Open-ended questions in Web surveys

–

Using visual and adaptive questionnaire design to improve narrative responses

INAUGURALDISSERTATION

zur Erlangung des Grades eines Doktors rerum politicarum (Dr. rer. pol.)
im Fachbereich Gesellschafts- und Geschichtswissenschaften
an der Technischen Universität Darmstadt

Referenten:

Prof. Dr. Marek Fuchs, Technische Universität Darmstadt
Prof. Dr. Michael Bošnjak, GESIS - Leibniz-Institut für Sozialwissenschaften

Genehmigte Dissertation von:

Diplom-Sozialwissenschaftler Matthias Emde
aus Bad Wildungen

Tag der Einreichung: 02.05.2014
Tag der mündlichen Prüfung: 14.10.2014

Darmstadt, 2014

D17

ABSTRACT

One of the most significant decisions when designing survey questions is whether the questions will be posed as closed-ended or open-ended. Closed-ended questions require respondents to choose from a set of provided response-options, while open-ended questions are answered by respondents in their own words. Open-ended questions offer the benefit of not constraining responses and allowing respondents to freely answer and elaborate upon their responses. Narrative open-ended questions are especially useful when there are no suitable answer categories available for a closed-ended question format, or if providing response options might bias the respondents. Open-ended questions are also powerful tools for collecting more detailed and specific responses from large samples of respondents. However, open-ended questions are burdensome to answer and suffer from higher rates of item-nonresponse.

This thesis aims to improve narrative open-ended questions in Web surveys by using visual and adaptive questionnaire design. Previous research on open-ended questions demonstrated that respondents react to the size and design of the answer box offered with an open-ended question in Web surveys. Larger answer boxes seem to pose an additional burden as compared to smaller answer boxes. At the same time larger answer boxes work as a stimulus that increases the length of the response provided by those respondents who actually answer the question. By varying the visual design of answer-boxes this thesis seeks ways to improve narrative open-ended questions.

Despite the influence of different answer-box sizes, the effectiveness of a counter associated with the answer box that continuously indicates the number of characters left to type is tested. In addition dynamic in size growing answer-boxes were compared to answer-boxes that where adjusted in size by respondents themselves. Despite varying the visual appearance of narrative open-ended questions and the answer-boxes used, the interactive nature of the internet allows a multiplicity of ways to integrate interactive features into a survey. It is possible to adapt Web surveys individually to groups of respondents. Based on previous answers

it is feasible to provide specific designed questions to engage respondents. This thesis puts two adaptive design approaches to improve narrative open-ended questions to a test. The amount of information respondents typed into the response box of an initial open-ended question was used to assign them later in the survey to a custom-size answer box. In addition a follow-up probe was tested were respondents who didn't respond to a narrative open-ended question were assigned to the same question in a closed format to get at least some information from them.

Overall a well-designed and thoughtfully written question is still the best way to obtain high quality data. But the adaptive and visual designs do enrich the survey experience for respondents and at the same time improve data quality of narrative open-ended questions in Web surveys. Therefore, results show that it's worth paying attention to the visual design of open-ended questions.

ACKNOWLEDGEMENTS

First and foremost I want to thank my advisor, Prof. Dr. Marek Fuchs for his continuous support, guidance and encouragement over the last four years. I am truly indebted to the second examiner of my thesis Prof. Dr. Michael Bošnjak and additionally to Prof. Dr. Helmuth Berking, Prof. Dr. Andrea Rapp and Prof. Dr. Jens Krüger for acting as members of the examination committee.

My office mates over the years Dr. Britta Busse, Heide Zobel, Tanja Kunz, Simon Laub, Dennis Schumacher, Dayana Bossert and Cornelia Neuert were always a pleasure to work with and talk to. There are also many friends and colleagues at TU Darmstadt that helped me through the dissertation: Angela Graf, Jochen Schwenk, Gunter Weidenhaus, Dr. Christian Schilcher and Carsten Hohmann.

I also want to thank Dr. Frederik Funke for his suggestions during the development of the experiments, Jill Bailin for proof reading and all members of the TU Darmstadt Onlinepanel for their participation in numerous Web surveys that made this thesis possible.

Finally, I want to thank my family. I would not have been able to complete this thesis without your support. Thank you!

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

1 INTRODUCTION

Surveys seek to elicit information about a specific group or category of people, simply by asking questions. Whether the questions are asked in person, over the telephone, via a paper questionnaire or a Web survey, all surveys rely heavily on respondents correctly interpreting a pre-established set of questions to supply the information these questions seek (Groves, et al., 2009). Despite the content and the function of any particular question in a survey, one major distinction can be made between closed- and open-ended questions. Closed-ended questions require respondents to choose among a set of provided response options, while open-ended questions are answered by respondents in their own words. Thus, closed-ended questions are quick to answer and easy to analyze, while open-ended questions are more burdensome to answer and require extensive coding (Reja, Manfreda, Hlebec, & Vehovar, 2003). Open-ended questions also suffer from higher item-nonresponse and answers always tend to lose some of their original meaning in the process of coding. Overall closed-ended questions are increasingly popular across all survey modes, whereas open-ended questions are used less frequently (Krosnick, 1999, p. 544).

However, closed-ended questions are not ideal for every purpose or in any situation. For example, answer categories may fail to provide an adequate set of response alternatives for a closed-ended question. Respondents may then break-off the survey, skip the question, or choose the best answer provided, but in any case they cannot express their originally intended answer. On the other hand, open-ended questions do not force respondents into an answer category, but rather allow for qualifying and quantifying responses, and overall responses can be more detailed. Despite missing answer categories influencing responses to closed-ended questions, answer categories in general are likely to influence responses. They might give away the answer too easily or even bias responses. In contrast, open-ended questions resemble natural everyday communication in a far better way than frequently used closed-

answer formats do: we usually ask questions freely and don't provide a set of response options when we are in conversation.

Open-ended questions can be used to ask for frequencies, to obtain short responses or enumerations such as "Which shows do you love most on television?" or even shorter ones asking only for a single word or phrase, such as "What is your city of origin?" They can also be used for narrative questions such as "What is the biggest problem facing the nation?" where respondents are encouraged to freely answer in their own words, creating the possibility of long, thick and rich responses. This thesis focuses on narrative open-ended questions and ways to improve them in Web surveys, using visual and adaptive questionnaire design.

In recent years, Web surveys have rapidly emerged as a common and important form of data collection. Web surveys have mimicked paper questionnaires with respect to their layout and appearance for a long time, but the Internet and its data collection methods offer various opportunities to improve responses via visual design. Overall visual design aspects, such as labels (Schwarz, Grayson, & Knäuper, 1998), pictures (Toepoel & Couper, 2011), the arrangement of answer categories (Smyth, Dillman, Christian, & Mcbride, 2009), the use of colors (Tourangeau, Couper, & Conrad, 2007a), and the ordering of answer categories (Tourangeau, et al., 2007a) affect responses. Various studies have focused on the design of associated answer-boxes with respect to open-ended questions. Among other variables, the size of the input box (Christian & Dillman, 2004; Smyth, et al., 2009), any label or mask attached to the box (Couper, Kennedy, Conrad, & Tourangeau, 2011; Fuchs, 2007, 2009b) or the number of boxes offered in association with a question (Fuchs, 2009a; Keusch, 2012) have been proven to affect the quality and the length of responses to open-ended questions.

Improving narrative open-ended questions in Web surveys using visual design starts with the answer-box in which respondents are supposed to type their answer. The size of the answer-boxes (or input fields) becomes important for the respondent when interpreting the question and calculating their response length and how elaborate to make their

response. If a large box is displayed, respondents might get the impression that they have to provide a longer answer, while a small box might encourage them to shorten their response. Further, small boxes seem to pose a lower response burden while large answer-boxes increase this burden and therefore might provoke increasing item-nonresponse. Finding the ideal box size seems to require a trade-off, balancing item-nonresponse and the length of the answer. In order to improve narrative open-ended questions this thesis is aimed to demonstrate the effectiveness of using visual design by (1) varying the answer-box size; (2) using a counter that indicates the number of characters left to type; (3) dynamic in size growing answer-boxes and (4) answer-boxes where respondents can adjust the answer-box size themselves by a plus and a minus button.

Despite varying the visual appearance of narrative open-ended questions and the answer-boxes used, it is also possible to adapt Web surveys individually to groups of respondents. Based on previously supplied answers, it is feasible to provide specifically designed questions to respondents in an adaptive design. This thesis tests two adaptive design approaches to improving narrative open-ended questions. First, the amount of information respondents type into the response box of an initial open-ended question is used later in the survey to assign them to a custom-size answer-box. Second, a follow-up probe is tested where respondents who didn't answer a narrative open-ended question were assigned to the same question in a closed format, to get at least some information from them. The Web provides multiple opportunities for extending the range of questionnaire features available to the survey designer (Couper, 2001, p. 4) and the adaptive designs make use of them. Adaptive and visual design is supposed enrich the survey experience for respondents and at the same time improve data quality in Web survey.

This thesis is structured into 11 chapters. The second chapter provides an introduction to Web surveys by first comparing the different modes of data collection. Common types of Web surveys are discussed, and their particular strengths weaknesses highlighted. Then a brief view of how

Web surveys work and how they can be administered is presented, as well as requirements for designing effective Web surveys.

Chapter 3 is concerned with the total survey error framework in order to classify the sources of errors that affect inference in surveys. The framework differentiates between errors of representation and errors that relate to the measurement process. Discussing the total survey error perspective is especially important because the quality of data obtained by Web surveys is often questioned. Since this thesis is focused on ways to improve responses to narrative open-ended questions in Web surveys, it is important to review the sources of error, especially in the measurement process first.

In chapter 4, the question–answer process is highlighted and discussed by focusing on the way respondents process a question. To answer a question, respondents have to decode the question's meaning before they retrieve information to form and provide response. The difficulty of a question, the respondent's motivation, and personal characteristics all affect responses. Since narrative open-ended questions are especially burdensome to answer, they require a certain degree of skill and motivation on the part of the respondent. When improving narrative open-ended questions, the question–answer process has to be considered to identify the best places for improvement.

Chapter 5 provides an overview on different question-and-response formats and their implementation in Web surveys. Closed-ended questions require respondents to choose among a set of response options, while open-ended questions are answered in the respondents' own words. After elaborating various types of closed-ended questions and their strengths and weaknesses, this chapter discusses open-ended questions. They can be used to ask for frequencies, short responses, or enumerations, but they can also be designed to elicit narrative responses. After this look at the different types of open-ended questions possible, their strengths and weaknesses are discussed in great detail.

The sixth chapter, on visual questionnaire design, emphasizes the importance of numerical, graphical and symbolic language in Web

surveys. In interviewer-administered modes of data collection, verbal language is the primary channel of communication and source of information. In Web surveys, however, communication is through the visual channel. Therefore, the visual design language affects the way respondents read, interpret, and answer questions in Web surveys, and can be used to improve responses.

Chapter 7 explores the first part of the research question. The literature reviewed in the prior chapters is integrated into a framework that describes how open-ended questions in Web surveys can be improved by using visual design language. Four approaches that should improve data quality in narrative questions are described: answer-box size, a counter indicating the number of characters left, dynamic, growing answer-boxes and respondent-adjusted answer-boxes. All four approaches are put to a test in three experiments reported in chapter 8.

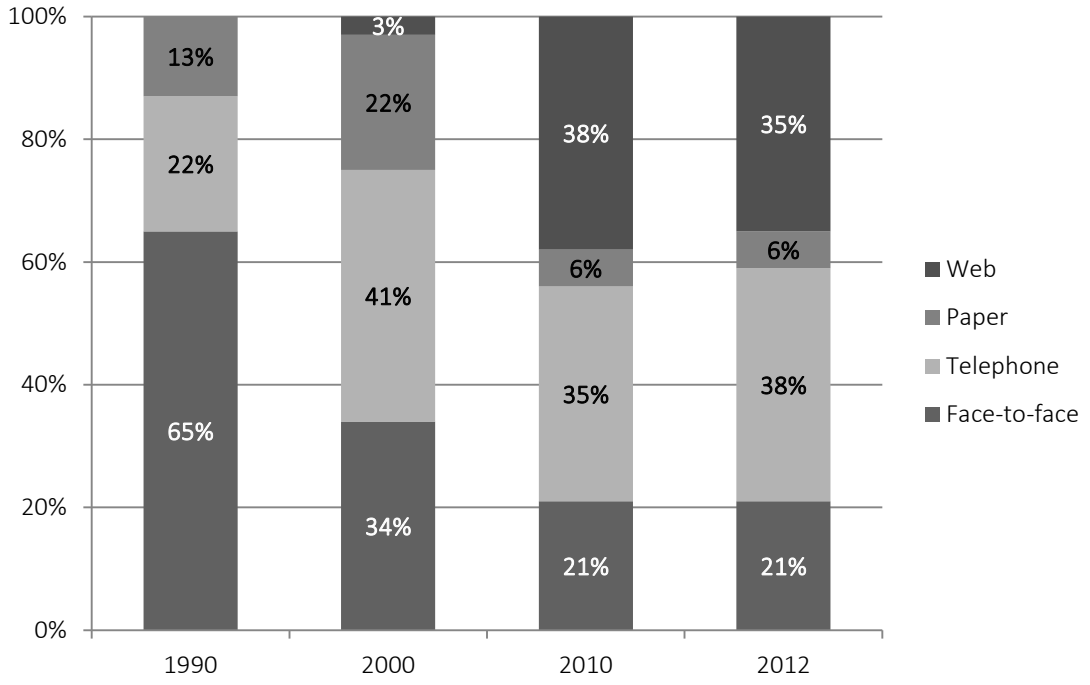
Beyond using visual design to improve narrative open-ended questions, Web surveys offer multiple opportunities to improve responses. Chapter 9 introduces adaptive questionnaire design as an additional approach to improve responses to narrative open-ended questions. By adapting the visual appearance of open-ended questions, based on prior answers, responses should be enhanced. In chapter 10, two adaptive design experiments are reported on. In the first experiment, the answer-box sizes for open-ended questions are adapted based on the length of prior responses, while in the second experiment a closed follow-up probe is used if respondents did not answer an open-ended question.

Finally, chapter 11 summarizes the results of the experiments conducted for this thesis. Major findings are discussed and implications for survey design and evaluation are specified. The chapter closes with an outlook on future research and additional opportunities to further improve open-ended questions in Web surveys.

2 WEB SURVEYS

In the relatively short time that it has taken the Internet to reach widespread penetration in the world, Web surveys have also rapidly emerged as a major form of data collection (Couper, 2008a, p. 1). Especially in market research, many companies switched to Web surveys to collect their data. The German association of private market and social research agencies (ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.) asked German market research companies about the number of interviews carried out in the different modes of data collection (see Figure 1).

Figure 1. Percentage of interviews for each mode conducted by members of the German association of private market and social research agencies (ADM, 2012)



While in 1990 almost two-thirds of all interviews were carried out face-to-face, in 2012 only a fifth (21 percent) of interviews were administered in person by an interviewer. Across the same period of time, the percentage of telephone interviews almost doubled, replacing face-to-face interviews to some extent as the predominant mode of interviewer-administered survey. The same pattern also seems to be apparent for

self-administered modes of data collection, where paper-and-pencil studies were replaced by Web surveys. While the number of surveys using paper questionnaires increased up until 2000, their percentage has become vanishingly small in recent years. At the same time, Web surveys have become more important in market research, accounting for approximately 35 percent of all interviews carried out in 2012 (ADM, 2012).

Before we take a closer look at different types of Web surveys and how they can be administered, it is important to point out how Web surveys compare to other common modes of data collection in the next section.

2.1 Comparing Web surveys to other modes of data collection

The most basic difference between different modes of data collection can be made between surveys that are administered by an interviewer and surveys that are administered by the respondent. Interviewer-administered surveys can be face-to-face surveys as well as telephone surveys, while self-administered surveys use paper or Web questionnaires.

Face-to-face surveys administer the questionnaire in person by an interviewer, typically in respondents' homes. The interviewer asks questions and records the answers. Almost any type of question can be used, and even complex and long questionnaires can be administered effectively if the interviewer is well-trained and the questionnaire well-designed and tested. The presence of an interviewer facilitates convincing a person to participate in an interview. The interviewer can also monitor the respondent, keep him motivated and involved, and provide assistance if needed. However, incorporating an interviewer into the administration of a questionnaire also has its downside. An interviewer's presence can influence responses, especially to sensitive questions. Also, the way an interview is administered will vary among multiple interviewers and therefore affect responses differently. Another major disadvantage of face-to-face surveys is that they are expensive to

administer. Interviewers must be hired and trained, and go from one address to another to record the interviews. The fieldwork will therefore take more time in comparison to any other survey mode. Samples for face-to-face surveys are usually drawn from population registers of the whole population (Bethlehem & Biffignandi, 2012). In addition, face-to-face interviewing usually does not exclude any persons from the target population. Overall, the data quality from well-designed face-to-face surveys is high, but the costs to administer face-to-face surveys are high as well. Balancing survey costs and data quality is essential to the process of deciding in which mode a survey will be conducted.

Telephone surveys share many features with face-to-face interviewing, since both modes are administered by an interviewer. But interviews conducted via landlines or mobile phones compromise survey data quality to some extent, because not every household has a telephone and can be contacted in this way. The sampling of telephone numbers is often realized by random digit dial (RDD) procedures generating valid random numbers in order to include even unlisted telephone numbers in the sample (e.g. Glasser & Metzger, 1972; Waksberg, 1978). Of course, sampling can also be realized using telephone directories but not every household is listed in these directories, making undercoverage a serious threat to data quality (see chapter 3 for a detailed description of coverage error).

Even though telephone surveys are administered by interviewers, the interviewer's is less pronounced in comparison to face-to-face interviews. Since communication includes only verbal (not nonverbal) language, an interviewer has fewer opportunities to monitor and support respondents. It is also harder for an interviewer to convince respondents to take part in a survey and keep them involved in the questionnaire. Therefore, the questions asked should be as simple as possible to answer, or broken up into multiple questions to facilitate the response process. In general, burdensome questions (such as open-ended questions) should be avoided in telephone surveys since respondents' cooperation is lower in comparison to face-to-face surveys (Bethlehem & Biffignandi, 2012). One major advantage of telephone surveys is the

lower cost to administer them, because interviewers don't have to travel, and multiple interviews can be realized in a short time.

