1999): Panel Effects in the American National Election Studies. *Political Analysis*, 8:1–20.

Biemer, P. and L. E. Lyberg (2003): *Introduction to Survey Quality.* Wiley series in Survey Methodology: New Jersey.

Bizer, G. Y., J. A. Krosnick, A. L. Holbrook, S. C. Wheeler, D. D. Rucker, and R. E. Petty (2004): The Impact of Lifeity on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate. *Journal of Lifeity*, 72:996–1028.

Blumberg, H., C. Fuller, and P. Hare (1974): Response Rates in Postal Surveys. *Public Opinion Quarterly*, 38:113–123.

Bodenhausen, G. and R. Wijer (1987): Social Cognition and Social Reality: Information Acquisition and Use in the Labaratory and the Real World. In H. Hippler, N. Schwarz, and S. Sudman, editors, *Social Information Processing and Survey Methodology*, pages 6–41. Springer-Verlag: New York.

Borgers, N., J. Hox, and D. Sikkel (2004): Response Effects in Surveys on Children and Adolecents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point. *Quality and Quantity*, 38:17–33.

Bowker, D. and D. A. Dillman (2000): An Experimental Evaluation of Left and Right Oriented Screens for Web Questionnaires. *Paper presented at the 55th annual conference of American Association for Public Opinion Research. Portland, Oregon, May 18-21, 2000. http://survey.sesrc.wsu.edu/dillman/papers/AAPORpaper00.pdf.*

Bradlow, E. T. and G. J. Fitzsimons (2001): Subscale Distance and Item Clustering Effects in Self-administered Surveys: A New Metric. *Journal of Marketing Research*, 38:254–262.

Brannen, J. (1993): The Effects of Research on Participants: Findings from a Study of Mothers and Employment. *The Sociological Review*, 41:328–346.

Cacioppo, J. and R. Petty (1982): The Need for Cognition. *Journal of Personality and Social Psychology*, 42:116–131.

Chan, J. and K. McDermott (2007): The Testing Effect in Recognition Memory: A Dual Process Account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33:431–437.

Chang, L. and J. Krosnick (2003): Comparing Oral Interviewing with Self-Administered Computerized Questionnaires: An Experiment. *http://communication.stanford.edu/faculty/krosnick.html.*

Choquette, K. A. and M. N. Hesselbrock (1987): Effects of Retesting with the Beck and Zung Depression Scales in Alcoholics. *Alcohol and Alcoholism*, 22:277–283.

Christian, L. (2003): The Influence of Visual Layout on Scalar Questions in Web Surveys. *Unpublished Master's Thesis on http://survey.sesrc.wsu.edu/dillman/papers.htm.*

Christian, L. M. and D. A. Dillman (2004): The Influence of Graphical and Symbolic Language Manipulations to Self-Administered Questions. *Public Opinion Quarterly*, 68:57–80.

Christian, L. M., D. A. Dillman, and J. D. Smyth (2005): Instructing Web and Telephone Respondents to Report Date Answers in Format Desired by the Surveyor. *Social and Economic Sciences Research Center (SESRC), Technical Report 05-067.*

——— (2007): Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys. *Public Opinion Quarterly*, 71:113–125.

Clinton, J. (2001): Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel. *Working paper www.knowledgenetworks.com.*

Coen, T., J. Lorch, and L. Piekarski (2005): The Effects of Survey Frequency on Panelists' Responses. *ESOMAR, www.websm.org.*

Coombs, L. (1973): Problems of Contamination in Panel Surveys: A Brief Report on an Independent Sample, Taiwan, 1970. *Studies in Family Planning*, 4:257–261.

Couper, M. (2000): Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly*, 64:464–494.

Couper, M. P., R. Tourangeau, and K. Kenyon (2004): Picture This. Exploring Visual Effects in Web Surveys. *Social Science Computer Review*, 68:256–266.

Couper, M. P., M. W. Traugott, and M. J. Lamias (2000): Web Survey Design and Administration. *Public Opinion Quarterly*, 60:230–253.

Crawford, S., M. Couper, and M. Lamias (2001): Web Surveys: Perceptions of Burden. *Social Science Computer Review*, 19:146–162.

Das, M., V. Toepoel, and A. van Soest (2007): Can I use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys. *CentER Discussion Paper, Tilburg University*, 56.

De Leeuw, E. (2005): To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21:233–255.

Dennis, M. (2001): Are Internet Panels Creating Professional Respondents? *Marketing Research*, 13:34–39.

Deutskens, E. (2006): From Paper-and-Pencil to Screen-and-Keyboard: Studies on the Effectiveness of Internet-Based Marketing Research. *PhD dissertation, University of Maastricht.*

Deutskens, E., K. de Ruyter, M. Wetzels, and P. Oosterveld (2004): Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study. *Marketing Letters*, 15:21–36.

Dillman, D. A. (2007): *Mail and Internet Surveys. The Tailored Design Method.* Wiley: Hoboken NJ.

Dillman, D. A., S. Caldwell, and M. Gansemer (2000): Visual Design Effects on Item Nonresponse to a Question About Work Satisfaction That Precedes the Q-12 Agree-Disagree Items. *Paper supported by the Gallup Organization and Washington State University http://survey.sesrc.wsu.edu/dillman/papers.htm.*

Dillman, D. A. and L. M. Christian (2002): The Influence of Words, Symbols, Numbers, and Graphics on Answers to Self-Administered Questionnaires: Results from 18 Experimental Comparisons. *http://survey.sesrc.wsu.edu/dillman/papers.htm.*

——— (2005): Survey Mode as a Source of Instability in Responses across Surveys. *Fieldmethods,* 17:30–52.

Dillman, D. A., J. D. Smyth, L. M. Christian, and M. J. Stern (2006): Multiple Answer Questions in Self-Administered Surveys: The Use of Check-All-That-Apply and Forced-Choice Question Formats. *Public Opinion Quarterly,* 70:66–77.

Dillman, D. A., R. Tortora, and D. Bowker (1998): Principles for Constructing Web Surveys. *SESRC Technical Report 98-50, Pullman, Washington. http://survey.sesrc.wsu.edu/dillman/papers.htm.*

Duan, D., M. Alegria, G. Canino, T. McGuire, and D. Takeuchi (2007): Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats. *Health Services Research,* 42:890–907.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources,* 33:251–299.

Friedman, L. W. and H. H. Friedman (1994): A Comparison of Vertical and Horizontal Rating Scales. *The Mid-Atlantic Journal of Business,* 30:102–107.

Friedman, L. W. and J. R. Leefer (1994): Label Versus Position in Rating Scales. *Journal of the Academy of Marketing science,* 9:88–92.

Fuchs, M. (2005): Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options. *Journal of Official Statistics,* 21:701–725.

Golob, T. (1990): The Dynamics of Household Travel Time Expenditures and Car Ownership Decisions. *Transportation Research,* 24A:443–463.

Griffith, L., D. Cook, G. Guyatt, and C. Charles (1999): Comparison of Open and Closed Questionnaire Formats in Obtaining Demographic Information From Canadian General Internists. *Journal of Clinical Epidemiology,* 52:997–1005.

Hausman, J. and D. Wise (1979): Attrition bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. *Econometrica,* 47:455–474.

Hirano, K., G. Imbens, G. Ridder, and D. Rubin (2001): Combining Panel Data Sets with Attrition and Refreshment Samples. *Econometrica,* 69:1645–1659.

Hofmans, J., P. Theuns, S. Baekelandt, O. Mairesse, N. Schillewaert, and W. Cools (2007): Bias and Changes in Perceived Intensity of Verbal Qualifiers Effected by Scale Orientation. *Survey Research Methods*, 1:97–108.

Hurd, M., D. McFadden, H. Chand, L. Gan, A. Merrill, and M. Roberts (1998): Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias. In D. Wise, editor, *Frontiers in the Economics of Aging*, pages 353–387. University of Chicago Press: Chicago.

Jarvis, W. and R. Petty (1996): The Need to Evaluate. *Journal of Personality and Social Psychology*, 70:172–194.

Jenkins, C. R. and D. A. Dillman (1997): Towards a Theory of Self-Administered Questionnaire Design. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, editors, *Survey Measurement and Process Quality*, pages 165–169. Wiley: New York.

Kalton, G., D. Kasprzyk, and D. McMillen (1989): Nonsampling Errors in Panel Surveys. In D. Kasprzyk, G. Duncan, G. Kalton, and M. Singh, editors, *Panel Surveys*, pages 249–270. Wiley: New York.