Even though telephone surveys are less expensive than face-to-face surveys, they are still quite expensive since both modes require interviewers to collect the data such as mail or Web surveys. Instead of administering a survey in-person by an interviewer, a questionnaire could be completed by the respondent alone. The reading of questions and recording the answers would be carried out by the respondent. Such surveys could be conducted at very low costs and address a large number of respondents at the same time. But the absence of an interviewer would also mean that no additional guidance or support could be provided to respondent. All the information necessary to respond to the survey would have to be transmitted through the questionnaire, and the question would have to be formulated in advance to be as easy to understand as possible. In general, long questionnaires should be avoided, and reminders and incentives used to increase response rates, which are usually lower in self-administered surveys (e.g. Dillman, et al., 2009). Overall, self-administered surveys do not allow the flexibility of interviewer-administered modes. However, they are less intrusive and respondents are less prone to feel pressured to provide socially desirable responses, since the perceived privacy is higher when no interviewer is present. While all these aspects relate to both mail and Web surveys, there are some differences between these most common self-administered modes as well. For example, in paper questionnaires respondents are totally on their own, while Web surveys allow for more assistance in formatting an answer and routing the respondent through the questionnaire based on prior responses rather than by using arrows and instructions (Couper, 2008a). In addition, inviting participants to a Web survey is much easier and more economical to realize in comparison to mail surveys, which require more time to send out questionnaires and recode the responses. Another main differentiator between mail and Web surveys is the sampling process. Sampling for mail surveys requires a list of postal addresses that are often available from a country's postal service organization or the residents' registration office. However, a

sampling frame of e-mail addresses is only available for specific populations, such as university students, for example.

All these modes of data collection have strengths and weaknesses that must be weighed before reaching a decision on which mode to choose. In addition, it can be helpful to mix different modes of data collection to counterbalance their various advantages and disadvantages. For example, providing an online option in a telephone survey can increase response rates (e.g. de Leeuw & Hox, 2010; Dillman, et al., 2009). Combining paper and Web questionnaires can also be sensible. Inviting respondents via mail and providing a Web alternative in the cover letter can increase response rates because a paper questionnaire will not have to be sent back to the researcher. Moreover, sample persons without valid e-mail addresses can be contacted via their postal addresses, if available, to improve data quality (e.g. Dillman, Smyth, & Christian, 2008; Kaplowitz, Hadlock, & Vevinde, 2004; Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2007; Messer, Edwards, & Dillman, 2012).

At first glance, mixing different modes of data collection appears to be promising. But different modes also imply differences between the communication channels used to collect data. This is so, especially when self-administered surveys (that rely on a visual channel) are combined with interviewer-administered modes (that use an aural communication channel). To what extent the data collected from different modes can be compared and combined is a valid question (for a review on mixed-mode surveys see de Leeuw & Hox, 2010; Dillman & Messer, 2010)

This is a rather brief description of the most common survey modes and how they compare. The closer look at the total survey error concept in chapter 3 will further help to illustrate the factors that contribute to Web data quality vis-of-vis other survey modes.

2.2 Types of Web surveys

The key challenge in Web surveys is deciding on the sampling process and therefore how to actually invite respondents to a Web survey. “The population of Internet users is dynamic and difficult to define”, meaning

that extrapolating to the full population (not just Internet users) based on Web surveys will be difficult (Couper, 2011, p. 5). Therefore, Web surveys are not ideal for every target population. The following closer look at the different types of Web surveys will help to define when Web surveys are suitable for data collection, and when their results might be questionable.

Overall, Web surveys can be differentiated by their sample selection using either probability sampling or non-probability sampling methods (Couper, 2000). Probability-sampled Web surveys can be list-based surveys of high-coverage populations, such as students at a university or employees of a company. Usually for students as well as for employees, a list of e-mail addresses is available in order to invite all students or employees, or to draw a random sample based on that list of e-mail addresses. Either way the targeted population is easy to define, identify, and invite to a Web survey, as long as every student or employee has an e-mail address and access to the Internet. If a list of e-mail addresses is incomplete, the sampling process and the later survey results are compromised.

When conducting Web surveys on the general population, it is much harder to identify the target population since there is no register of e-mail addresses for drawing a sample and no method for generating e-mail addresses randomly. In order to deal with this dilemma, researchers borrowed probability sampling methods from telephone or face-to-face surveys to pre-recruit respondents for Web surveys; the recruiting is administered offline, while the survey itself is conducted online. Since this way of recruiting respondents is rather elaborate, it is best used for developing probability-based online panels where individuals agree to participate in future studies as well, such as the Dutch LISS (de Vos, 2010), French ELIPPS (Cornilleau, 2013) or the German Internet Panel (Blom, et al., 2013). Online panels are basically a managed community of Internet users who have agreed to participate in several Web surveys over an extended period of time. Since the sample persons are supposed to participate in multiple Web surveys, the expenses of the sampling process can be spread over a larger base.

Probability-based Web surveys can also be part of a mixed-mode design where the online survey is an available option. For example, respondents in a probability-based recruited telephone survey can switch the mode, if they feel more comfortable with a Web option or prefer to respond to the questions at another time. In such a case, the interviewer will provide a link for the respondent where he can easily continue the survey in the Web format. Of course, a mixed-mode design is not limited to telephone surveys and can easily be provided in paper-pencil studies or face-to-face interviews as well. The purpose of a mixed-mode design is to reduce the response burden, decrease the survey costs, and minimize coverage-related errors (Manfreda & Vehovar, 2008, p. 266).

Finally, probability-based Web surveys can be intercept surveys where visitors to a Web site are systematically sampled. For example, every *n*th visitor to a Web page would receive an invitation to take part in a Web survey, via a pop-up window in the browser (Couper, 2000). Another intercept approach is so-called "river-sampling", where potential respondents are asked a few screener questions at a Web site that direct them to a Web survey without having them join a panel (Couper, 2011, p. 7).

While probability-sampled Web surveys can cover a targeted population adequately if same sort of sampling frame is used to draw a sample, non-probability-based Web surveys do not use a sampling frame and their results usually don´t hold for a specific population even when the data are weighted (see chapter 3 for a closer look at sampling and weighting data).

Non-probability Web surveys can be online access panels (or opt-in panels) and are especially common in market research (Manfreda & Vehovar, 2008). Access panels use almost any method possible, such as Web site adds, banners, or e-mail, to contact people who might be interested in joining a panel and participate in surveys (Couper, 2011). Therefore, the potential respondent decides whether to participate in the panel or not. The research has no supervision over the sampling process and survey results will not extend to the general population.

The same holds true for unrestricted self-selected Web surveys where respondents are invited to a Web survey via open invitations on different Web sites. They then participate in one Web survey (not multiple surveys as in the case of online access panels). As a result, data quality is compromised and inferences from the self-selected Web survey remains of questionable value, even though many of these studies claim otherwise (Couper, 2011).

Finally, Couper (2000) names polls for entertainment as a non-probability-based Web survey type. Usually answered by visitors to a specific Web site, these polls do not claim to be representative. Instead, they are meant to be entertaining, and they provide feedback to the site's audience. The "question of the day" on many news sites is an example of entertainment polls which are not really "surveys", since they only ask one question about a current topic (Manfreda & Vehovar, 2008, p. 267).

Overall, the main differentiator between these different types of Web surveys is the underlying sampling process. That's especially important when it comes to Web surveys whose results are supposed to hold for the general population. Results of non-probability-based Web survey are more likely to be biased, since the proportion of the population with no chance of being part of the survey is larger in comparison to probability-based Web surveys (Couper, 2011). For example, inviting students using their enlisted e-mail addresses ensures that almost everybody in a sample gets the invitation, while a banner ad on a university Web site won't be seen by every student, and may be addressed by someone outside the university who is not supposed to be part of the survey population at all. While broadly classifying Web surveys as probability-based or non-probability-based approaches may appear to be valid, these two dimensions can also be thought of as ends of a continuum, with one end being self-selected volunteer surveys. At the other end of the continuum are probability-based surveys, where those without Internet access are provided access to accomplish high response rates. "In practice, most Web surveys lie somewhere between these two end-points" of the continuum (Couper, 2011). Before taking a closer look at

sampling and data quality in chapter 3, the next section describes how Web surveys actually work.

2.3 Web survey administration

Web surveys are computerized. The technical implementation of Web surveys can be realized client-side on the respondent's computer, or server-side on a Web server that hosts the questionnaire. Client-side surveys can be downloadable forms that are later sent back, perhaps via e-mail. Here the Internet is used for transmitting a questionnaire that is answered on the respondent's computer to be returned to the researcher after the questionnaire is answered. However, this type of online survey was only used at the very early days of Web survey research. Currently it is rarely used, because of technical limitations and security concerns (Couper, 2008a, p. 3).

The common administration of Web surveys is based on server-side designs like the surveys conducted for this thesis. Generally, this means that respondents answer the questionnaire through their browser, with the answers being transmitted to the server whenever the "submit" or "next" button is pressed. The survey itself is located on a Web server during the entire process, requiring a continuous Internet connection with the respondent. By clicking a link in an e-mail or at a homepage, for example, a request to deliver the Web survey to the respondent's browser is sent to the server hosting the questionnaire. Server-side programmed surveys also guarantee, at least to some extent, that the survey runs as expected since no special hard- or software is required from the respondent (Birnbaum, 2004). The content of Web surveys – the questionnaire – is usually structured via HTML (Hyper Text Markup Language), which allow only rather static designs, at least until HTML 5 brings substantial changes in the near future (Funke, 2010). However, by using JavaScript or Flash, this lack of design opportunities can be compensated for, and more highly styled questionnaires realized. The tradeoff in using JavaScript or Flash is the necessity of enabling the respondent's browser, via plug-ins, to read these scripting languages. If a

respondent's computer does not provide these plug-ins, the Web survey user's experience will be restricted, or the questionnaire might even not run at all. New, more sophisticated, questionnaire designs might exclude targeted respondents from a survey if their hard- or software is not capable of carrying them out (Dillman & Bowker, 2001). Therefore, Web surveys should be conducted by using basic and common technology, rather than demanding approaches that might be posed by new hardware or software (Reips, 2002, p. 248).

Despite using HTML to fashion a questionnaire and address its meaning, HTML forms are used to capture the responses. A HTML form has a front end that displays a type of button, and a back end that handles the forms submission. Web survey respondents therefore visit a page where the questions are presented on a form (using a radio button, check box, text box, etc.). The respondent fills out the form in a browser and submits the response. Afterwards, the browser sends the submitted form's data to a Web server, where the data are processed and stored in a log file or a database (Funke, 2010). At the same time, the Web survey's next page is sent back to the respondent, and that process continues through to the conclusion of the questionnaire.

This thesis concerns the visual design of open-ended questions and how it can be improved in Web surveys. Visual design affects responses in many ways (see chapter 6 for an overview), while it depends heavily on the respondent's hardware and software, and especially the Web browser (Funke, 2010). Varying display sizes and screen resolutions will result in different appearances of the same Web survey across respondents. With the growth in smartphones and tablet computers, the researcher has even less control over the way a survey is perceived. As well as the ongoing fragmentation at the hardware side, the software on a respondent's device also determines if and how the questionnaire is displayed (for example, when the survey contains flash animation, sound, videos or other content requiring browser plug-ins). The Web browser particularly affects the appearance of Web surveys since they may interpret the HTML code differently and display diverging style elements. For example, answer-boxes displayed with open-ended questions can be

presented with scroll-bars from the beginning, or later, depending on the need. Grab-handles for re-sizing the answer space are standard in some browsers while not available in others. Using cascading style sheets (CSS) is the best way to provide almost identical visual appearances across all different Web browsers. Despite the technical implementation, there are more aspects to be considered when designing Web surveys and Web survey questionnaires.

2.4 Web questionnaire design

Web surveys are self-administered, which means that there is no interviewer to administer the questionnaire, ask the questions verbally, or record the answers (de Leeuw & Hox, 2008). Instead, the respondent is on his own when answering the survey, making questionnaire design extremely important. In any survey, well-worded questions are essential, but in Web survey all information, such as the question itself, or a decision to skip instructions or explanations have to be executed by the respondent. “As a consequence, the visual presentation of questions and the general layout of the questionnaire are far more important in self-administered questionnaires” (de Leeuw & Hox, 2008, p. 240).

It is important to point out that designing Web surveys is different from designing Web sites in terms of their structure and goals (Bethlehem & Biffignandi, 2012, p. 190). And even though Web surveys often look a lot like paper questionnaires pinned to a screen, not all principles of designing paper-and-pencil questionnaires apply to Web surveys. One of the most apparent differences between both modes is that in Web surveys researcher can choose between a paging or a scrolling design. A scrolling design displays all the questions on a single screen, more like a classic paper-and-pencil questionnaire. The respondent scrolls though the questionnaire and answers one question after another until he submits the results at the end of the survey. In a paging design, the questions are presented on multiple pages, each showing only a limited number of questions at the same time. In a paging design the respondent goes through the questionnaire and submits every page one at a time,

rather than submitting the whole questionnaire at once, as in the scrolling design.

The scrolling design was especially common in the early days of Web survey research because it resembled the user experience of Web pages at that time (Dillman, 2007). In a scrolling design, respondents can easily scan the entire questionnaire, determine its length, and move back and forth within the questionnaire. But there are some downsides to scrolling designs. For example, the questionnaire must be completed all at once, and until the submit button is pressed, no data are stored. Also, the opportunity to glance at later questions might influence prior responses in an unintended way, and skipping or routing procedures had to be presented as instructions. In addition, the order in which the questions are answered cannot be controlled in a scrolling design, and consistency checks of responses not applied to a single page survey (Couper, 2008a).

Longer and more complex surveys are better realized via multiple screen designs. Because routing or skipping instructions can be automated, instead of instructing respondents how and when to skip a question, with multiple screen designs, respondents are automatically directed to the questions they are supposed to answer. Overall, the advantages of a paging design correspond to the disadvantages of the scrolling design. Respondents don't have to scroll as much as in the paging design. Since the responses provided on every single page of the survey are stored, partial interviews can be stored. Interactive feedback such as running tallies or probing questions can be implemented into paging surveys. Several studies compared paging and scrolling designs (e.g. Peytchev, Couper, McCabe, & Crawford, 2006) and found only minor differences between scrolling and paging designs in response rates, break-offs, and response time. In general, the scrolling design should be used for short surveys that contain only a few skip conditions and require a fair amount of scrolling from the respondent. Longer surveys (such as the ones conducted for this thesis) that contain a lot of skip conditions or randomized assignments to different questions are best realized in a paging design. All Web surveys carried out for this thesis relied on a

paging design in order to randomly assign respondents to different designs of the same question.

One advantage of the scrolling design is that the respondent can easily note the questionnaire's length. Even without scrolling to the bottom of a questionnaire, the scrollbar at the right side of the browser provides an indication of the questionnaire's length (Manfreda & Vehovar, 2008). In a paging design, the respondent is usually unaware of the number of upcoming questions. However, a progress indicator that shows the percentage of questions a respondent has already answered can address this issue (e.g. Conrad, Couper, Tourangeau, & Peytchev, 2010).

Despite the overall arrangement of questions, a Web survey's welcome screen is very important when it comes to designing effective Web surveys. It's the first impression a respondent gets of the survey. Introducing the survey and naming its purpose as well as emphasizing the ease of responding and instructing respondents about how to proceed to the next question, will increase participation (Dillman, 2007). In addition, addressing surveys specifically to sample persons (such as inviting them by name) can further improve respondents' cooperation (Kaczmirek, 2009). In general, participation should be made appealing to the respondent across the whole questionnaire.

Similar to the importance of the welcome screen, the first questions asked are also critical, since it defines for the user to some extent whether the questionnaire will be easy to answer or not. Therefore, "the initial question should be interest-getting, and confirm to the respondent that it is worthwhile to continue" (Dillman, Tortora, & Bowker, 1998, p. 8). Of course, subsequent questions should be designed carefully as well, but in comparison they are less essential. The wording of Web survey questions does not differ too much in comparison to other survey modes as long as the same standards for formulating good questions are considered in every case (Manfreda & Vehovar, 2008). A closer look at possible and suitable question-and-response formats in Web surveys is provided in chapter 5.

In his Tailored Design perspective, Dillman (2007) describes ways to establish respondents' trust and perception of increased rewards, while at the same time lowering the costs of participating in a survey. By providing a token of appreciation in advance or stressing the importance of the survey, trust can be established while rewards can be provided by interesting questions, asking for the respondent's advice, or just by saying "thank you" (Dillman, 2007, p. 27). The cost of participation can be reduced by a short, easy to read and answer questionnaire that requests not too much personal information and is consistent over the whole questionnaire. The strength of this design perspective is the integration of social exchange theory to better understand when and why individuals respond to a survey. However, these design principals apply to other survey modes as well as does the need for thoughtful and well-written questions.

Beyond the question itself, a questionnaire carries a lot more information than just plain text, which is another reason that the nonverbal (or visual design) language is very important (e.g. Christian & Dillman, 2004; Christian, Dillman, & Smyth, 2007; Fuchs, 2010; Stern, Dillman, & Smyth, 2007). This holds especially true for Web surveys where various graphic and multimedia design features can be easily implemented (Emde & Fuchs, 2012). In addition, in some cases, visual design is the only way to communicate to a Web survey respondent. The absence of an interview does not allow for correcting respondents' mistakes or the provision of motivation when needed (see chapter 6 for a closer look at visual design). Internet technology greatly extends the opportunities in Web surveys in comparison to other survey modes. Thus, one might easily suggest that the use of colors and an overall visually pleasing experience will increase participation and data quality. Mahon-Haft and Dillman (2010) compared the results of two versions of the same survey. The first one carried an aesthetically pleasing design, while the second one used an unattractive color scheme and positioned the question text in rather awkward ways. Surprisingly, the impact on response and completion rates of the aesthetically-displeasing screen design was only minor. However, in more burdensome questions, such as open-ended

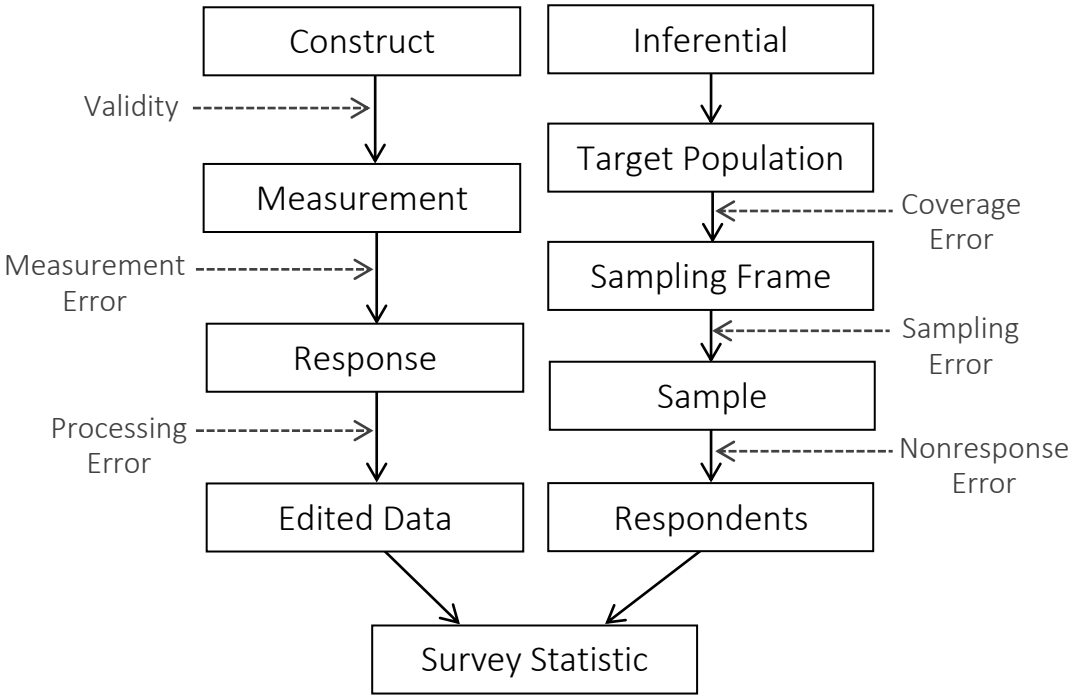
questions, respondents to the unappealing design were less likely to provide complete responses (Mahon-Haft & Dillman, 2010).