Knauper, B., N. Schwarz, and D. Park (2004): Frequency Reports Across Age Groups. *Journal of Official Statistics*, 20:91–96.

Krosnick, J. and F. Alwin (1987): An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51:201–219.

Krosnick, J., S. Narayan, and W. Smith (1996): Satisficing in Surveys: Initial Evidence. *New Directions for Program Evaluation*, 70:29–44.

Krosnick, J. A. and L. R. Fabrigar (1997): Designing Rating Scales for Effective Measurement in Surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, editors, *Survey Measurement and Process Quality*, pages 141–164. Wiley: New York.

Kwak, N. and B. Radler (2002): A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality. *Journal of Official Statistics*, 18:257–273.

Lee, M. (2002): *Panel Data Econometrics*. Academic Press: Boston, 1st edition.

Little, R. and D. Rubin (2002): *Statistical Analysis with Missing Data*. Wiley: New York.

Lohse, G., S. Bellman, and E. Johnson (2000): Consumer Buying Behavior on the Internet: Findings from Panel Data. *Journal of Interactive Marketing*, 14:15–29.

Lozar Manfreda, K., Z. Batagelj, and V. Vehovar (2002a): Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer-Mediated Communication*, 7, 3:http://www.ascusc.org/jcmc/vol7/issue3/vehovar.html.

——— (2002b): Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer-Mediated Communication*, 7,3:http://jcmc.indiana.edu/vol7/issue3/vehovar.html.

Lynch, J., D. Chakravarti, and M. Anusree (1991): Contrast Effects in Consumer Judgments: Changes in Mental Representations or in the Anchoring of Rating Scales? *Journal of Consumer Research*, 18:284–297.

Lynn, P. (1991): Data Collection Mode Effects on Responses to Attitudinal Questions. *Journal of Official Statistics*, 14:1–14.

Manski, C. (1989): Anatomy of the Selection Problem. *Journal of Human Resources*, 24:343–360.

——— (1995): *Identification Problems in the Social Sciences*. Harvard University Press: Cambridge MA.

Mathiowetz, N. and T. Lair (1994): Getting better? Changes or Errors in the Measurement of Functional Limitations. *Journal of Economic & Social Measurement*, 20:237–262.

McFarland, S. (2001): Effects of Question Order on Survey Responses. *Public Opinion Quarterly*, 45:208–215.

Mehrabian, A. and J. Russell (1974): *An Approach to Environmental Psychology*. MIT Press: Cambridge.

Menon, G., P. Raghubir, and N. Schwarz (1995): Behavioral Frequency Judgments: An Accessibility-Diagnosticity Framework. *Journal of Consumer Research*, 22:212–228.

Meurs, H., L. Van Wissen, and J. Visser (1989): Measurement Biases in Panel Data. *Transportation*, 16:175–194.

Nicoletti, C. (2006): Nonresponse in Dynamic Panel Data Models. *Journal of Econometrics*, 132:461–489.

Norman, K. L., Z. Friedman, K. Norman, and R. Stevenson (2001): Navigational Issues in the Design of On-Line Self-Administered Questionnaires. *Behavior and Information Technology*, 20:37–45.

Pennell, S. and J. Lepkowski (1992): *Panel Conditioning Effects in the Survey of Income and Program Participation. Proceedings of the Survey Research Methods Section.* American Statistical Association: Washington, DC.

Petty, R. and W. Jarvis (1996): An Individual Differences Perspective on Assessing Cognitive Processes. In N. Schwarz and N. Sudman, editors, *Answering Questions*, pages 221–257. Jossey-Bass Publishers: San Francisco.

Peytchev, A., M. P. Couper, S. E. McCabe, and S. D. Crawford. (2006): Web Survey Design. Paging Versus Scrolling. *Public Opinion Quarterly*, 70:596–607.

Redline, C., D. Dillman, L. Carley-Baxter, and R. Creecy (2003): Factors that Influence Reading and Comprehension in Self-Administered Questionnaires. *Paper presented at the workshop on Item-Nonresponse and Data Quality, Basel, Switzerland, October 10 2003. http://survey.sesrc.wsu.edu/dillman/papers.htm.*

Rockwood, T., R. Sangster, and D. Dillman (1997): The Effect of Response Categories on Questionnaire Answers: Context and Mode Effects. *Sociological Methods and Research*, 26:118–140.

Rubin, D. (1976): Inference and Missing Data. *Biometrika*, 63:581–592.

Sanchez, M. (1992): Effects of Questionnaire Design on the Quality of Survey Data. *Public Opinion Quarterly*, 56:206–217.

Saris, W. and I. Gallhofer (2007): Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions. *Survey Research Methods*, 1:29–43.

Schonlau, M., R. D. Fricker, and M. N. Elliott. (2002): *Conducting Research Surveys via E-mail and the Web.* RAND, Santa Monica.

Schuman, H. and S. Presser (1981): *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording and Content.* Quantitative Studies in Social Relations: New York.

Schwarz, N. (1996): *Cognition and Communication. Judgmental Biases, Research Methods, and the Logic of Conversation.* Lawrence Erlbaum Associates Publishers: New Jersey.

Schwarz, N., C. E. Grayson, and B. Knauper (1998): Formal Features of Rating Scales and the Interpretation of Question Meaning. *International Journal of Public Opinion Research*, 10:177–183.

Schwarz, N. and H. Hippler (1987): What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives. In H. Hippler, N. Schwarz, and S. Sudman, editors, *Social Information Processing and Survey Methodology*, pages 163–178. Springer-Verlag: New York.

Schwarz, N., H. Hippler, B. Deutsch, and F. Strack (1985): Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49:388–395.

Schwarz, N., B. Knauper, H.-J. Hippler, E. Noelle-Neumann, and L. Clark (1991): Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, 55:570–582.

Sharot, T. (1991): Attrition and Rotation in Panel Surveys. *The Statistician*, 40:325–331.

Sharpe, J. and D. Gilbert (1998): Effects of Repeated Administration of the Beck Depression Inventory and Other Measures of Negative Mood States. *Personal Individual Differences*, 24:457–463.

Smith, T. W. (1995): Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses. *Proceedings of the American Statistical Association*, Survey Research Section:1046–1051.

Smyth, J., D. A. Dillman, L. M. Christian, and M. McBride (forthcoming): Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? *Public Opinion Quarterly*.

Smyth, J., D. A. Dillman, L. M. Christian, and M. J. Stern (2006): Effects of Using Visual Design Principles to Group Response Options in Web Surveys. *International Journal of Internet Science*, 1:6–16.

Stern, M. J., D. A. Dillman, and J. D. Smyth (2007): Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey. *Survey Research Methods*, 1:121–138.

Strack, F. and L. Martin (1987): Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H. Hippler, N. Schwarz, and S. Sudman, editors, *Social Information Processing and Survey Methodology*, pages 123–148. Springer-Verlag: New York.

Strack, F., N. Schwarz, and M. Wanke (1991): Semantic and Pragmatic Aspects of Context Effects in Social and Psychological Research. *Social Cognition*, 9:111–125.

Strahan, R. and K. C. Gerbrasi (1972): Short, Homogeneous Versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 28:191–193.

Strube, G. (1987): Answering Survey Questions: The Role of Memory. In H. Hippler, N. Schwarz, and S. Sudman, editors, *Social Information Processing and Survey Methodology*, pages 86–101. Springer-Verlag: New York.

Sturgis, P., N. Allum, and I. Brunton-Smith (2007): Attitudes Over Time: The Psychology of Panel Conditioning. In P. Lynn, editor, *Methodology in Longitudinal Surveys*, pages 1–13. Wiley: Chischester.

Sudman, S., N. Bradburn, and N. Schwarz (1996): *Thinking about Answers*. Jossey-Bass Publishers: San Francisco.

Thomas, R. and D. Klein (2006): Merely Incidental?: Effects of Response Format on Self-reported Behavior. *Journal of Official Statistics*, 22:221–244.

Toepoel, V., M. Das, and A. van Soest (2006): Design of Web Questionnaires: the Effect of Layout in Rating Scales. *CentER Discussion Paper 2006-30, Tilburg University*.

——— (2009a): Design of Web Questionnaires: The Effects of the Number of Items per Screen. *Field Methods*, 21 (2).

Toepoel, V., C. Vis, M. Das, and A. van Soest (2009b): Design of Web Questionnaires: an Information-Processing Perspective for the Effect of Response Categories. *Sociological Methods and Research, Special Issue on Web Surveys*.