Overall, Web survey design is not that much different when compared to paper questionnaires, and most of the design principles for designing paper questionnaires apply to both survey modes (see Dillman, 2007 for an overview). And even though the opportunities Web surveys offer are extensive, their use should always reflect the purpose of increasing data quality. It is important to design Web surveys to some extent in ways that mimic the conventional format usually used in self-administered surveys, to accommodate the expectations a respondent holds towards a questionnaire (Manfreda & Vehovar, 2008). At the same time, it is important to point out that paper and Web surveys are also different in many ways. The flexibility of Web surveys can improve the quality of data obtained in comparison to paper questionnaires, and offer a great opportunity in terms of an improved measurement process (Couper, 2011, p. 47). At first glance, the questionnaire seems to be the most essential aspect of designing and conducting a survey. But the actual measurement process is another important aspect that contributes to data quality. What else is important to obtaining high-quality data, especially in Web surveys, is discussed in the light of the total survey error concept in chapter 3.

3 THE TOTAL SURVEY ERROR CONCEPT

Developing a survey, whether it is carried out online or offline, typically follows a required sequence of steps, starting with defining the relevant research objectives (Groves, et al., 2009). The sampling frame and the mode of data collection, which influence each other, are chosen next. Then the sample has to be drawn from the sampling frame and the questionnaire designed and pre-tested based on the prior survey-mode decision. After recruiting and assigning respondents to the questionnaire, the data are collected. Finally, the responses are edited and coded, and if necessary, weighting and analysis performed. This very brief description of the survey process helps to convey what single steps have to be considered to finally provide solid survey estimates. In addition, this process perspective helps to identify the sources of error and how they relate to each other.

Figure 2. Survey lifecycle from a qualitative perspective (Groves & Lyberg, 2010, p 856).



The total survey error concept (Deming, 1944; Groves, et al., 2009; Groves & Lyberg, 2010; Kish, 1965) refers to all kinds and sources of errors that may arise in this process. Sources of error can be conceptualized in several ways. “The deviation of a survey response from its underlying true value defines the survey error” (Biemer, 2010b, p. 817) and cannot be entirely prevented. According to this concept, errors can be allocated to the measurement or representation stage in the total survey error framework, displayed in Figure 2.

The construct represents the researcher’s intended scope of information, that (for example) results in the selection of a survey question in order to obtain this information. If the question used is not adequate to measure the construct, its validity is limited. But even if a question is thoughtfully chosen, the responses provided by respondents can deviate from the true value and result in measurement error. In addition to the previous noted error sources, processing error relates to false edited data and contributes to survey error in the overall survey statistics. Despite measurement-related errors, the representation of results affects the overall error as well. First, the population that the results are applied to must be defined and the target population identified. Afterwards the sampling frame collects all targets possible, although it may miss some, resulting in coverage error. Next, a sample has to be drawn from the sampling frame, which results in sampling error when targets have no chance to be included to the sample, for whatever reasons. Finally, if respondents refuse to participate or were not assigned to the survey, nonresponse error can affect results. “Minimizing total survey error subjects to cost and timeliness constraints” (Biemer, 2010b, p. 821) is the goal of optimal survey design. It is important to point out that errors in the whole survey process do not necessarily compromise data quality as long as these errors are random. Systematic errors can be more prevalent in inducing a bias to the data obtained. The sources of error in the lifecycle of a survey are discussed in the next section, and focus on the Web survey mode.

3.1 Coverage error

When a survey's target population is identified, the next step in the process is defining the sampling frame. The sampling frame is "a list of target population members or another mechanism used for drawing the sample" (Biemer, 2010b, p. 824). Coverage errors occur when the sampling frame does not exactly cover the targeted population. People that are not part of the target population but falsely included in the sampling frame result in overcoverage. Undercoverage describes the opposite situation, when people that are part of the target population have no chance to be included in the sampling process, for example, mobile-onlys in landline telephone surveys, or people without Internet access in Web surveys. Despite being falsely included or excluded from the frame, duplicates are also problematic since they heighten a subject's chance to be drawn from the sampling frame. Overall, the difference between the target population and the sampling frame is what defines the coverage error (e.g. Couper, 2000).

Table 1. Demographics of Internet users in 2000 and 2011 in the United States (Zickuhr & Smith, 2012)

	2000	2011
Adults	47%	78%
Men	50%	80%
Women	45%	76%
Age		
18-29	61%	94%
30-49	57%	87%
50-64	41%	74%
65+	12%	41%
Household income		
Less than $30,000/year	28%	62%
$30,000-$49,999	50%	83%
$50,000-$74,999	67%	90%
+$75,000	79%	97%
Education Level		
No high school diploma	16%	43%
High school grad	33%	71%
Some college	62%	88%
College +	76%	94%

Coverage error in Web surveys is particularly affected by undercoverage since many people still have no Internet access. Even though online penetration rates in developed countries are at a high level and still growing, between 10 and 30 percent of the population are not yet online (Couper, 2011). Across the 27 countries of the European union, reported rates for individuals using the Internet at least once a week increased from 36 percent in 2004 to 70 percent in 2012. Usage in Iceland, Sweden, Norway, and the Netherlands was above 90 percent in 2012 ,while in Portugal, Italy, Greece, Romania, and Bulgaria, reported usage was below 55 percent. In 2012, in Germany 78 percent used the Internet at least once a week, in the UK 84 percent, in France 78 percent, and in Spain 65 percent (Eurostat, 2012).

Despite Internet penetration affecting coverage error, more important is the way in which those individuals with Internet access differ from those who are not online (see Table 1). Most Internet users are younger, male, and better educated (e.g. Bethlehem, Cobben, & Schouten, 2011; Couper, 2000; Couper, Kapteyn, Schonlau, & Winter, 2007; Fuchs & Busse, 2009; Zickuhr & Smith, 2012). This systematic difference between the online and not-online population is what makes undercoverage a serious threat to data quality in Web surveys.

By adjusting the weighting, these demographic differences can be compensated for somewhat, but Internet users also differ in their perception of their own health, activities of daily living, and attitude measures, limiting the power of data adjustment (Couper, Kapteyn, et al., 2007; Fricker, Galesic, Tourangeau, & Yan, 2005; Schonlau, van Soest, Kapteyn, & Couper, 2009).

But even if everyone in the target population does have Internet access, the target population is not necessarily easy to identify. Since there is no register of e-mail addresses, Web surveys of the general population are not suitable with respect to the total survey error. In order to compensate for this problem, several general population online panels recruited their panel members via telephone, and provided online access to those persons who were not online yet (e.g. the Dutch LISS panel and

the German Internet Panel). However, providing Internet access does not completely prevent coverage error since, for example, women and the elderly remain underrepresented in the RDD-recruited LISS panel in comparison to national statistics data (de Vos, 2010). In addition, Bosnjak et al. (2013) compared the characteristics of a probability online panel-based sample to a probability national sample survey, and found differences for age and education as well as a few personality traits, resulting in a sample composition bias.

While Web surveys of the general population struggle to deal with coverage issues, they are especially suitable for restricted target populations such as employees, students, or establishments where an exhaustive e-mail list of the target population and a sampling frame are available (Kaczmirek, 2009).

3.2 Sampling error

Instead of surveying all people in the sampling frame, in most cases a sample is drawn using either probability or non-probability-based methods, in order to save on costs or due to logistical infeasibility (Groves, et al., 2009). Sampling requires a sampling frame, which is a list of all elements in the target population and a method to extract elements from this frame. While coverage error refers to people missing from the sampling frame, sampling error arises when not all members of the sampling frame are measured (Couper, 2000). The sampling design depends heavily on the sampling frame, which can be carried out in many ways and for several purposes (e.g. Bethlehem & Biffignandi, 2012; Bethlehem, et al., 2011; Kish, 1965).

Ideally, probability-based samples for Web surveys (and other survey modes as well) can be drawn that allow every person in the sampling frame the same chance to be selected. Probability-based sampling methods include Web surveys using list-based sampling frames, intercept surveys recruiting via pop-up windows, as well as Web options for mixed-mode surveys or pre-recruited online panels for special target populations or the general population (see chapter 2 for different types

of Web surveys). The ideal Web survey sampling frame would be a complete list of e-mail addresses for every person in the target population. For specific surveys such an exhaustive list of e-mail addresses is often available, such as when conducting a job satisfaction survey at a company, or a student survey at a university. As mentioned before, there is no national register of e-mail-addresses and therefore no sampling frame available to draw a general population sample.

Finally, there a multiple ways to draw a sample. Simple random samples ensure that every possible subset of the target population have the same probability of being selected. Simple probability samples are especially useful if little is known about the target population, for example, if only a list of e-mail addresses without further information is available. Even though simple random samples are often considered to be ideal, the sampling process can be prone to errors, especially when the sampling frame holds only a small number of objects, or only a small sample is drawn from the frame. Sampling variance as a source of error refers to the fact that even with an ideal probability sampling process, results will still be subject to variance since they are based on a subsample of the whole sampling frame. Results can differ between two random samples drawn from the same frame just by chance, especially if the frame or the sample drawn consists of a small number of individuals. One of the ways to reduce sampling variance as a source of error is by increasing the sample size. The sampling process can be also adjusted in other ways to reduce variance related errors.

Stratified sampling methods divide the target population into smaller sampling frames that are proportional to the total targeted population. These so-called strata are formed using characteristics of the individuals that are sampled, such as gender or education. The procedure reduces sampling variance as a source of error since all members of every stratum selected will be apparent in the sample. The actual sampling process within the strata selected then is usually drawn probability-based. Simple random samples require a sampling frame that provides a contact list (e.g. street addresses, e-mail addresses, or phone numbers) of the persons in the target population. When these contact lists also

include information such as gender, employment status, or age, stratified sampling methods can improve the efficiency of the sampling process.

Cluster sampling methods use different-sized samples to survey a target population. For example, for an education survey, one might first draw a sample for the different states, then the schools in those states, and lastly, for classes within the schools. Cluster sampling can also be combined with unequal probability-based sampling methods. Instead of selecting every school probability-based from the sampling frame, the chance of being selected might be set in proportion to the number of pupils, in order to provide a greater chance for large schools to be selected than smaller ones. However, the downside to any of these cluster sampling procedures is the fact that individuals who are apparent in one cluster are often related and their responses therefore similar. To compensate for this similarity, more individuals must be sampled than when using a simple random sample. In addition, it is possible to combine different sampling methods into multistage sampling procedures (Lohr, 2010).

However, random errors do not compromise data quality as much as systematic errors bias survey results. Sampling bias arises when some members of a probability sample have a reduced or no chance being selected. If these individuals differ systematically from those selected for the survey, results will depart from the corresponding frame population. Web surveys that use non-probability sampling are especially prone to sampling biases. In non-probability samples it is left up to the individual to participate while in probability samples, the researcher selects the individual from the sampling frame. Biases in entertainment polls as well as self-selected surveys, volunteer panels, or opt-in panels are therefore highly likely (Couper, 2000).

Overall the largest problem of conducting large-scale Web surveys is the lack of a sampling frame, since no registry or list of e-mail addresses exists. List-based random sampling is only feasible for specific populations since it is not possible (at least for now) to construct e-mail addresses randomly (although constructing lists of phone numbers for

telephone surveys is possible). Therefore, only telephone-recruited sampling frames allow Web surveys that generalize to the general public, like the Dutch LISS panel (e.g. de Vos, 2010) or the German Internet Panel (e.g. Blom, et al., 2013). Another opportunity for general population random sampling arises from the rapidly growing number of smartphones. Since mobile phone numbers can be randomly computed, there is a chance that Web surveys in the near future might use phone numbers to invite respondents to answer surveys via their smartphone browser, to accommodate sampling related errors in Web surveys (e.g. Emde & Fuchs, 2013).

3.3 Nonresponse error

Drawing a sample and inviting respondents does not imply that every unit selected will answer the survey. If a targeted person does not respond to the survey, whether he or she refuses to participate or was unable to be contacted at all, the consequence is unit nonresponse. Item-nonresponse on the other hand refers to not answering single or multiple questions and is therefore related to the measurement process and error described later.

Unit nonresponse can be caused by a failure to contact a sample unit because the person refuses to participate, is not able to respond, or can't be contacted at all. Unit nonresponse only refers to the eligible sample, while ineligible units of a sampling frame do not contribute to nonresponse. Several aspects contribute to the likelihood of respondents to participate in a survey, such as the use of incentives (Singer, 2012), the sponsor of a survey (de Leeuw & De Heer, 2002), its topic (Voogt & Saris, 2005), its length (Galesic & Bosnjak, 2009), and the mode it is carried out in (Groves & Peytcheva, 2008). Overall, the willingness to respond to a survey varies over persons and is affected by the overall survey design (Groves, et al., 2009). Despite noncontact and refusal, the inability to participate increases unit nonresponse, for example, if a respondent's reading or writing skills are limited. In addition, technical problems can contribute to an inability to participate and researchers are challenged to

create Web surveys that run on many different devices (tablets, smartphones, and laptops) and across many different browsers (e.g. Internet Explorer, Firefox, Chrome, Safari, or Opera). In order to accommodate this problem, many Web surveys were often built under the low-tech paradigm to ensure that a survey would run on any device and browser (Couper, 2008a). It is important to point out that the sheer number of people not responding, as well as the response rate, is not necessarily the most relevant factor regarding data quality. As long as nonresponse is the result of pure chance –, in other words, if nonresponse is completely random – then there is no real problem (de Leeuw, Hox, & Dillman, 2008). More important (in respect to data quality) is if those individuals responding to the survey differ in their answers systematically from those who don't. If they do differ, they cause response biases and are thereby detrimental to data quality (e.g. Fuchs, Bossert, & Stukowski, 2013; Groves & Peytcheva, 2008; Stoop, Billiet, & Vehovar, 2009).

It is generally agreed that nonresponse is very important with respect to data quality and that response rates are an important descriptor of the quality of a survey measuring the percentage of eligible sample cases that were measured (AAPOR, 2011; Groves, et al., 2009). In Web surveys, as in other self-administered survey modes, response rates have declined in recent years (Groves, et al., 2009) and tend to be lower in comparison to face-to-face or telephone surveys. In a meta-analysis, Manfreda et al. (2007) compared Web survey response rates to other modes of data collection (mail, telephone, face-to-face) and found that Web surveys overall yield an 11 percent lower response rate. Shih and Fan (2007) compared e-mail and mail-survey response rates and found only minor differences in nonresponse when younger college student populations were surveyed, while overall response rates to Web surveys were, on average, 20 percent lower. There are several ways to compute a response rate for a survey. The AAPOR definitions list six different ways of calculating response rates (AAPOR, 2011, pp. 44-45), which differ predominantly in the treatment of partial interviews and units of unknown eligibility. The response rate provided for the studies in this

thesis is computed via the AAPOR Response Rate 2 (RR2), which represents the number of complete and partial interviews (at least 50 percent of all items answered) divided by the number of invitations delivered to the individuals in the sample.

3.4 Measurement error

While validity refers to whether the questionnaire or survey measures what it intends to, measurement error in the total survey error concept refers to the sources of error relevant while measuring. For that matter, measurement error accommodates a very broadly defined source of error and can be affected by the way a survey is administered, the survey mode, the entire questionnaire, or one single question (Biemer, 2010a; de Leeuw & Hox, 2010; Fuchs, 2008; Groves, et al., 2009). In general terms, measurement error is the deviation of the response from the true value, for whatever reason. Despite unsystematic random errors, systematic errors that create a bias are especially a threat to data quality.

Interviewer characteristics, such as interviewing experience and style, as well as age, gender and race/ethnicity, are the most important causes of interviewer effects reported in the literature (Hox, De Leeuw, & Kreft, 1991; Schwarz & Oyserman, 2001). Respondents frequently tend to agree often with interviewers in order to be polite. This tendency to agree (acquiescence) also appears in self-administered surveys like Web surveys, but at a lower level and especially when respondents are bored or less motivated (see chapter 4 for a detailed view of the question–answer process). Since it is often more comfortable to agree than to disagree, Schuman and Presser (1981) suggest avoiding “agree” or “disagree” answer categories entirely whenever possible. Interviewer effects related to their socio-demographic properties are driven by the extent of social distance between interviewers and respondents in a survey situation, but are of course only relevant if the questionnaire is administered by an interviewer.

Measurement error is further affected by the mode in which a survey is conducted. For example, respondents who are used to working with

computers and the Internet will most likely have no problems in filling out a Web survey, while those who only rarely use those devices might have more problems in self-administering the survey (Fuchs, 2003). In addition, the absence of an interviewer might entice respondents to do other tasks on their computer simultaneously, reducing their attention to survey questions. On the other hand, self-administered survey modes lowers any need for social desirability since potentially displeasing answers don't have to be spoken aloud to an interviewer. For example, respondents rate their health more positively when an interviewer is asking the questions, while they are more likely to admit to taking drugs or other undesirable behaviors in self-administered survey modes (Dillman & Christian, 2005).