Tormala, Z. L. and R. E. Petty (2001): On-Line Versus Memory-Based Processing: The Role of "Need to Evaluate" in Person Perception. *Pers Soc Psychol Bull*, 27:1599–1612.

Tourangeau, R., M. P. Couper, and F. Conrad. (2004): Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68:368–393.

——— (2007): Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, 71:91–112.

Tourangeau, R., L. Rips, and K. Ransinki (2000): *The Psychology of Survey Response*. University Press: cambridge.

Trivellato, U. (1999): Issues in the Design and Analysis of Panel Studies: A Cursory Review. *Quality & Quantity*, 33:339–352.

Van der Vaart, W. and T. Glasner (1999): Applying a Timeline as a Recall Aid in a Telephone Survey: A Record Check Study. *Applied Cognitive Psychology*, 21:227–238.

Van der Zouwen, J. and T. Van Tilburg (2001): Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses. *Sociological Methods & Research*, 30:35–56.

Vella, F. (1998): Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources*, 33:127–169.

Voogt, R. and W. Saris (2005): Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, 21:367–387.

Wang, K., D. Cantor, and A. Safir (2000): Panel Conditioning in a Random Digit Dial Survey. In *Proceedings of the Section on Survey Research Methods*, pages 822–827. American Statistical Association: Alexandria VA.

Waterton, J. and D. Lievesley (1989): Evidence of Conditioning Effects in the British Social Attitudes Panel. In D. Kasprzyk, G. Duncan, G. Kalton, and M. Singh, editors, *Panel Surveys*, pages 319–339. Wiley: New York.

Weng, L. and C. Cheng (2000): Effects of Response Order on Likert-type Scales. *Educational and Psychological Measurement*, 60:908–924.

Williams, W. (1970): The Systematic Bias Effects of Incomplete Responses in Rotation Samples. *Public Opinion Quarterly*, 33:593–602.

Williams, W. and C. Mallows (1970): Systematic Biases in Panel Surveys Due to Differential Nonresponse. *Journal of the American Statistical Association*, 65:1338–1349.

Winter, J. (2002a): Bracketing Effects in Categorized Survey Questions and the Measurement of Economic Quantities. *Discussion Paper No. 02-35, Sonderforschungsbereich 504, University of Mannheim.*

——— (2002b): Effects in Survey-Based Measures of Household Consumption. *Discussion Paper No. 02-34, Sonderforschungsbereich 504, University of Mannheim.*

Nederlandse samenvatting

Iedereen gebruikt informatie bij zijn dagelijkse bezigheden. Informatie om te bepalen welk product men gaat kopen, in de dagelijkse gesprekken etc. Ook respondenten gebruiken informatie bij het invullen van een vragenlijst. Ze gebruiken informatie uit de vragenlijst (bijvoorbeeld uit de vraag of uit de antwoordcategorieën), informatie uit voorgaande vragenlijsten die ze hebben ingevuld, en informatie die ze hebben verkregen in hun dagelijkse leven. Al deze informatie wordt gebruikt bij het kiezen van een antwoord op een vraag uit een vragenlijst. Bij het ontwerpen van een vragenlijst maakt een onderzoeker allerlei keuzes die de informatie die respondenten tot hun beschikking hebben beïnvloedt. Respondenten gebruiken deze informatie als signalen en daardoor kunnen keuzes met betrekking tot het ontwerp van een vragenlijst de antwoorden van respondenten beïnvloeden. Om vragen optimaal te formuleren en om de antwoorden van respondenten correct te interpreteren, is het voor onderzoekers belangrijk om te weten welke informatie respondenten gebruiken bij het invullen van een vragenlijst.

Het doel van dit onderzoek was na te gaan welke keuzes met betrekking tot het ontwerpen van een Internetvragenlijst van invloed zijn op de antwoorden die respondenten geven. Allereerst moet een onderzoeker kiezen wat voor soort steekproef hij gaat nemen, met andere woorden wie zijn vragenlijst gaat invullen. De keuze van de steekproef kan de data beïnvloeden. Zo kunnen getrainde respondenten de vragenlijst anders invullen dan ongetrainde respondenten; zij hebben meer inhoudelijke en procedurele kennis. Ook kan de wijze van bevraging van invloed zijn op de data: communiceren via telefoon, papier, of door een persoonlijk interview kan andere antwoorden opleveren bij dezelfde persoon dan communiceren via de computer. Verder kunnen keuzes in een vragenlijst (zoals het soort vraag en het soort antwoordcategorieën) informatie bevatten die respondenten gebruiken bij het kiezen van een antwoord. Ook kunnen de kenmerken van respondenten zelf van invloed zijn op de data. Mannen, vrouwen, ouderen, jongeren, mensen met een hoge opleiding of een lage enzovoorts kunnen vragen op een andere manier interpreteren. Respondenten kunnen ook verschillen in de mate waarin zij nadenken of in de behoefte om hun mening te uiten. In dit proefschrift gebruiken wij vier factoren die het antwoordproces kunnen beïnvloeden: het type steekproef, de wijze van bevraging, kenmerken van de vragenlijst, en kenmerken van de respondenten zelf.

Met betrekking tot het type steekproef is gekeken naar herhaalde bevraging van respondenten (in de context van een panel). Door herhaaldelijk dezelfde respondenten te vragen een vragenlijst in te vullen krijg je een schat aan informatie van individuen, je kan ze door de tijd heen volgen, en het is goedkoper dan steeds nieuwe respondenten te zoeken. Maar aan de andere kant heeft een panel dat herhaaldelijk wordt bevraagd ook nadelen: de uitval van respondenten kan selectief zijn (waardoor de mensen in het panel niet dezelfde kenmerken hebben als de mensen die het panel verlaten) en de respondenten kunnen leren van het invullen van vragenlijsten (waardoor hun antwoorden vertekend kunnen zijn). In dit onderzoek bleek selectieve uitval in panelonderzoek minimaal. Een 'leereffect', waarbij respondenten inhoudelijk leren van het invullen van een vragenlijst, was alleen aanwezig in kennisvragen. Hoe minder mensen het (juiste) antwoord weten op een kennisvraag, hoe groter het 'leereffect' van panelleden. Voor kennisvragen kan een frisse cross sectie (nieuwe steekproef), of bij een bestaand panel een 'nieuwe' aanwas van respondenten, gebruikt worden om meetfouten door herhaaldelijke bevraging te vermijden. We vonden dat getrainde respondenten vaker de eerste de beste antwoordmogelijkheid die in de richting komt van hun eigen mening selecteren, waardoor antwoorden die in het begin van een lijst staan vaker worden gekozen. Een manier om hiervoor te corrigeren is het verschillend (random) voorleggen van antwoordmogelijkheden. Een voordeel van getrainde respondenten is dat zij minder gevoelig zijn voor visuele informatie in een vragenlijst. We vonden dat visuele elementen in een Internetvragenlijst zoals de presentatie van antwoordmogelijkheden op het computerscherm, het toevoegen van cijfers aan verbale antwoordmogelijkheden en dergelijke juist nieuwe respondenten beïnvloedt. Zij vinden het wellicht moeilijker om vragen te beantwoorden en laten zich daardoor eerder beïnvloeden door signalen in een vragenlijst.

In dit proefschrift staan Internetvragenlijsten als wijze van bevraging centraal. Internetvragenlijsten staan bekend om hun visuele eigenschappen. We zijn nagegaan of deze visuele eigenschappen de antwoorden van respondenten beïnvloeden. Wanneer antwoordschalen over meerdere rijen/kolommen verdeeld worden, zijn respondenten eerder geneigd om het tweede antwoord van de eerste regel te kiezen (in vergelijking met één rij/kolom antwoorden). Het is daarom beter om antwoordmogelijkheden in één rij/kolom te presenteren. Dit onderzoek laat ook zien dat respondenten geneigd zijn het eerste antwoord te kiezen dat hun mening in redelijke wijze weergeeft. Antwoordmogelijkheden die meer beweging (van ogen of handen) vergen lijken minder vaak gekozen te worden. Verder laat het onderzoek zien dat als antwoordmogelijkheden omgedraaid worden (bijvoorbeeld van eens-oneens naar oneens-eens) dit de data vertekent; een negatieve toon van het eerste antwoord zorgt voor meer negatieve antwoorden. Een manier om hiervoor te corrigeren is het willekeurig voorleggen van antwoordmogelijkheden. Grafische wijzigingen, het toevoegen van cijfers en symbolen worden gezien als visuele taal in aanvulling op de verbale taal in een vragenlijst. Om meetfouten te voorkomen is het beter om deze visuele taal tot een minimum te beperken en hier uiterst zorgvuldig in een vragenlijst mee om te gaan om de kwaliteit van de data (en vergelijkingen tussen onderzoeken) te waarborgen.