The order of the response options and the overall layout and visual design also affect the measurement process (e.g. Sudman, Bradburn, & Schwarz, 2010; Tourangeau, Rips, & Rasinski, 2000). When asking specific questions that respondents know the answers to, context and order effects are highly unlikely, while vague questions encourage respondents to seek additional information from the context (and perhaps prior questions) when interpreting the question (Dillman & Christian, 2005). The most important response order effects influencing measurement error are primacy and recency effects. Primacy describes the tendency of respondents to choose among the first-offered answer categories provided with a survey question. Especially in self-administered surveys (such as Web or paper-pencil), respondents are more likely to choose the first category that seems adequate and then proceed to the next question, ignoring the remaining response options (Smyth, Dillman, Christian, & Stern, 2006). Recency is the tendency to choose one of the last-offered answer categories provided, and this occurs predominantly in interviewer-administered surveys because it is easier for the respondent to remember the last response-options (Krosnick & Alwin, 1987).Visual design aspects such as labels, pictures, the arrangement of answer categories, the use of colors, and the order of answer categories affect responses and likewise measurement error in self-administered surveys (e.g. Christian & Dillman, 2004; Smyth, et al., 2009; Toepoel &

Couper, 2011; Tourangeau, et al., 2007a). A closer look at the influence of visual questionnaire design on the measurement process is provided in chapter 6.

As noted earlier, measurement error is further exacerbated by item-nonresponse. Item-nonresponse occurs when respondents only partially answer a questionnaire, or perhaps leave single items unanswered. Most factors that contribute to unit nonresponse affect item-nonresponse in the measurement process as well. For example, respondents might have difficulties in processing a question and decoding its meaning, or they may not able to retrieve the information needed, or are not able or willing to share the response outcome (Krosnick, et al., 2002). Respondents could feel forced to not answer a question when there is a missing response category or a privacy-threatening question (for reasons of social desirability see also chapter 4), or when a question offers no "don`t know" option that could permit respondents not to express their non-opinion. Overall, motivating respondents, reducing the response burden and providing exhaustive answer categories are the most effective ways to keep item-nonresponse at a minimum. In comparing different designs of open-ended questions in this thesis, item-nonresponse is an important criterion for evaluating which of the tested designs increase the willingness to respond, and reduce measurement error.

This is only a very brief description of the error sources in the measurement process; a closer look at the question–answer process in chapter 4 as well as an exhaustive overview of the influence of visual design in self-administered surveys in chapter 6 will help to better describe which aspects of questions and the questionnaire contribute to measurement errors in Web surveys.

3.5 Processing error

Processing errors appear predominantly in the data entry, coding, and editing processes. In paper-pencil studies, most data entry errors occur while transferring data from the questionnaire into an electronic dataset.

These errors may vary by the types of questions, with numeric questions being easier to transfer in comparison to written answers, where coders might have difficulty reading poor handwriting. Scanning paper-pencil questionnaires using optical scanning recognition or keying the answers into a database are also prone to errors (Fuchs, 2008, p. 900). In using computer-assisted interview modes, data entry errors can be reduced by allowing interviewers to type in only properly formatted answers. In Web surveys the data are entered directly by the respondent, making the transferring process avoidable.

The coding process is error-prone especially when responses to open-ended questions must be transferred to a categorical variable for analysis. Usually a coding scheme defines the categories a coder has to assign responses to. However, different coders come to different judgments about how to classify answers, especially when a bad coding scheme (e.g. categories not exhaustive, not exclusive), poor coding instructions, or inadequate coder training lessen the reliability and affect processing error (Kaczmirek, 2009, p. 26). By using multiple coders for the same set of responses, the percentage of times the coding differed between coders can be displayed as a measure of processing error (Niedomysl & Malmberg, 2009). In a study of coder variability, Kalton and Stowell (1979) found correlations between six professional coders to be relatively low (r =.65 to r =.80), suggesting coding can substantially affect survey error.

Editing data can further affect overall survey error. For example, incorrectly formatted answers are often edited to the desired format, with partial interviews and outlier excluded. Editing a response can always create a threat to data quality, when it is not executed properly. In order to compensate for item-nonresponse, missing data can be imputed and erroneous data replaced by computed values (e.g. Thompson & Washington, 2012). If these new imputed values are not flagged, they appear as legitimate responses and cannot be identified later when analyzing data. Overall errors created by data entry, coding, and editing can be reduced by carefully carrying out each single step

(multiple times if necessary), and by documenting the procedures used when coding responses and editing data.

Adjusting data via weighting is another error source in the editing process. Survey weights are used to try to compensate for everything that might go wrong – for coverage, sampling, and nonresponse errors in particular. But adjusting data is not always possible, nor is it ideal. If results are applied to the general population, weighting can be a powerful tool, but only if enough variables exist in the dataset to apply a proper weight. Most often, demographic variables like gender, age, or education of a general population survey are used to increase the weights of the respondent cases that are underrepresented in said survey. The so-called adjustment error that arises in the processing of data will be discussed in greater detail next.

3.6 Adjustment error

The basic idea behind correcting for biases resulting from coverage, sampling, and nonresponse errors for probability-based as well as for non-probability-based Web surveys is to use auxiliary variables from a high-quality probability-based survey to calibrate another survey. This procedure is used to apply survey results to the general population (or the targeted population) by assigning a value to each case in a dataset that indicates how much the case will count in a statistical analysis. For example, a sample might include 65 percent women, while in the general population they may only make up 50 percent of the population. In such a case, the oversampled women will result in a bias since survey results will reflect the higher “weight” of women in the dataset. Weighting procedures mainly use socio-demographic variables in order to correct differences between Web samples (as well as other samples) and the general population. Since only one weight can be used per case, weights for different factors (e.g. gender, age, education, or region) must be combined into a single weight by multiplying all individual weights into a total weight.

In order to calculate weights, the results of a general population survey (such as the Allbus in Germany or the NHIS in the US) are used as an auxiliary dataset to which the sample data are compared. If the (socio-demographic) data obtained in a survey resemble those of the general population survey, weighting procedures are not required. If differences are present, sample weights should be considered. However, these so-called post-stratification weighting techniques have been criticized in past as not ideal for weighting Web surveys, since Internet users also differ in some of their attitudes from non-Internet users, and not exclusively according to demographics (Bandilla, Bosnjak, & Altdorfer, 2003; Schonlau, et al., 2003). Coverage error is also more problematic for Web surveys (especially in comparison to RDD Telephone surveys) since "Internet use is not yet equally spread among all socioeconomic and demographic groups [so] the coverage problem is likely to lead to biased estimates on variables related to socioeconomic status" (Schonlau, et al., 2009, p. 293). Even with growing Internet penetration rates, sampling and nonresponse error still compromise data quality, especially in non-probability-based Web surveys.

To address this situation and reduce cofounding effects of selection mechanisms arising in Web surveys, propensity weighting approaches were developed that balance the covariates between comparison groups (see d'Agostino, 1998; Rosenbaum & Rubin, 1983 for a review). The propensity score resembles "the conditional probability that a respondent is a Web survey respondent rather than the respondent of a reference survey given observed covariates" (Schonlau, et al., 2003, p. 6). Regarding Web surveys, the propensity score reflects the probability that an individual to have Internet access as opposed to not having access. In order to estimate propensity weights, a Web survey sample is combined with a reference survey sample (e.g. RDD or other probability-based surveys). The reference survey is supposed to have no (or only minor) biases, allowing for adjusting the Web survey. After combining the Web and the reference surveys, the propensity scores of the cases that are part of the Web survey sample are estimated via logistic regression analysis; then the scores are ordered by their size and divided into five

subsamples. For each of these five samples, weights are computed to approximate the distribution of the reference sample (e.g. Lee & Valliant, 2008; Schonlau, et al., 2003).

At first glance, weighting survey data appears to be a compelling notion, but the actual improvements realized can be deceiving. The adjustments made sometimes actually increase biases. And while weighting might increase the accuracy of some variables, it will unlikely work across all variables in a dataset. In general, the initial quality of the survey that is adjusted, and especially the quality of the reference survey, determine whether the adjustment will improve survey results. As an example, Lensvelt-Mulders, Lugtig and Hubregtse (2009) tried to correct for differences between a random and volunteer Internet sample; after adjusting the volunteer self-selected Web survey, the results failed to resemble the results of the probability-based Web survey.

Web surveys that target the general population often require post-survey adjustments. This holds especially true when the sampling process is based on non-probability procedures. Overall, the total survey error framework highlights the sources of error apparent in survey research. Therefore, avoiding these errors in the first place remains the best way to improve survey data. Weighting can help to correct for errors, but only to a degree.

4 ANSWERING SURVEY QUESTIONS

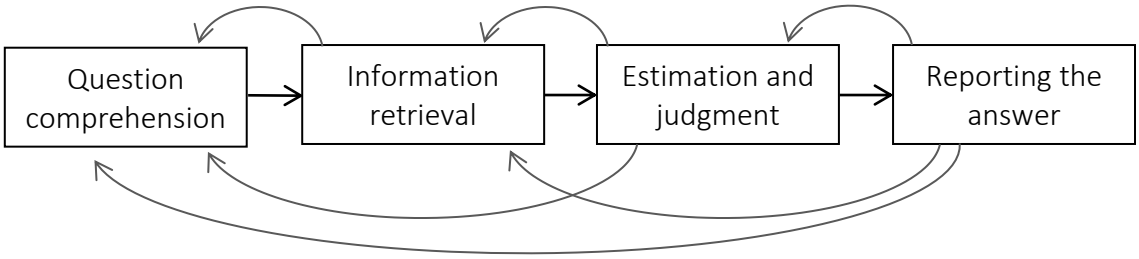
Answering survey questions implies that as a first step a question and the information it carries are administered to the respondent; then the respondent cognitively processes the question and provides an response (Callegaro, 2005). Once the question's information is processed it can be stored, retrieved, and transformed into a response with the human operating in much the same way as a computer (Deutsch & Deutsch, 1963). Essentially a computer has input devices, a central processing unit, and hard disk storage, similarly to the way that individuals have a sensory register, short-term or working memory, and long-term memory. A computer and a human can each report processed results. However, there are certainly several ways in which individuals differ from a computer, and limit this analogy. A computer observes every kind of input while the human brain pays attention only to a small amount of information. The human brain more easily loses information, yet can't delete unpleasant information from its mind. Also, a person's emotions have a strong impact on information processing, in comparison to an entirely emotionless machine. While a computer can only work with the information given, the human brain can attempt to make corrections and fill in gaps. Therefore the computer metaphor has some limitations, but it does help one to better understand how information is processed and what factors contribute to the processes involved when the respondent answers a survey question and provides a response.

4.1 The question–answer process

The completion of a questionnaire requires asking questions in whatever mode is used to present them. Despite the possible difference when presenting a questionnaire, all surveys rely heavily on respondents interpreting a pre-established set of questions to supply the information these questions seek (Groves, et al., 2009). The underlying process respondents undergo when answering a survey question can be rendered as four essential steps (e.g. Biemer & Lyberg, 2003; Groves, et al., 2009; Presser, et al., 2004). Tourangeau's model (1988) is the most

common in the literature, classifying the question–answer process in the following four steps: First, respondents interpret the question and decode its meaning when comprehending the question. Second, they retrieve all the information necessary to answer the question. Third, the information recalled is summarized and combined. Finally, the answer is formulated and put in the required answer format.

Figure 3. Model of the four-step survey response process (Groves, et al., 2009, p. 218)



Ideally, respondents will process a survey question as described above, but at the same time it is highly unlikely that they will complete every single step in the process in a linear sequence. Often, respondents skip some of the steps, jumping back in the process of generating an answer, or even stop at some point in the process and report no answer at all.

4.1.1 Question comprehension

In order to answer a question, respondents first have to interpret and decode its meaning to understand what they are actually being asked. Ideally the respondent will interpret the question in the same way the researcher does (Sudman, et al., 2010). Comprehension starts with a string of words separated into syntactical units that are understood, which means that the meaning encoded in the linguistic units is extracted by a process that is still poorly comprehended (Bradburn, 2004). Respondents parse the question, its components, and their relation to each other in order to assign meaning to the substantive components of the survey question. In Web surveys, where questions are presented via written words, word recognition does not depend heavily on the context

(in comparison to surveys where the question is transmitted vocally). However, using the same words for every single respondent will not necessarily result in identical representation and comprehension of that question.

Comprehension problems occur predominantly when words are ambiguous, have multiple meanings, or are used in different ways. For instance, when teachers were asked how many children they have, some included the children in their classes, even though the initial intention of the question was to evaluate the number of children in their families (Bradburn, 2004). The context in which a question is asked can be used to resolve ambiguity and help to better understand the question. But at the same time, context can influence question comprehension in an unintended way. As the previous example about the teachers illustrates, when asked in a school context, teachers will likely include children in their classes in their answer. When the same question is asked in a personal context, teachers in all likelihood will only include children in their households. If ambiguous words are included in a question, the interpretation that comes to mind most easily is used. Until an interpretation is proven wrong, respondents proceed with the most accessible one and do not process multiple interpretations simultaneously (Sudman, et al., 2010). Even though respondents encode the literal meaning of a question, they also have to infer the researcher's intention. Therefore, respondents not only rely on the question's wording, but they also use any information that comes with a question and consider presented response alternatives, the context in which a question is asked, and their own previously provided answers (Sudman, et al., 2010).

Studies on visual questionnaire design (see chapter 6) revealed that the response options provided with a question have a strong influence on how respondents interpret and answer one and the same question. For instance, high- or low-frequency answer scales, adding numbers to answer categories, or the use of smileys instead of text-labeled response options all affect responses (e.g. Emde & Fuchs, 2012; Knäuper, Schwarz, Park, & Fritsch, 2007; Tourangeau, Couper, & Conrad, 2004). The

positioning of answer elements, the placing of instructions, and their visual appearance also have an impact on responses (e.g. Christian & Dillman, 2004; Israel, 2010; Redline & Dillman, 2002; Toepoel & Dillman, 2010). Altogether these results imply that respondents infer different meanings based on the additional information (or cues) provided via response options and their visual appearance. The vaguer a question is, the more respondents rely on the answer categories and other information to determine the question's meaning. Of course, for open-ended questions, no additional information can be drawn from the response options since there are none supplied. Unambiguous wording and easy understandable questions are therefore essential when asking open-ended questions.

Despite the question asked and the answer options provided, respondents also look at preceding questions and for further additional cues available from the whole questionnaire and its context. This is true especially in self-administered surveys where respondents can easily skip to preceding and subsequent questions to gather additional information from the entire context. Overall, question comprehension is a broad process including all aspects of a question and the questionnaire. In order to assign meaning to a question and infer its purpose, respondents may go beyond the information provided by the question wording and use whatever information and cues they can.

4.1.2 Information retrieval

Once a question is understood, respondents have to recall information required from their memory. Long-term memory has a vast capacity, storing autobiographical memories as well as general knowledge (Groves, et al., 2009). Overall, more information is stored in human memory than actually used in any given moment, and only information that has been encoded and stored in long-term memory can be retrieved for later use. Some events or attitudes cannot be properly retrieved because they have never been transferred or encoded to the long-term memory (Callegaro, 2005). Therefore, survey results can only be as accurate as the memories of the respondents. Retrieval cues that prompt memory can help trigger

the recall of information. In the retrieval process, the cues available in a given question activate pathways of association leading to the desired information (Bradburn, 2004). While sometimes the answer can be drawn directly from long-term memory, in many cases only some of the required information is present in the memory and the cues retrieved will only help to lead to an answer. Once the retrieval process begins, activation spreads automatically from one concept or memory to related other ones, connecting them in order to become additional cues guiding the recall process (Tourangeau, et al., 2000). Retrieval cues such as words, images, or emotions activate and direct the information-seeking process. This cycle of generating cues and retrieving information continues until the respondent finds the necessary information or gives up (Groves, et al., 2009).

While respondents rely on preceding and subsequent questions when decoding a question's meaning, these context effects must be attributed to the retrieval rather than the comprehension stage. Several studies indicate that prior questions activate concepts or feelings that are, once activated, carried over to the following question because they are easier to access (Tourangeau & Rasinski, 1988). In a paper-pencil survey, Strack, Martin and Schwarz (1988) asked respondents for their overall life-satisfaction, and before or after they asked them how often they went out on a date. In a third control condition, the two questions were separated by a more general question. When the specific question about dating was asked prior to the general life-satisfaction question, both answers were highly correlated, while in the opposite order this correlation was reduced. When both questions were asked separately, the correlation diminished further, suggesting that the specific item had a strong influence on the vaguer general life-satisfaction item, a slightly weaker influence in the other direction, and no influence if both questions were separated. The authors assume that the dating question made dating more accessible in the memory, priming the answer to the subsequent life-satisfaction question. The experiments reported in chapter 6 on anchor effects further exemplify how accessible information affects responses, even if this information is irrelevant. As a result, recall

tasks provide two distinct sources of information: the content that comes to mind, and how easily it comes to mind (Strack, et al., 1988).

Altogether, inference and knowledge influence each other mutually. The success in retrieving relevant information to answer a survey question depends on further aspects. Some events are harder to remember than others; events that are less distinctive, numerous, took place long ago, or offer fewer or unelaborated cues require more time to retrieve. On the other hand, important and emotionally-involving events are easier to recall. Multiple and rich cues help retrieval further, as well as do backwards search and longer time frames in which retrieval can be carried out (Tourangeau & Rasinski, 1988). If and how respondents retrieve (desired) information further depends on their motivation and ability to process the information. Instead of retrieving all possible information, respondents will truncate the search process as soon as enough information has come to mind (Sudman, et al., 2010). Overall, the more complex and demanding a question and the required retrieval are, the less accurate it will be (Groves, et al., 2009; Tourangeau, et al., 2000).

4.1.3 Estimation and judgment

As a next step, respondents had to use the retrieved information to render a judgment. Based on the information previously recalled, respondents filled in gaps in what was recalled, combined products or adjusted omissions in retrieval (Groves, et al., 2009). Therefore, they used all information provided by the question, activated by the question's cues and context (Bradburn, 2004). When respondents were asked for their date of birth or gender, because this kind of information is generally easy to access, they will retrieve the answer directly from memory. But even when information is pre-stored, easy to access and no judgment is necessary, respondents still need to determine if their answer matches the response categories provided or the reference period specified (Sudman, et al., 2010). Especially when questions are more complex and the retrieval process more demanding, such as when respondents were asked for certain behaviors or attitudes, they have to

build answers rapidly from all information accessible and present in their short-term memory. This construction process uses permanently available information as well as temporarily accessible information from the question, the context, or any other source as the respondent answers a question (Bradburn, 2004).