Kenmerken van Internetvragenlijsten kunnen ingedeeld worden in de manier waarop een vragenlijst is opgezet, het soort vragen en het soort antwoordmogelijkheden. In Internet vragenlijsten kunnen vragen scherm-per-scherm gepresenteerd worden, kunnen er enkele vragen per scherm staan, of kan alles op één enkel scherm staan zodat de respondent moet scrollen om de vragen te beantwoorden. Het plaatsen van meerdere vragen op het computerscherm heeft geen invloed op de resultaten. Wel beoordelen respondenten de lay-out van het scherm slechter. Om respondenten tevreden te houden is het daarom aan te bevelen om niet te veel vragen op een scherm te zetten. De keuze van antwoordschalen heeft grote invloed op de antwoorden die respondenten geven in vragen waarbij de respondent het antwoord moeilijk voor de geest kan halen. Zo gaf (in Hoofdstuk 3) 22% van de respondenten aan meer dan 2 en een half uur per dag televisie te kijken toen ze een schaal met lage antwoordcategorieën voorgelegd kregen, maar was dit 54% toen ze een schaal met hoge antwoordmogelijkheden voorgelegd kregen. Respondenten gebruiken de informatie in de schaal om in te schatten hoe vaak zij iets doen en denken dat het midden van de schaal normaal (gemiddeld) gedrag weergeeft. In makkelijke vragen, waarbij respondenten zich het antwoord goed kunnen herinneren, is de invloed van schalen veel kleiner. Voor moeilijke vragen kunnen beter open antwoorden gebruikt worden. Als dit niet gewenst is (respondenten vinden open antwoorden soms moeilijker en vergeten vaker een antwoord te geven) kan een vooronderzoek de antwoordschalen helpen bepalen. Antwoorden worden namelijk minder snel beïnvloed door antwoordschalen als de schalen dichtbij de verdeling in de populatie liggen.

Kenmerken van de respondenten zelf zijn de vierde factor die het antwoordproces kan beïnvloeden. Omdat dit proefschrift gebruikt maakt van heterogene (representatieve) steekproeven, kan het effect van persoonlijke kenmerken bekeken worden. Dit effect was niet zo duidelijk als van tevoren verwacht. Onze data lieten bijvoorbeeld niet duidelijk zien dat de kwaliteit van data samenhangt met de hoogte van de opleiding (de capaciteit van het werkend geheugen), of de leeftijd van respondenten (ouderen produceren geen slechtere data vanwege een verminderd geheugen), iets dat in de literatuur wel gesuggereerd wordt. Persoonlijkheidsfactoren zoals de behoefte om na te denken of een mening te hebben, dragen bij aan de variatie in antwoordgedrag. Over het algemeen worden de antwoorden van respondenten met een lage behoefte om na te denken en een lage behoefte om een mening te hebben eerder beïnvloed door het ontwerp van een vragenlijst, bijvoorbeeld de keuze van een antwoordschaal.

Dit proefschrift resulteert in de volgende aanbevelingen voor het ontwerpen van Internetvragenlijsten:

1. Gebruik enkele vragen per scherm, maar niet te veel (zodat de respondent niet hoeft te scrollen).

2. Gebruik geen antwoordcategorieën maar open antwoordmogelijkheden bij moeilijke vragen. Als de voorkeur toch uitgaat naar antwoordcategorieën, gebruik dan een vooronderzoek om de antwoordschalen te bepalen.

3. Beperk visuele taal tot een minimum; gebruik geen nummers als aanvulling op verbale labels bij antwoordcategorieën en presenteer de antwoorden steeds op dezelfde manier (om vergelijking mogelijk te maken).

4. Gebruik één rij/kolom voor de antwoordmogelijkheden.

5. Randomiseer antwoordmogelijkheden om te voorkomen dat antwoorden in het begin van een lijst eerder worden gekozen.

6. Gebruik geen panel (met respondenten die herhaaldelijk bevraagd worden) voor kennisvragen. Als dit toch noodzakelijk is, controleer dan voor eventuele leereffecten met behulp van een nieuwe aanwas van respondenten.