Attitude questions a more complex to answer and the judgment process involves scaling of beliefs, an assessment of their importance into an overall judgment unless the respondent has a well-defined attitude or no attitude (Callegaro, 2005). The same appears true for many open-ended questions. When asking for the most important problem facing the nation today respondents commonly omit problems that are important but not salient like nuclear war for example since they are not accessible when forming the answer (Jabine, Straf, Tanur, & Tourangeau, 1984). Frequency and probability judgments are also difficult to form and respondents rely heavily on the answer scales if they are provided with question.

Since respondents rather truncate the information retrieval as soon as enough information has come to their mind respondents often use the information that comes to mind most easily. Other less accessible information is unlikely to be considered (Sudman, et al., 2010). Overall the accessibility and availability of concepts can result in more or less differential interpretations and evaluations because only this information is used to form an answer. Therefore the context in which it is build has an important influence on the estimation process and the response. Information used to answer preceding questions comes easier to mind and can priming following responses. When forming a judgment responses can further be affected by assimilation and contrast effects. Not relevant but accessible information included in the judgment process can influence responses into the direction of this information. When Schwarz et al. (1991) asked respondents about their martial happiness first and for their life satisfaction afterwards, answers to the second question were affected by the preceding question (similar to the experiment of Strack et al. (1988) in the prior Chapter). Adding unrelated questions between the question on martial and life happiness reduced

this influence as well as asking for the more vague life happiness question first. Despite including irrelevant information that might lead to assimilation effects respondents could also exclude temporary representations that will result in contrast effects. As a result excluding positive information will have a negative influence on the response and vice versa. Mood and emotions are also affecting the memory and judgment process in that way that respondents in a good mood tend to report more positive judgments especially when they are unaware of their feelings or when those feelings are result from the theme of the survey they are responding to (Banaji, Blair, Schwarz, & Sudman, 1995).

In summary, responses are based on accessible information that form a mental representation of the attitude or behavior that is used by the respondents to come to an answer. These representations can be chronically accessible in the memory leading to stable reports whenever a respondent is asked that question, while others are only temporarily accessible and therefore contribute to the instability of reports over time (Sudman, et al., 2010).

4.1.4 Reporting the answer

Once a judgment is formed, respondents have to fit it into the response options provided for the question. Closed-ended questions provide a limited number of response options. If the response formed does not fit in the options provided, respondents might go back in the question–answer process in order to form an adequate response. Even though the question–answer process appears to be linear, respondents often skip back and forth between different stages of the question–answer process. Response options affect every single step of the question–answer process: they can guide as well as influence the respondent (Sudman, et al., 2010). If the response options provided do not accommodate the prior formed answer, respondents might choose the item closest to their initial judgment that seems to be satisfactory; they might eliminate response options to narrow down the number of possible responses, or employ ranges or round values to report numeric quantities (Callegaro,

2005). Respondents might even skip a whole question or provide a random response if a question is too burdensome.

Open-ended questions do not restrict respondents and there is no need for them to adapt their answers to a provided set of response options. On the other hand, answering open-ended question require respondents to formulate a response in their own words, making the actual reporting a much more burdensome task in comparison to just checking a closed-ended response option. Therefore, respondents will more likely leave open-ended questions unanswered even though they might have an answer. Overall the question–answer process is affected by respondents' abilities and the difficulty of the response task, which is why we will take a closer look at information processing in the following chapter.

But even when a judgment is formed and an adequate response option is provided by the researcher, respondents might edit their responses in order to portray themselves more admirably, especially on questions about socially desirable or undesirable matters (Krosnick & Presser, 2010). Questions on sensitive topics such as sexual behavior, drug and alcohol abuse, criminal offences and attitudes towards politics, abortion, and suicide, amongst others, are prone to social desirability biases (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). Some behaviors are over-reported, such as voting or church attendance, while others are frequently underreported, such as racist attitudes or drug use (Tourangeau, et al., 2000).

The extent to which topics are perceived as sensitive varies between societies, ethnic groups, and age cohorts. The editing of a response therefore depends on situational factors such as the presence of an interviewer, the topic of a question, and the respondents' attitudes (Tourangeau, et al., 2000). The perceived anonymity of a survey and the social distance between a respondent and an interviewer influence the occurrence and level of social desirable responses (Fuchs, 2009c). In face-to-face surveys, the perceived anonymity is lower in comparison to telephone interviews, making socially desirable responses more likely. If the characteristics like sex, age or education of an interviewer a similar to

those of a respondent the social distance between both is lower and the appearance of social desirability responses less likely. Confidentiality endorsements, forgiving wording, the placement of sensitive questions in contexts that promote accurate answers can help reducing the threatening character of sensitive questions (Tourangeau, et al., 2000). In addition randomized response techniques can be useful. Instead of asking directly a sensitive question (e.g. have you done drugs in the last 6 month), the respondent is than asked randomly the sensitive question or he is just instructed to answer affirmatively. Only the respondent knows to which of these two instructions (question vs. affirm) his response refers to. But based on the proportions it is later possible for the researcher to compute a response (for an overview on randomized response techniques see Chaudhuri, 2011; Lensvelt-Mulders, et al., 2005).

However the involvement of any interviewer makes sensitive questions more threatening for respondents. Therefore self-administered surveys like paper-pencil or Web surveys are less prone to social desirability biases. But even without an interviewer involvement questions that are sensitive and stressful for the respondents will still be subject of social desirable responses.

4.2 Cognitive information processing

The question–answer process as described before names the processing of information as one of the most important and crucial aspects when it comes to forming a survey response. Answering a single question requires a lot of cognitive work and confronting respondents with a whole questionnaire enlarges this cognitive processing task even more (Krosnick, 1999). At the interpreting stage respondents have to build a mental representation of the question, search their memory afterwards for information to estimate a judgment and translate it into a response. Anything that goes wrong within this process will lead to measurement errors and consequently affect data quality and survey results (Bethlehem & Biffignandi, 2012). And while answering a single

thoughtfully written question might not be too challenging for most respondents, their motivation will decline when answering an entire questionnaire. Instead of carefully retrieving all information possible, respondents truncate the search process and use shortcuts (Sudman, et al., 2010). In addition, they are distracted, especially in a Web survey, since they can easily multitask, read e-mails, or surf the Web while answering a questionnaire in their browser (Heerwegh & Loosveldt, 2008). Generally speaking, the more complex and demanding the question–answer process appears for a respondent, the less accurate it will be (Groves, et al., 2009; Tourangeau, et al., 2000).

The computer analogy described at the beginning of this chapter implied a very rational approach by individuals when processing information. Even though the question–answer process appears to be shaped that way, its underlying processes are not. For example, Schwarz and Clore (1983) have found that judgments of life satisfaction can be influenced by one's mood. As a result, respondents rated their life satisfaction higher on sunny days than on rainy days (Schwarz & Clore, 1983). So how do individuals actually process the information needed to come to an answer?

Human information processing is complex and over the years several theories have tried to condense the processes involved into a theoretical framework. Two-Process theories like the Heuristics-Systematic Model (HSM) by Chaiken and Eagly (1989) as well as the Elaboration-Likelihood Model (ELM) by Petty and Cacioppo (1986) postulate two different routes of processing information when forming an attitude: a systematic (HSM) respectively central route (ELM), as well as a heuristic (HSM) or peripheral route (ELM). Both models note, even though different in some details (for an overview see Strack 1999 and Chen & Chaiken, 1999), that persuasion is accomplished by two essentially dissimilar routes. Heuristic (as well as peripheral) processing is described as a limited mode of processing that demands much less cognitive effort (Chen & Chaiken, 1999). Individuals then only use easy available information and simple inferential rules or heuristics to come to a decision. Since heuristics are pre-stored and easily accessible, their use is less demanding if those

heuristics are applicable in the respective context. On the other hand, in the systematic (or central) route of processing, opinions and attitudes are formed by carefully processing information, ideas, and relevant arguments. But central processing will only occur when individuals have the motivation and ability to think about an addressed message and its topic. Lack of motivation or distraction will force people to process information rather than use heuristics. In addition, cognitive resources are always limited and individuals always tend to execute tasks by the principle of least effort in order to balance cognitive effort and "satisfying their motivational concerns" (Chen & Chaiken, 1999, p. 74). Overall, the two-process models state that it comes down to three basic aspects when processing information: abilities, motivation, and the difficulty of the processing task.

However, the two routes of processing drawn by Chaiken, Eagly, Petty, Cacioppo, and their colleagues suggest that individuals take either the peripheral or the central route. This rather static model assumption raised concerns that the dynamics in information processing should be captured by a model that allows for the flexibility that "characterizes our mental functioning" (Strack, 1999, p. 169). In their Unimodel, Kruglanski (1989) integrated the two routes of persuasion into one, since in their opinion, the two routes do not fundamentally differ. The authors argue that heuristics and arguments can be both complex and simple, and that both are processed in the very same way with the same mechanisms involved.

This is only a rudimentary description of these models of persuasion. The complexity of provided cues as well as the motivation and ability to process information are the factors involved when individuals decide which information is processed and how it is processed. Even though these models were not specifically designed to apply to the question–answer process, cognition is in most terms evaluative and social objects are almost impossible to view "without at the same time making an assessment on dimensions closely corresponding to good/bad, pleasant/unpleasant, high/low, etc" (Markus & Zajonc, 1985, p. 210).

In his theory of bounded rationality, Simon (1957) developed the idea that individuals in a decision making process satisfice rather than optimize. As in the two-process models, Simon noted that individuals accept a "good-enough" approach because it is almost impossible for everyone to inventory all possible information. Instead, people search their memory until they come to a solution that satisfies their needs. Krosnick (1991, 1999) borrowed Simon's satisficing terminology in order to explain how survey respondents actually answer questions. However, optimal answers can be expected when respondents execute all four stages of the question–answer process carefully. Satisficing occurs when at least one of the four steps is in the question–answer process is compromised (Krosnick, Narayan, & Smith, 1996). Krosnick further describes different levels of satisficing. Weak satisficing arises when all four stages of the question–answer process are processed superficially. Respondents might "be less thoughtful about a question's meaning, they may search their memory less thoroughly, they might integrate retrieved information more carelessly, and/or they may select a response option more haphazardly" (Krosnick, et al., 1996, p. 31). Selecting the first response option acceptable in a list of closed-ended response options or using a "don't know" or "no opinion" option when offered, simply because it is easy and less demanding, are examples of weak satisficing. Omitting entire steps of the question–answer process and the selection of a response option based on easily accessible cues from the question itself are what Krosnick describes as strong satisficing. Overall, some sort of satisficing will likely appear in most surveys and from many respondents. Krosnick advances his view, that optimizing and strong satisficing should be seen as the "two ends of a continuum the degrees of thoroughness with which the four response steps are performed. The optimizing end of the continuum involves complete and effortful execution of all four steps. The strong satisficing end involves little effort in the interpretation and answer reporting steps and no retrieval or integration at all. In between are intermediate levels" (Krosnick & Presser, 2010, p. 266).

In general, three basic factors contribute to the likelihood of satisficing (Krosnick, et al., 1996). First, the greater difficulty of the question, the more challenging the processing is for respondents. Secondly, respondents' motivation influences how and even if the information provided via the question is processed. Third, the ability to process is important when formulating an answer. The same factors that influence information processing in the two-process models also contribute to the strength of satisficing. Therefore, we will take a closer look at these factors and how they contribute to data quality in surveys.

4.2.1 Question difficulty

The question–answer process as described earlier made clear that a great deal of cognitive work is required from the respondent to come to an answer. Difficulties in question comprehension can appear if words used are unfamiliar, vague, ambiguous, or if their arrangement is complex and hard to understand from a semantic perspective. Information retrieval can be difficult if an event was not very distinctive, or events were numerous, took place long ago, or offer few or unelaborated cues. Conversely, important and emotionally involving events are easier to recall since they provide more and richer cues to help retrieval (Tourangeau & Rasinski, 1988). A question about the number of jobs a respondent had during the last three years might be much easier to answer, compared to asking for the number of doctor visits during the same time frame. More frequent events (such as visiting a doctor) should be asked across shorter time frames, such as "during the last six months". Van Der Vaart and Glasner (2007) found more recall errors by respondents when the tasks were less salient or less recent. Further, subjects tend to overestimate the frequency of common events (Schwarz & Hippler, 1987) and are prone to scale effects if events asked for are not regular (Menon, 1994).

When rendering their judgment, respondents use the cues provided via the response options. For instance, answer scales are easier to use if they are fully labeled rather than just numerically numbered, and if the words used are familiar, unambiguous, and easy to understand. Open-ended

questions can be particularly difficult to answer since the lack of response options allows for no additional cues. Providing a response scale would result in fewer non-substantive responses ("don't know") in comparison to an open-response format (Knäuper, Belli, Hill, & Herzog, 1997; Schwarz, 1990). The survey mode also affects the difficulty of questions. In self-administered Web surveys, difficult questions can be even more challenging since no additional information or probing is provided via an interviewer (Conrad & Schober, 2000; Couper, et al., 2011). In order to compensate for the absence of an interviewer, clarifying features like instructions, definitions, or examples can help to facilitate the respondents' task (Kunz & Fuchs, 2012). And while in self-administered surveys the respondents can take their time to process a question, in face-to-face or telephone surveys the interviewer dictates the pace of an interview.

Task difficulty also depends on the respondent's level of distraction while answering a question, such as when other people are present while a respondent is answering survey questions, or when respondents simultaneously browse the Internet while answering a Web survey. However, the best way to reduce question difficulty is to provide solid and well-designed questions. A variety of guidelines for writing are available in the literature (e.g. Fowler, 1995; Schuman & Presser, 1981; Sudman & Bradburn, 1982) and there is conventional wisdom that the words used should be specific, concrete, simple and easy to understand, and response categories exhaustive and exclusive. Ambiguous words or meanings should be avoided, as well as double negatives, and loaded questions that suggest a certain response.

4.2.2 Respondents' motivation

If a question is easy to understand, respondents may have no problems in providing a proper answer. However, if the question is involved and complex, less motivated respondents won't put much effort into their answers, especially when they have to look forward to several more upcoming questions. Therefore, the length of a questionnaire influences respondents' motivation. Galesic & Bosnjak (2009) found lower

participation rates when they announced that a survey wouldtake 30 minitus to complete, in comparison to one that would take 10 minutes. The authors further discovered that burdensome matrix questions were answered faster when positioned at the end of a questionnaire, and responses to open-ended questions were longer when positioned at the beginning.

Oudejans and Christian (2010) found respondents more likely to respond to narrative open-ended questions when they were interested in the topic of the questionnaire. In addition, individuals who found the survey interesting delivered longer responses. However, the authors also discovered that the quality of responses to narrative open-ended questions declines as respondents progress through the survey, suggesting a decrease in motivation along the process of answering a questionnaire (Oudejans & Christian, 2010). Open-ended questions are especially difficult and burdensome to answer, which is why motivating respondents can be essential. Smyth, et al. (2009) used verbal prompts to improve response quality. Simply highlighting the importance of a narrative open-ended question improved the response quality, especially for late respondents who tend to be less motivated. Israel (2013) used the same motivational statement in two narrative open-ended questions and found the statement to be effective at increasing the percentage of respondents who answered the question. But citing the importance of a question might only work once or twice, since the effect of such a prompt will wear out.

In general, respondents' motivation to answer survey questions is driven by "desires for self-expression, interpersonal response, intellectual challenge, self-understanding, feelings of altruism, or emotional catharsis" (Krosnick, 1999, p. 547). Additionally, the interest in a specific topic or in surveys in general affects respondents' willingness to participate (Groves, Presser, & Dipko, 2004).

The most basic differentiation can be made between intrinsic and extrinsic motivation. While intrinsic motivation in answering survey questions is based on implicit factors, extrinsic motivation is always tied

to some kind of reward (Stiglbauer, Gnambs, & Gamsjäger, 2011). Providing incentives (e.g. lotteries, vouchers, cash prizes) can help to improve the motivation to participate in a survey (Bosnjak & Tuten, 2002; Frick, Bächtiger, & Reips, 2001; Singer, 2012), but it questionable whether these rewards really help the respondents' motivation in the question–answer process. Sending out reminders, pointing out the value of a study and what the answers are needed for are other examples of extrinsic motivators. To strengthen intrinsic motivation, questionnaires should be designed to provide a pleasant and positive experience, suggesting to respondents that their participation is appreciated.

Individuals differ in their need for cognition (Krosnick, 1991; Petty & Cacioppo, 1986). Respondents with a high need for cognition are more eager to think and enjoy engaging in cognitive tasks more than others (Toepoel, Vis, Das, & van Soest, 2009). Generally speaking, individuals with a high need for cognition will be less likely to satisfy, since their motivation for cognitive processing of information is superior to those who are lower in their need for cognition. Despite a person's need to think,(Toepoel, Vis, et al., 2009) further point out that a respondent's need to evaluate can affect motivation in the same way since people differ in their willingness to be involved into evaluation processes. The need for cognition and evaluation both affect individuals' motivation to process survey questions; both are personal traits anticipating the influence of personal characteristics on information processing, described in the next section.

4.2.3 Personal characteristics and the ability to process information

The motivation to respond varies among respondents, as does their ability to process information. Respondents' personalities (and personality traits) might also affect their response behavior. Despite the need for cognition and evaluation and their influence on motivation, respondents also vary in their reasonable accuracy to carry out tasks properly (Goldberg, 1990). In addition, individual differences in education and cognitive abilities will lead to differences in the ability to answer

more burdensome survey questions. Research on survey methodology (such as the studies presented in this thesis) is often collected using highly-educated student samples rather than surveys of the general population. That may also be the reason why sufficient studies on the influence of personal characteristics are rare. Tourangeau et al. (2007a) varied the layout of response scales and found no differences in the response behavior between sex, age, and education. On the contrary, Krosnick and Alwin (1987) found less-educated respondents and their limited vocabularies to be more prone to response-order effects in answer scales.

Knäuper et al. (2004, p. 94) found increasing and decreasing scale effects for older as compared to younger respondents. The authors argue that this conflicting result is due to varying motivation and memory performance. Addressing topics that are important to older people, such as health-related questions, will compensate for deficient cognitive abilities, while low motivation and abilities add up to scale effects as a sign of weak satisficing for older respondents. Further, respondents with lower cognitive abilities were more affected when questions were ambiguous, retrospective, or asked for quantitative reports, and provided more frequent "don't know" responses (Knäuper et al., 2004). When surveying children, Fuchs (2005) found a prevalence of response order effects as well as scale effects associated with response categories decreasing with the children's age and their cognitive abilities.

Open-ended questions are especially affected by personal characteristics and respondents' abilities to process information. Smyth et al. (2012), for example, found higher item-nonresponse rates among low literacy respondents for any question type, suggesting the influence of personal abilities on the question–answer process. Respondents low in literacy also made more errors in skipping instructions, provided range rather than absolute responses, and wrote shorter answers to narrative open-ended questions. In a Web survey, Denscombe (2008) discovered that women provided longer responses to open-ended questions than men. Stern, et al. (2007) found responses to narrative open-ended questions to be shorter when respondents were male or did not hold a college

degree. Oudejans and Christian (2010) found no gender differences regarding the likelihood of answering narrative open-ended question, but responses provided by women were longer. More highly-educated respondents were more likely to respond to narrative open-ended questions using more characters in elaborating on their answer. In addition, Scholz and Zuell (2012) found higher item-nonresponse to open-ended questions when respondents had lesser cognitive abilities or low interest in the question's topic, while Geer (1988) found only minor differences between cognitive abilities and response rates.

Overall, the processes involved when answering survey questions are complex, with answers usually obtained by respondents in a matter of seconds (Callegaro, 2005). The success of a questionnaire, or even a single question, depends heavily on the respondents' ability and motivation as well as the question itself. Any efforts to improve questionnaires should therefore reflect at least one of these three aspects. How different types of questions affect responses is discussed in greater detail in the next chapter.

5 QUESTIONS AND RESPONSE FORMATS: CLOSED- VS. OPEN-ENDED QUESTIONS

When setting up a questionnaire, researcher can choose among various response-formats and questions types. Questions can be differentiated according to their function in a questionnaire. Opening questions, for example, may be easy to understand and enjoyable to answer, and are designed to get respondents involved in the survey and its topic. Filter questions determine whether a respondent is qualified to answer the next question, or has to skip to a later one. Buffer questions help to connect different topics in a survey, and final or farewell questions conclude a questionnaire.

Question types can also be differentiated by the content they seek. For example, questions can ask for behaviors and respondents' factual circumstances, for attitudes concerning the opinions of a respondent, or facts or the knowledge a respondent possesses. Behavioral questions ask about personal characteristics, things people have experienced or done, which could also be observed by a third party or an external observer (Sudman & Bradburn, 1982). Knowledge questions seek respondents' knowledge use of their cognitive abilities on a specific topic. Further, knowledge questions "are often combined with attitude or behavior questions to gauge the saliency of an issue or the outcome of a program" (Sudman & Bradburn, 1982, p 28). Since attitudes are not available to an external observer, questions are the only way to assess psychological states and gain an individual's opinions.

Despite the content sought and the function a question has in a questionnaire, one major differentiation can be made between closed-ended questions, which require people to choose among a set of provided response-options, and open-ended questions, which respondents answer in their own words (Krosnick, 1991). Somehow the term "open-ended question" is misleading, since "it is really the answers that are left open or closed" (Couper, 2008a).

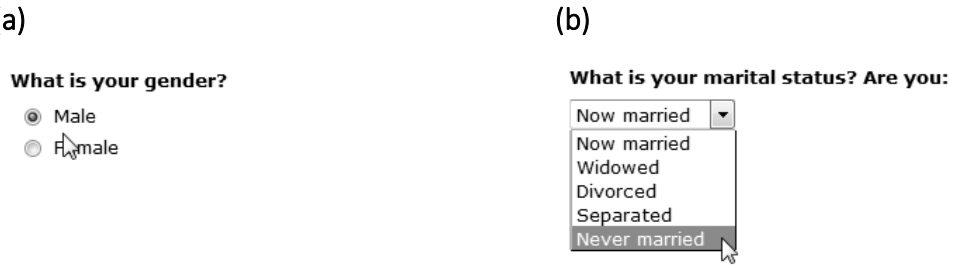
The purpose of this thesis is to find ways to optimize and improve narrative open-ended questions in Web surveys. First we will take a look

at closed-ended questions, and then focus on open-ended questions in greater detail.

5.1 Types of closed-ended questions

Closed-ended ended question provide respondents with a set of response options. These options help respondents decode a question's meaning and provide additional guidance on how to answer a question (as described in chapter 4). In order to preserve their purpose, answer categories have to be exhaustive, mutually exclusive, and easy to understand. However, the answer options provided for a closed-ended question vary based on the question's purpose. In Web surveys, researchers generally make use of three HTML input forms to display closed-ended questions: radio buttons, drop-down boxes, and check-boxes. By using active scripting such as JavaScript or Flash, additional input types can be made available, like slider or visual analog scales, drag-and-drop designs, pictures or smilies as response options, and much else besides. And while at first glance Web surveys mimic paper questionnaires, the input tools in online-administered questionnaires serve as both a visual guide and a forcing function to gain responses (Couper, 2008b). Radio buttons, for example, allow respondents to select just one of a limited number of response options. They are used for simple closed-ended Yes/No questions, or for example, when asking respondents for their gender by providing the answer categories "male" and "female". Of course, the list of response options can be expanded and also displayed by using a drop-down menu, constraining the selection of one and only one response option (Figure 4).

Figure 4. Single-choice question with (a) radio buttons and (b) drop-down menus



Once selected, respondents cannot unselect a radio button, but they can change their response to another response option. Radio buttons can also be arranged horizontally as a rating scale, using labels from Agree to Disagree, Frequently to Never, Important to Unimportant, Good to Poor, Likely to Unlikely, and so on. The number of scale points can be varied, and made even or uneven to provide or omit a midpoint in the scale. Further, scales can be bipolar, reflecting two opposite dimensions with a clear conceptual midpoint, for example, when asking for attitudes. Unipolar scales, in contrast, reflect varying levels of the same dimension, with a zero point on one end, and can be used when asking for the level of importance, for instance. Providing a midpoint allows respondents with a moderate opinion to assign their response, and respondents are not forced to randomly select points closest to the middle, the way they are when a midpoint is omitted. The results of studies on using or omitting midpoints are mixed (Christian & Dillman, 2004; Toepoel & Dillman, 2010; Tourangeau, et al., 2004).

So there is no overall rule regarding the use of a midpoint, but as a researcher it is important to ensure that when it is legitimate for respondents to hold a neutral opinion towards an attitude, or when they don't know the answer to factual questions, they can express their response by a neutral scale point or an additional "don't know" response option.

In addition, scales can be fully labeled, or only the endpoints labeled in order to save space in online surveys and reduce the amount of text respondents have to read. Scales can also make use of numbers instead of labels, or combine both (Schwarz, Knäuper, Hippler, Noelle-Neumann,

& Clark, 1991). Over the years several studies have tried to find the ideal answer scale design and there is agreement that bipolar scales should use from 7–9 balanced scale points that are fully labeled, while unipolar scales should be realized by using 5–7 labeled scale points (for an overview see Krosnick & Presser, 2010). Overall, the number of scale points affects the way respondents can differentiate and is primarily limited by the combination of possible label options. Further decoding and interpreting the meaning of provided response options is easier for respondents when the scale is fully labeled, as compared to a scale using numbers (Tourangeau, Couper, & Conrad, 2007b).

Figure 5. Types of answer scales: (a) fully labeled 5-point rating scale,(b) numeric 7-point scale,(c) slider scale, (d) Visual analogue scale, (e) matrix rating scale

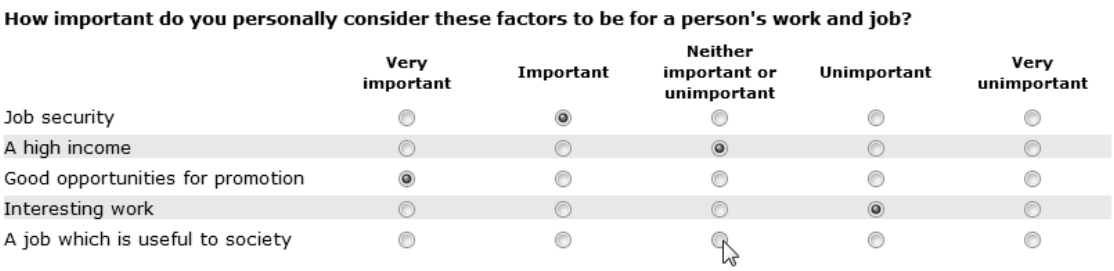
(a)

(b)

(c)

(d)

(e)



How important do you personally consider these factors to be for a person's work and job?

	Very important	Important	Neither important or unimportant	Unimportant	Very unimportant
Job security	○	◉	○	○	○
A high income	○	○	◉	○	○
Good opportunities for promotion	◉	○	○	○	○
Interesting work	○	○	○	◉	○
A job which is useful to society	○	○	○	○	○

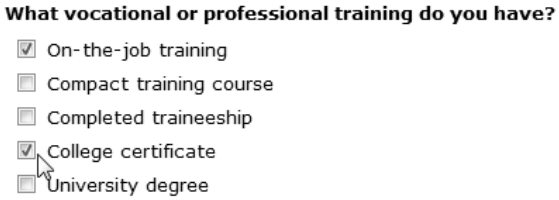
Multiple scale questions can also be combined into a matrix (or so-called grid) question design to save further space and ask respondents several questions in a short time frame, using the same response options for several items. However, encouraging respondents to move quickly through a questionnaire might also result in higher levels of non-differentiation, item-nonresponse, and lower data quality (Toepoel, Das, & Van Soest, 2009). Grid questions also require respondents to match information in rows (items) to the columns (response options), a process that can be error-prone and complex, as well as comply with the instructions on how to answer such a question (Dillman, et al., 2008). Since several items are arranged in one grid, respondents also interpret the items as related, perhaps making one item per screen a better choice in comparison to a grid question (Couper, 2008a).

Instead of using a set of radio buttons, scales can be displayed as slider bars or visual analog scales. These graphical rating scales are incorporated in Web surveys using Java, JavaScript, or Flash, and are displayed in Figure 5. Slider bars allow respondents to simply slide a moveable handle with the mouse to their desired response. Visual analog scales, however, are displayed by a horizontal line between two labeled endpoints. Instead of sliding a handle, the visual analog scales require respondents to click on the line to indicate their response (Funke, 2011; Funke, Reips, & Thomas, 2011). Even though these scales allow respondents to graduate their responses finely, a discrete rating scale using a series of radio buttons obtains very similar results, making slider scales less compelling for Web surveys (Couper, Tourangeau, Conrad, & Singer, 2006).

All types of closed-ended questions described above are single-choice questions that rely on a set of radio buttons to provide a response (except slider and visual analog scales). Unlike radio buttons, check boxes can be checked and unchecked, and respondents can check none, a selection, or all response options, making checkboxes the ideal form element for “choose all that apply” questions (Couper, 2008a).

Figure 6. Multiple choice questions with (a) checkboxes or (b) radio buttons

(a)

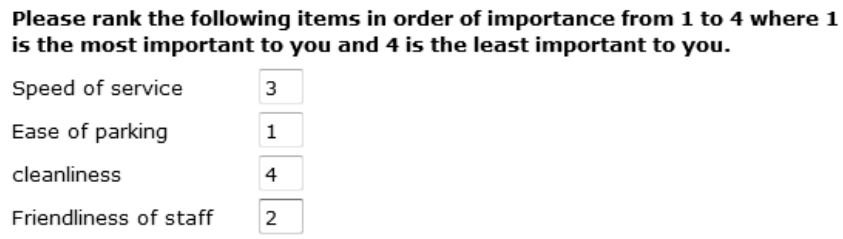


(b)

Check boxes can be arranged in similar ways to radio buttons, but the problem of their functioning is that nobody can tell if a check box was left unchecked by a respondent deliberately or for another reason. Therefore, forcing respondents to answer an "all that apply" question by using a matrix question layout with the response options Yes/No (see Figure 6) might be superior, even though it will take respondents more time to answer (Dillman, 2007; Sudman & Bradburn, 1982). Rather uncommon is the use of drop-down boxes for "all that apply" questions, where respondents can check multiple items from a list within the drop-down menu.

Finally, closed-ended questions can also be designed as ranking (rather than rating) questions, where respondents are asked to order a list of different objects (response options) according to their importance. The example design in Figure 7 is a very basic one using input boxes, while in Web surveys often interactive drag-and-drop designs are used to order objects. However, ranking tasks that are complex to administer take respondents longer to answer and have higher drop-outs in comparison to rating scales (Neubarth, 2006).

Figure 7. Ranking questions



The types of closed-ended questions discussed in this chapter provide an overview of the most common formats (for an overview, see Couper, 2008a). Because Web surveys are self-administered, no interviewer support is possible, rendering the questionnaire and the questions used even more important when collecting data on the Internet. Closed-ended questions can be more specific, take less time to respond to, and are easy to analyze, which is why they are more frequently in comparison to open-ended questions.

5.2 Advantages and disadvantages of closed-ended questions

The widespread use und the superiority of closed-ended questions is due to other factors as well. Answers to closed-ended questions can easily be fitted to the categories provided and can therefore be compared easily between respondents. While the answers to open-ended questions range widely, they are not necessarily able to be tapered to a single category, or perhaps any answer category at all. In addition, open-ended questions require extensive coding, in comparison to closed-ended question formats that do not have to be coded at all, and which save time and money (Reja, et al., 2003). Since no coding is required in closed-ended formats, there are no coding-related biases. While these advantages mostly apply to researchers, there are several advantages for respondents too. First, understanding the question is easier as the answer categories provide additional information on the question's intention. Therefore, question comprehension is improved by the answer categories, and negative response styles like satisficing or item-

nonresponse more unlikely. Answering closed-ended questions is easier as well. Respondents do not have to formulate an answer in their own words. Instead, they only have to check the response option that applies to their answer, making the answering process much easier and less demanding, resulting in better response rates to closed-ended questions. Not having to explain a response can also be less intimidating and therefore it is easier for respondents to answer questions about sensitive topics. However, closed-ended questions are only easier to process and answer when all relevant answer categories are included. If not, respondents can get frustrated by the fact that their expected answer category is not provided, or not provided in sufficient detail, resulting in a feeling that they have no chance to express their intended answer.

No irrelevant responses are produced, given by the restriction of a closed-ended format. Further, respondents who have no opinion or do not know the answer can easily guess a response or even answer randomly (Krosnick, 1991). Overall, closed-ended questions give no indication as to whether a respondent interpreted the question in the intended way, while open-ended questions allow detecting misinterpretations made by the respondents. Next we visit this in greater detail when discussing the advantages of open-ended questions, since those can only be made visible in comparison to closed-ended questions.

5.3 Types of open-ended questions

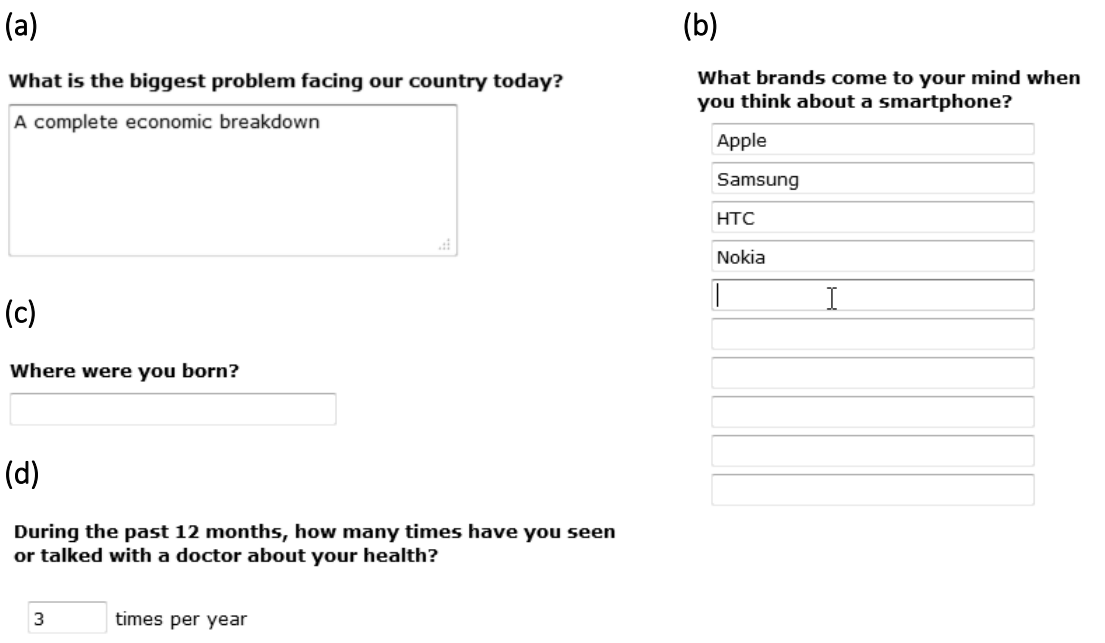
While closed-ended questions in Web surveys are predominantly realized by radio buttons or check boxes, text or numeric input is captured by either text boxes or text areas. Text boxes that use the “input” form element are suitable for rather short text or numeric input, since they allow only for a single row. Text areas, however, are useful for longer texts since they can be displayed in various sizes and with multiple rows. Of course, text boxes and areas can be combined and displayed in various ways, depending on the information they try to obtain.

These two form elements (text areas and text boxes) allow for a variety of open-ended questions and different types of design (Couper, et al.,

2011). There are narrative ones, e.g. “What is the biggest problem facing the nation?” where respondents are encouraged to freely answer in their own words in order to get long, thick, and rich responses. Since the answer-box size should reflect the size of the response intended by the researcher, its size should be larger than that for a closed-ended question, and realized via a text area box. As this thesis focuses on narrative open-ended questions, the following chapters will provide a closer look at this type of an open-ended question.

Open-ended questions are used for narrative responses and can also be used to obtain short responses, e.g. “What is your favorite television show?” or even shorter questions that seek only a single word or phrase, e.g. “Where were you born?” Short text responses are usually captured by using text boxes rather than text areas, and can be also combined into list-style questions asking for enumerations, e.g. “Please list the names of all foreign countries you have visited in the past 12 months” (Fuchs, 2009b). List-style open-ended questions therefore often use multiple text boxes to illustrate the enumeration task as displayed in Figure 8.

Figure 8. Types of open-ended questions: (a) narrative, (b) list-style, (c) short and (d) frequency



Frequency questions can also be asked by providing a text box rather than a range of closed-ended response options. Instead of checking a radio button, respondents just type their response into the answer-box provided with the question. Using open-response formats asking for frequencies is preferred, since response categories bias responses especially in frequency questions, and the coding of numbers is not costly (Krosnick, 2006; Schwarz, Hippler, Deutsch, & Strack, 1985).

Surveys in social science, with few exceptions, are mostly based on closed-ended questions (Schuman & Presser, 1981). When narrative open-ended questions are included in surveys, they are mostly used for questions where other formats are not possible. Further, open-ended questions are frequently presented at the very end of a questionnaire to give respondents the chance to enter some final thoughts.

So far, most of the experience and the arguments for the superiority of closed-ended question formats are based on common sense and anecdotal experience rather than on sustainable studies (Schuman & Presser, 1981). The reason that open-ended questions are often excluded from surveys is a result of their inefficiency in processing them, rather than any negative influence on data quality.

5.4 Advantages and disadvantages of open-ended questions

Overall, Schuman and Presser (1981) name two important reasons to choose open-ended questions over closed-ended formats. First, answer categories constructed in a preparatory stage might fail to provide an adequate set of response alternatives in a closed-ended question. Respondents will then break-off the survey, skip the question, or choose the best answer-option provided, but in any case, they were not able to express their originally intended answer. Second, respondents are influenced by the categories provided with a question (see chapter 6).

In a series of experiments, Schuman and Presser (1981) compared the answers given to a question using either a closed- or an open-ended response format. In their first study, the authors asked participants for

the “Most important problem facing the United States”. Based on answers from previous studies, Schuman and Presser constructed eight closed items that were integrated into a survey where respondents were randomly assigned to an open-ended or a closed-ended answer format for the same question, concerning the most important problem facing the nation. The results of their study indicate that out of the eight closed-answer categories, only responses for the categories of crime, inflation, and unemployment resemble the answers provided by respondents using the open-ended answer format. The crime and violence categories yielded twice as many responses when offered in the closed format, while in the open-ended answer format unemployment was indicated as the most important problem facing the nation.

In a second study, on work values, Schuman and Presser (1981) carried out the same experimental design as before. In the closed-ended design, respondents were asked, “Which thing on this list would you most prefer in a job?” or “What would you most prefer in a job?” in the open-ended format. Five closed-answer responses, “High income”, “No danger of being fired”, Working hours are short, lots of free time”, “Chances for advancement” and “The work is important, and gives a feeling of accomplishment” were adopted from Lenski (1963). Almost 60 percent of the open responses fell outside the five answer categories provided for the closed format, indicating gross differences between answers provided. Respondents to the open-ended format produced a much more diversified set of answers. Despite this diversity, the authors discovered two additional categories in the open responses, namely “pleasant and enjoyable work” and “work conditions”, suggesting that the closed-answer categories provided were insufficiently developed. In a next step Schuman and Presser extracted their own five categories. These were: Work that (1) pays well, (2) gives a feeling of accomplishment, (3) does not have too much supervision and you make most decisions yourself, (4) is pleasant and where the other people are nice to work with, and (5) is steady with little chance of being laid off. These were based on the answers to the open-ended questions, and included in a third survey asking the exact same work values question.

Even though the new five categories were more suitable and responses to the open-ended format did not spread as widely as they did in the previous study, there were still a variety of substantive answer categories missing. On the other hand, answers provided to the open-ended answer format were sometimes very vague, did not always match the question's intention, and therefore could not be coded properly. Thus the authors conclude that open-ended questions are essential to frame the reference of responses and for finding suitable answer categories. When this is accomplished, there is no need to use the open-ended format instead of closed-answer categories (Schuman & Presser, 1981).

The experiments by Schuman and Presser were all carried out via landline telephone surveys. While interviewer-administered surveys allow for probing, self-administered surveys do not, and results can therefore not be fully applied to a Web survey setting. A comparison between narrative open- and closed-ended answer formats in a Web survey design was realized by Reja et al. (2003). The authors asked participants to name the most important problem the Internet is facing today. Respondents were randomly assigned to answer the question via: (1) open-ended format; (2) closed-ended single-choice format using radio buttons with 10 answer categories plus one "other" category with a short text-entry box; or (3) closed-ended multiple choice format using 10 check boxes as well as the "other" category. The open-ended "other" category was only used by 5 percent of the respondents to the closed-ended single choice format, and 3 percent of the respondents on the closed-ended multiple choice format. Based on the answers given to the "other" category, the authors could only formulate one additional answer category that was absent from the 10 answer categories provided. The open-ended response format, in contrast, produced a more diversified set of answers. An additional 8 answer categories to the 10 pre-coded answer options were obtained after coding the answers to the open-ended response format. Overall, 63 percent of respondents to the open-ended format provided these 8 additional categories. Therefore open- and closed-ended response formats show extensive differences in the answers they elicit (Reja, et al., 2003). It also becomes apparent that an

additional open-ended “other” category cannot secure comparable answers to an open-ended question. While substantive differences occurred in the frequency distributions between the three designs, differences in the ranking of answer categories were minor. For instance, all three designs show that “data transmission safety” is the most important problem the Internet is facing today. While in the closed-ended single choice condition, 24.6 percent of the respondents name “data transmission safety” as the biggest problem, 60 percent gave this answer to the multiple choice closed-ended question, and 32.4 percent in the open-ended format. Despite some answers that could not be coded at all, several answers provided to the open-ended question were not sufficiently exclusive. While the coding of open-ended questions can be challenging, the difference in item-nonresponse and missing data are the larger problem. Overall, 41 percent of the respondents assigned to the open-ended format either skipped the question or provided an invalid answer, while in the closed-ended conditions, item-nonresponse was negligible.

Closed and open-ended response formats both reveal strengths and weaknesses. Schuman and Presser recommend open-ended questions especially when proper closed-ended are not available to pose a question in a closed-ended format. However, there are additional factors in the decision as to whether to use a closed or an open-ended response format.

Open-ended questions allow respondents to answer freely in their own words and in the detail required to clarify and quantify their answer. Respondents are not forced into a set of answers and they can indicate the intensity of a response. Even though analysis of open-ended questions can be time-intensive, researchers will get more information from open-ended questions. If results reveal that the most commonly elicited responses to closed-ended questions are only provided by a small proportion of respondents when answering in an open-ended format, these results would imply that forcing respondents to use a closed format will not necessarily include the aspects that respondents generally associate with a particular question (Haddock & Zanna, 1998). Therefore,

predetermined answer categories may fail to capture respondents' actual intended answers since they cannot pre-determine what aspects respondents will initially associate with the question.

The studies by Schuman and Presser show how to use open-ended questions to better understand which answer categories might be suitable for a question. Open-ended questions can be used to get an exhaustive list of possible answer categories based on free responses. Therefore, such narrative open-ended questions are ideal for exploratory research. Open-ended questions are also useful for complex issues that cannot be condensed into a few answer categories (Couper, et al., 2011). Expanding the number of answer categories so as not to miss any valid response cannot be a solution because too many categories will confuse respondents, lower response rates, and lead to negative response behaviors such as satisficing. In addition, open-ended questions don't give away an answer and they don't allow for guessing, which is especially important on knowledge questions. They provide more room for creativity and self-expression and are easy to answer and more natural (Geer, 1988). "What do you think about your new phone?" is an example of a type of open-ended question we hear every day, while it is highly unlikely in a natural conversation to be asked, "On a scale from 1 to 10, how would you rate your new phone?" Because open-ended questions often resemble natural, everyday communication much better than closed-answer formats frequently do, some respondents will judge open-ended questions as less boring and more enjoyable. On the other hand, several studies have demonstrated that narrative open-ended questions suffer from higher item-nonresponse (Reja, et al., 2003). In a Web survey conducted by Crawford, Couper, and Lamias (2001), 56.5 percent of all abandonments occurred on open-ended questions. Higher item-nonresponse on open-ended questions was also found by Scholz and Zuell (2012) and Smyth, et al. (2012).

Open-ended questions require more consideration, time, and effort to answer than simply selecting from a list of items (Holland & Christian, 2009). The burden of responding to open-ended questions is therefore somewhat higher, compared to closed-ended formats, and respondents

are less motivated to answer them (Galesic, 2006). In a study on employment status (Galesic, 2006), questionnaire sections containing open-ended questions were judged by the respondents to be less interesting and more burdensome. In addition, respondents needed more time to answer sections that included narrative open-ended questions. Further, when a previous section was judged as enjoyable, item response rates were higher and answers to open-ended questions longer. Oudejans and Christian (2010) showed that respondents who said that the subject of the questionnaire was interesting were more likely to answer, and also provided longer responses. But even if respondents are willing to answer an open-ended question, they might have difficulties articulating their own views, which is why these questions may measure people's education and not necessarily their attitudes (Geer, 1988, 1991). Narrative, open-ended question require superior writing skills and the ability of the respondent to verbalize feelings. In fact, education is a strong predictor of an actual response as well as the number of words provided by respondents in narrative open-ended questions (Denscombe, 2008; Oudejans & Christian, 2010). Further item-nonresponse is affected by respondents' ability and willingness to answer (Beatty & Herrmann, 2002).

Besides an educational influence on the responses on narrative open-ended questions, the literature also revealed effects of sex and age. While females are not in general more likely to respond, on average their answers are longer in comparison to answers from male respondents (Denscombe, 2008; Oudejans & Christian, 2010; Stern, et al., 2007). In a study by Stern, Dillman and Smyth (2007), women provided longer responses; respondents over 60 years of age provided responses that were at least one word longer, compared to younger respondents. In addition, Smyth et al. (2012) found higher item-nonresponse rates across persons of low literacy, who also gave shorter answers to narrative open-ended questions.

Based on these findings, narrative open-ended questions especially struggle with high item-nonresponse. The fact that cognitive abilities influence responses makes the use of open-ended questions even harder

to justify. Any improvement of narrative open-ended questions therefore must start with the response burden. Lowering the burden and getting as many participants as possible to respond an open-ended question in a Web survey is exactly what the designs tested in this thesis are aimed at. After looking at the influence of visual questionnaire design in the next chapter, we will focus on how to use visual design in order to improve responses to narrative open-ended questions.

6 VISUAL QUESTIONNAIRE DESIGN

While in interviewer-administered telephone and face-to-face surveys, verbal language is the most important source of information, self-administered surveys contain graphical paralanguage as well. Every aspect of a question affects respondents. In addition to the question itself, the verbal and visual language of response categories, e.g. buttons, checkboxes, input-fields and their arrangement on a screen are also influential. Christian and Dillman (2004) name three distinctive visual design languages. 1. Numeric language includes all numbers displayed and used in queries and answer categories. 2. Graphical language characterizes the size, spacing, and location of information, while 3. Symbolic language contains arrows, pictures, and answer-boxes. It is important to point out that all three languages affect the way respondents read, interpret, and answer survey questions. They never stand alone and combine to create meaning for respondents (Redline & Dillman, 2002).

6.1 Numeric language

Numerical language contains all information displayed by numbers in a question. Various experiments on anchoring effects revealed rather strong influences of numbers on survey responses. Anchoring describes a bias towards a certain value and the tendency of respondents to rely heavily on that piece of information, even if the anchor values are absolutely uninformative and inadequate. Tversky and Kahneman (1974) let respondents spin a manipulated “wheel of fortune” with the possible outcomes 10 and 65. After spinning the wheel respondents were asked whether the percentage of African countries in the United Nations was more or less than 10 percent (low anchor condition) or more or less than 65 percent (high anchor) of all countries represented in the UN. Next participants were asked to give an accurate estimate of the percentage of African countries in the UN as compared to the whole body. Respondents in the low anchor condition (10 percent) estimated the percentage of African countries in the UN compared to the whole body,

on average, at 25 percent, while those in the high anchor condition (65 percent) answered that African countries represented 45 percent of countries in the UN (Tversky & Kahneman, 1974). Even though respondents had to assume that the outcome of the wheel was pure chance, it influenced their answer. Since Traversky and Kahnemann, several slightly similar experiments had been carried out. For example, Mussweiler, Englich, and Strack (2004) asked respondents, "Is the Brandenburg Gate taller or shorter than 150 meters?" vs. "Is the Brandenburg gate taller or shorter than 25 meters?" Respondents were then asked to give an estimate of the height of the Brandenburg gate with the result that respondents in the high anchor condition (150 meters) estimated the Brandenburg gate to be taller than did respondents in the low anchor condition (25 meters). Critcher and Gilovich (2008) showed respondents a picture of the football player Stan Fischer wearing a number 54 shirt (low anchor) or a number 94 shirt (high anchor). After that, they asked respondents "How likely do you think it is that Stan Fischer will register a sack in the conference playoff game?" Fifty-five percent of the respondents in the low anchor condition and 62 percent in the high anchor condition guessed Stan Fisher would register a sack in the next playoff game. In almost every study of anchor effects, high anchors resulted in higher estimates while low anchors resulted in underrated estimates. The sheer presence of a number alone seems to influence responses. However, some criticism can be leveled at the studies, beyond any artificial and unnatural design. By avoiding starting points, looking at previous questions and using clarifications, anchoring effects can be kept to a minimum. While anchor effects rely mostly on the whole question, scale effects are due to the answer categories as well, and even more relevant, as they cannot be easily avoided.

Frequency questions are especially influenced by the answer categories provided. Respondents always try to figure out what is expected from them and what the researcher intended by using numbers on an answer scale. Therefore, answer categories somehow imply a norm. Schwarz, Hippler, Deutsch and Strack (1985) asked respondents "How many hours

do you typically watch TV a day?" Researchers assembled two different answer scales to this question: a low-frequency and a high-frequency scale. In the low-frequency scale, respondents had to answer the question using six possible answer categories ranging from up to half an hour to more than two and a half hours, while in the high-frequency scale categories started with up to two and a half hours to more than four and a half hours. As a result, in the low-frequency condition, reported TV consumption was significantly lower in comparison to the high-frequency answer scale. While in the low-frequency condition 16 percent of respondents reported watching more than two and a half hours, in the high-frequency condition, 38 percent reported watching two and a half hours (Schwarz, et al., 1985). Knäuper, Schwarz and Park (2004) replicated the influence of low- vs. high-frequency scales, as did Toepoel, Das and Soest (2009). By using short time periods, providing recall cues, or using adequate categories that fit in the distribution of a targeted population, scale effects can be effectively reduced. By asking for frequencies along an open-format scale, effects can be prevented entirely.

Besides using numbers to gather frequencies, they are commonly used for rating scales as well. Schwarz, Knäuper, Hippler, Noelle-Neumann and Clark (1991) asked respondents in a face-to-face interview how successful they have been in life, using an 11-point rating scale with the labeled endpoints "not at all successful" and "very successful". The numbers attached to the scale were experimentally varied. Respondents had to map their answer to a scale ranging from 0 to 10, or from -5 to +5. Respondents reported higher rates of success along the -5 to +5 scale than along the 0 to 10 scale. The authors hypothesized that in the 0 to 10 condition, 0 might be interpreted as an absence of success, while the minus sign in the second scale could be interpreted as the presence of explicit failure. Fuchs (2003) replicated the experiment using the same question and also found higher reported success when respondents were required to answer the question using the scale -5 to +5.

All these studies suggest that numerical language influences response behavior. Respondents gain information about the researcher's

expectations and intentions from the response alternatives provided. Thus, beyond verbal labels, respondents may use numerical language as an additional source of information when interpreting the question's meaning (Christian & Dillman, 2004).

6.2 Graphical language

The arrangement of numbers, words, and symbols, as well as their position, sizing, color shape, and location affect the graphical language (Christian & Dillman, 2004). Verbal, numerical, and symbolic languages are transmitted through the visual channel via graphic paralanguage (Redline & Dillman, 2002). Respondents interpret questions differently when words are highlighted by color or when a bold font is used. The arrangement of question components also influences responses.

Tourangeau, Couper and Conrad (2004, 2007a) argue that respondents follow simple heuristics when answering survey questions. The authors distinguished five heuristics that respondents may follow when interpreting the visual questionnaire: 1. middle means typical; 2. Left and top means first; 3. Near means related; 4. Up means good; and 5. Like means close.

Referring to the first heuristic, respondents tend to interpret the middle point of a scale as the normal and most typical response option and use it as an anchor. The previously reported experiment by Schwarz et al. (1985) on TV consumption revealed the importance of anchor effects and numerical design language. However, this experiment illustrates the influence of graphical language as well. Despite the numeric labels used, the position of the answer categories affected responses. The midpoint of a scale represents the most typical response, a neutral point, or a conceptual midpoint, and constitutes the "middle means typical" heuristic (Tourangeau, et al., 2004, 2007a). To test the first heuristic the authors compared two attitude questions in a Web survey where the non-substantive responses "Don't know" or "No opinion" were presented simply as additional radio buttons or as scale points separated by a divider line or a dividing space. When a divider was used, (regardless

of whether it was a divider line or divider spaces), the visual midpoint of the scale fell at the conceptual midpoint. When the non-substantive answer categories were only displayed as additional radio buttons, the visual midpoint fell on the end of the scale and affected the distribution of the answers provided in comparison to when a divider was used. In another experiment the authors used even and uneven spacing for the scale points of a question, and found that when the visual midpoint was positioned in the lower end of a scale relatively to the conceptual midpoint, the mean response was lower as well (Tourangeau, et al., 2007a).

The “left and top means first” heuristic implies that in a list of items, left and top placed items will be seen first by respondents, corresponding to the reading order in most Western countries. When response categories are ordered from positive to negative or agree to disagree, for instance, respondents will expect a logical order for each of the following items. If the responses provided do not follow this heuristic, respondents may become confused, make mistakes, and take longer to respond (Tourangeau, et al., 2004). The authors varied the order and consistency of response options in behavioral frequency and attitude items. According to the “top and left means first” heuristic, it is easier to process consistent response categories rather than strongly inconsistent ones, and as a result, it took respondents more time to answer the question when answer categories were inconsistent.

The third heuristic indicates that items which are physically near each other on a screen appear to be related conceptually for respondents. Based on the Gestalt Law of Proximity, the authors suggest that “near means related” and response options that are further apart or presented on separate screens will be less likely seen as related, and influence responses accordingly. To put their third heuristic to a test, Tourangeau et al. (2004) randomly assigned respondents either to eight items on a single screen in a grid format, to four item grids on two screens, or to one question per screen in the third condition. Responses to the eight items were more highly inter-correlated when they were presented in a grid on a single screen, than when the items were presented in two grids

on two screens or when a single screen per item was used. Overall, when items were grouped into a single grid, respondents seem to infer greater similarity between them.

In their forth heuristic the authors present a variant of their second heuristic and state that in a vertically oriented list, the item on top of the list tends to be seen as the most desirable. In a response list, the authors varied the position of an unfamiliar car model. As a result, respondents made inferences about the unfamiliar item based on where the item appeared in the presented list. When they had to rate the prices of a series of cars ordered from most to least expensive, respondents inferred the price of an unfamiliar car model based on its position in the answer series (Tourangeau, et al., 2004).

The fifth heuristic is based on the Gestalt Law of Similarity and claims that response options that are visually similar will be seen as conceptually closer as well. Therefore, like in appearance means close in meaning, and items that look alike in aspects such as color, font or type size may convey a relationship between the response options provided (Tourangeau, et al., 2007a). The authors hypothesized that in a scale question respondents use cues such as differences in hue or font size when interpreting the meaning of scale points. When the two endpoints of a scale are displayed in different hues, respondents will see the scale as covering a broader conceptual range than when no color is attached to the scale (Tourangeau, et al., 2004). Overall the results indicated that respondents attend to all the details provided on a response scale. Especially when difficulties occur in the interpretation of the scale points' meaning, respondents use every possible cue (including secondary cues like color) to decode a scale's intended meaning.

Toepoel and Dillman (2010) replicated and extended the experiments on the heuristics of Tourangeau et al. (2004, 2007a) in order to find a hierarchy of the visual design aspects respondents use when answering a survey question. The authors varied verbal, numerical, and visual language of a 5-point rating scale. Respondents were randomly assigned to polar point or fully labeled scale with numbers added or not, and

uneven spacing of the response options or not, in order to test the "middle means typical" heuristic. Overall the results indicated that verbal information is more important than numerical language, and that numerical language is more important than other visual language. The "middle means typical heuristic" was only apparent in the polar point scale and the not fully labeled, scale while adding numbers helped to reduce the effect of visual design language. In a second experiment, Toepoel and Dillman (2010) tested the "like means close" heuristic by assigning respondents randomly to these different types of answer scales: polar point, polar point with numbers, fully labeled, fully labeled with numbers, polar point with color, and fully labeled with color. Again, verbal labels had a strong effect on the answers. Further, positive answers were more often provided when they were shaded in a green color or when the numbers were added to the response options. Despite these numerical and graphical design influences, visual design effects were overpowered by the verbal language of a fully labeled scale.

The notion of verbal language holding sway over visual language was also acknowledged by Emde and Fuchs (2012). In a Web survey, the authors varied the visual appearance of a faces scale question using animated smileys as symbolic labels for the answer categories. Their fixed design included no animation at all. In the affective design, the faces changed color and increased in size with the cursor hovering over, while in the cognitive version the faces did not change color and zoomed out and a text answer category was displayed. The radio button control design used the same answer categories as the cognitive design, but included neither faces nor animation. While there was no significant influence attributable to face color and size in comparison to answers for the fixed faces scale design, the cognitive faces scale design that combined faces and verbal labels provides corresponding answers to the radio button question. When using faces scales and verbal labels on the radio button design, the results show that the verbal design overruled the visual, even when the visual design is more complex and animated faces scales are used.

Christian and Dillman (2004) carried out a series of experiments varying the graphical design language. First the authors manipulated the location

of an instruction and placed it either after the question or after the yes/no response categories provided. When the instruction was placed after the question, respondents could process the instruction before mapping their answer to the provided answer categories. By placing the instruction after the answer options, it is not available when needed for formatting the answer. And indeed, the location of the instruction influenced the responses in the expected way, and even confused some respondents as they used the instruction placed after the answer categories for the subsequent question as well (Christian & Dillman, 2004). These findings imply that the question text, instructions, and answer categories should all be displayed consistently to the question–answer process.

Kunz and Fuchs (2012) used eye-tracking data to identify the ideal position of clarification features such as definitions, retrieval cues, motivational statements or formatting instructions. Their study showed that the actual position of clarification features affected responses and should be adapted to the question–answer process. Definitions should be displayed before the question text, and formatting instructions next to the response options (Kunz & Fuchs, 2012).

In a second study, Christian and Dillman (2004) compared two linear scales where all scale points were either listed vertically or nonlinearly, using double- or triple-blanked responses. The authors assumed that it easier for respondents to answer the linear layout. Even though double- or triple-blanked responses are commonly used to save space in (especially) paper questionnaires, respondents were more likely to select a response from the top line in the nonlinear version in comparison to the linear scale when they choose their answer from a continuum.

Christian and Dillman (2004) asked respondents, “On a scale of 1 to 5, with one being very satisfied and 5 being very dissatisfied, how satisfied are you with the classes you are taking this semester?” Respondents were assigned either to a numbered and endpoint-labeled polar point scale, or an input field requesting respondents to write in the number corresponding to their answer. Respondents provided much higher

values to the input field in comparison to the polar point answer scale, suggesting that the removal of verbal cues (scale endpoints) from the question might have confused respondents.

In addition, Stern, Dillman and Smith (2007) also compared responses to polar point scales of number input fields and added a “don’t know” option to both designs. Again, respondents provided more negative responses to the input field design. Further, the “don’t know” option was more frequently ticked when the input field was displayed, supporting the assumption of Christian and Dillman (2004) that the input field design seems to be more confusing.

Further, Christian and Dillman (2004) manipulated the space between response categories of nominal and ordinal answer scales using either equal or unequal spacing. The authors expect respondents to select categories that are set off from others more often. While answers to the nominal scale were affected by the spacing, the ordinal scale and the more set-off category was more frequently chosen, answers to the ordinal scale were not affected by the varied spacing. The ordinal scale question therefore is less susceptible to graphical spacing effects as the respondents better understand a question’s intention based on the question stem (Christian & Dillman, 2004).

Open-ended questions are also affected by graphical language. Fuchs (2009b) assigned respondents in a Web survey to differently designed open-ended list-style questions where respondents were requested to enumerate and list short entries in the response fields provided. Once the respondents start typing their first entry, another response field was provided, and after that a third, and so on. In the control group all input fields to enumerate the answer were displayed from the start posing a higher response burden in comparison when only one input field was displayed in the experimental group. The dynamic list-style design (stepwise adding additional input fields) increased item-nonresponse and thus did not reduce the burden to respond, while the number of characters increased in comparison to the control group (where all input fields were presented at once). Smyth, Dillman, and Christian (2007) also

compared multiple small answer-boxes to a single larger box and found a higher number of enumerations for the multiple list-style answer-boxes. In addition, Keusch (2013) found that list-style answer-boxes limit the maximum number of enumerations to the number of boxes presented. The number of boxes also influences the awareness of low-frequency responses, since respondents try to use the answer-boxes provided along with the question (Keusch, 2013).

The influence of graphical language on responses is also present in open-ended questions. Smith (1995) argued that allowing respondents more space for recoding narrative open-ended responses actually produces longer responses. Smith illustrates that by referring to the 1954 Stouffer study, which accidentally varied the answer space of two open-ended questions as the survey was carried out by two different companies who printed their questionnaires themselves. As a result, in one questionnaire the open-ended question allowed five times as much open space as the other, and the word count on the two different designs varied. Ultimately, the larger answer space facilitated and encouraged longer and more detailed responses (Smith, 1995).

Via a self-administered questionnaire, Christian and Dillman (2004) tested how increasing the size of answer spaces on open-ended questions affect responses. The authors randomly provided a small answer-box or a box twice that size to three open-ended questions. The number of words as well as the number of reported themes was affected by the different answer spaces. While for all three questions the number of words was significantly higher when the answer space was doubled, the number of reported themes was only higher in two questions.

Israel (2010) varied the answer-box height for two narrative open-ended questions. While in the first question the answer space was doubled, four different heights were used for the second question in order to better understand if there is an optimal size for a narrative open-ended answer-box. As a result, the mean number of words as well as the number of lines of text and sentences increased with the answer-box height. Based on 2,200 mail survey responses collected from 2003 to 2006, Israel

(2010) further investigated how varying the answer-box space of narrative open-ended questions affected the response length. While the width of the answer-boxes remained the same in every survey, the height was increased linearly measuring .28, .56, .84, 1.12, 1.4 or 1.68 inches. Overall the answer-box height had a strong effect on the length of responses, but not on the propensity to provide an answer. The mean number of words increased with the answer-box height, as did the number of themes reported. The number of additional details also grew along with the answer-box-size. When the smallest answer-box was used, almost 20 percent of the respondents continued writing outside the answer space, while in all larger answer-boxes, only 3 to 7 percent of the respondents did that.

In a mail survey, Stern, Dillman and Smyth (2007) used two different sized answer-boxes to ask respondents to elaborate on a previously given answer. Consistent with earlier studies, the larger answer-box elicited longer responses. Respondents to the large answer-box provided 17.17 words on average, while to the half-sized small answer-box the mean number of words was 14.95. The authors further controlled for respondents' characteristics and found the influence of box-sizes on response length to be apparent independent of respondents' age, gender, or education.

While previous studies focused on the effect of answer spaces in paper-based questionnaires, Smyth, Dillman, Christian and Mcbride (2009) could demonstrated the same response pattern for narrative open-ended questions in Web surveys only for less motivated late respondents. Smyth at al. (2009) argue this finding to be encouraging because it indicates that larger box sizes (and solely the visual design) can stimulate less motivated respondents. To further enhance responses to narrative open-ended questions, the authors varied the instructions in combination with differently sized answer spaces. Despite using no instruction at all, in a second condition, respondents were instructed that they were not limited in their response by the size of the answer-box, while in a third condition the motivational instruction on the importance of this particular question was given. Pointing out that the box does not

limit the response length resulted in longer responses for early and late respondents, while the motivating instruction produced longer responses among all respondents (Smyth, et al., 2009).

Fuchs (2009b) included two narrative open-ended questions to a Web survey and added a counter, or not, indicating the number of characters left to a small (2 rows), a mid-sized (4 rows), or a large (6 rows) answer-box. In line with the previous studies on different answer-box sizes, respondents provided more characters when exposed to a large answer-box in comparison to a mid-sized or a small box. The adding of a counter increased the number of characters reported, but it did not affect item-nonresponse.

6.3 Symbolic language

Symbolic language shares many similarities with graphical design language. It transmits information to respondents by using signs with a cultural meaning that can be used by respondents to interpret the question, and by researchers to make sure that respondents interpret the question as intended. For example, directing arrows might help to ensure that the respondent will check where they are pointing.

In a paper questionnaire, Christian and Dillman (2004) used a branching arrow to direct respondents to a subordinate question. The directional arrow was expected to reduce branching errors and increase the likelihood that respondents would actually answer the subordinate question. Indeed, the percentage of eligible answers was higher when using a directional arrow. On the other hand, the number of ineligible answers grew as well. The authors conclude that the symbolic language manipulation significantly influenced the respondent behavior (Christian & Dillman, 2004).

Israel (2006) also used arrows in a paper-based questionnaire to direct respondents to a follow-up question. The author hypothesized that the presence of directional arrows would help reduce nonresponse on the follow-up question when compared with not using any guiding symbols. The influence of the directional arrow on the initial question was not

significant compared to when it was omitted, and even the influence on the follow-up question was minor. On the other hand, the arrows helped respondents navigate correctly between the initial and the follow-up question, resulting in fewer mismatched responses and slightly higher response rates for both questions.

Redline, Dillman, Dajani and Scaggs (2003) also tried to optimize branching conditions in a paper-based questionnaire and besides simple changes in the verbal language, tested how graphic and symbolic language can help respondents handle branching instructions correctly. The authors tested how a larger bold font (graphic language) and directing arrows (symbolic language) reduce branching errors. While the graphical highlighting of the font slightly improved correct responses, the use of two directing arrows reduced commission errors by one-third and omission errors by one-fourth, suggesting that respondents extracted meaning from the visual design and used that to answer the questionnaire in the intended way.

Despite design opportunities that exist when adding color, symbols or lines, the ease of adding pictures to Web surveys offers various opportunities in the visual design of questionnaires, but at the same time influence respondents. In a Web survey, Couper, Tourangeau and Kenyon (2004) added explicative pictures to several behavioral frequency questions (e.g. traveling, shopping, eating out). For example, the question “About how many times have you eaten out?” was accompanied by a high-frequency picture (one person eating fast food in a car) or a low-frequency picture (a couple dining in a restaurant). As a result, high-frequency pictures resulted in a significantly higher number of reported events (e.g. eating out) while low-frequency pictures evoked fewer events. Analog results were found for the other behavioral questions in the survey, suggesting that images are used for the retrieval of relevant information and therefore systematically influence responses. Using the same pictures, Toepoel and Couper (2011) found similar results in a probability-based online panel Web survey. However, the authors extended the previous experiments and verbal instructions to the behavioral frequency questions. With the high- and low-frequency

instructions used to include restaurant and fast food dining as well in the eating out example, the influence of the picture (whether high- or low-frequency) diminished. When the instruction told respondents to include restaurants but exclude fast food (low frequency) fewer eating-out events were reported. Again, the verbal instructions, and hence the verbal language, had a stronger effect on the behavioral frequency reports than did the pictures and visual language.

In series of experiments embedded in a Web survey, Couper, Conrad and Tourangeau (2007) added pictures of a woman jogging or a woman in a hospital bed to a self-rated health question. Despite a control group, where no picture was displayed, they also varied the position and size of the pictures used. The pictures were displayed directly before the question text, in the header of the Web page, or to the left of and above the question. When the picture of the sick woman was displayed, higher levels of personal health were reported in comparison to respondents who were shown the picture of a jogging woman. The effect was stronger when the picture was displayed directly before the question or on the top left. When placed in the header, the contrasting effect of the pictures was smaller.

Symbolic language affects open-ended questions as well. Despite using arrows in branching questions, Christian and Dillman (2004) also added lines to the answer space of an open-ended question to motivate respondents to provide longer answers. As a result, respondents used the lines as expected while they did not provide more detailed answers. The number of words, as well as the number or reported themes, did not vary significantly in comparing the answer field with the lines to an answer field without any lines.

Couper, Traugott and Lamias (2001) varied the input field length of an open-ended frequency question and found more unwanted qualified responses and non-numeric input when the input field was larger (16 digits) than when it was only two digits wide. In addition to the input field length, Fuchs (2009a) also varied the labels in a Web survey and found no

differences between long and short input fields on the amount of correctly formatted answers.

Couper, Kennedy, Conrad and Tourangeau (2011) tried to optimize input fields via visual design adaptations as well. First the authors varied the size of an input field of a frequency question and found only a minor effect on behavioral frequency reports. In a second study they placed a "$" symbol to the left and the two decimal places ".00" to right side of an input-field and found fewer incorrectly formatted answers in comparison to a plain design ("$" and ".00" not displayed). In order to optimize answers provided to a numeric date question, the authors randomly assigned respondents to a single long input field or two separated input fields for month and year, or two separated drop-boxes for month and year. The use of two text boxes increased the well-formed responses to almost 98 percent, while only 76 percent of the answers to the single input field were correctly formatted. In a fourth experiment, Couper et al. (2011) experimented with different input field designs for a question on date of birth. A single short input field, a single long input field, three separated input fields (month, day, year), or a drop-down box had to be answered. Again, the design using three separated input fields produced fewer ill-formed answers (1.3 percent) in comparison to the long (7 percent) and the short (7 percent) input fields (drop-down 0 percent).

Christian, Dillman and Smyth (2007) altered the visual design of input fields to open-ended date questions (month and year) in a series of Web surveys, lowering formatting errors. Providing a smaller box for the month and a larger for the year, instead of using equal box sizes, increased the percentage of correctly formatted answers, as did the use of the symbols "MM" (for month) and "YYYY" (for year) in comparison to the verbal labels "Month" and "Year" (Christian, et al., 2007). As the symbols convey additional meaning in shorthand, they make correct formatting much easier.

In a Web experiment, Fuchs (2007) varied the appearance of an input field by either placing a correctly formatted default value (0,000.00), or not, inside the input field for a question on the amount spent on

alcoholic beverages in the last four weeks. As a result, respondents to the default value input field provided more answers in the desired format (integers), while others used alphanumeric characters to elaborate their answer.

The studies reported in this chapter all point to the importance of visual design. Open-ended questions (whether used for frequencies or narrative input) are especially influenced by visual design language. The hierarchy in languages clarifies to the influence of visual design, especially for ambiguous questions. In order to assign meaning to a question, respondents may go beyond the information of the question's wording. Since open-ended questions do not provide a set of response options to choose from, they are more difficult to answer. Hence, respondents more often need to use all the information and cues they can get, including visual cues.

7 USING VISUAL DESIGN TO ENHANCE NARRATIVE OPEN-ENDED QUESTIONS

Narrative open-ended questions are affected by a questionnaire's visual design and even though the literature is not fully conclusive, Web surveys seem to yield longer responses to open-ended question in comparison paper–pencil studies (Denscombe, 2008; Israel, 2013; Kwak & Radler, 2002). Visual design aspects such as labels, the arrangement of answer categories, or the use of colors affect responses in general. When it comes to open-ended questions, studies have focused on the design of the answer-box displayed (Christian & Dillman, 2004; Israel, 2010; Smyth, et al., 2009), whether instructing labels were attached to the answer-box (Couper, et al., 2011; Fuchs, 2007, 2009b), or whether multiple boxes were used instead of a single one (Fuchs, 2009a; Keusch, 2012, 2013). Whatever answer-box design is used, the intention behind it has to be the improvement of responses to open-ended questions in order to get long, rich, and detailed responses.

Item-nonresponse and the accuracy of responses are the primary challenges when using narrative open-ended questions in a survey. Generally speaking, narrative open-ended questions are relatively burdensome to answer. Since respondents would rather satisfice than optimize (see chapter 4), the response burden is essential to whether or not a narrative open-ended question is answered and how it is answered. Question difficulty, the motivation to respond, and the ability to process all contribute to likelihood of satisficing. And while the abilities of a respondent cannot be affected by a researcher, respondent motivation and task difficulty can be influenced in order to improve narrative open-ended questions.

The easiest way to reduce the task difficulty in an open-ended question is simply to provide good and easily understandable question. Further instructions or examples can compensate for the lack of closed-ended guiding response options and help respondents understand a question. Even though these aspects seem to be a matter of course, open-ended questions should be chosen very carefully. The task difficulty is further

aided by the level of distraction a respondent has while filling out the questionnaire (e.g. multi-tasking on a PC). A researcher usually cannot control the environment in which a respondent is answering a survey, but can encourage respondents to take the time they need or finish the survey at a later time if that would be more convenient. Just pointing out the importance of the whole questionnaire, or even a single question, can improve respondents' willingness to participate (Israel, 2013; Smyth, et al., 2009). Overall the difficulty of the task and the respondent's motivation and abilities are essential when it comes to improving narrative open-ended questions (as well as other question types). Any alteration of an open-ended question should involve one of these three aspects in order to enhance data-quality.

In the hierarchy of the efficiency of visual design languages (see chapter 6), changes in verbal language are supposed to have the biggest influence on respondents. Despite the actual question and instruction wording, there are opportunities to further enhance narrative open-ended questions by using graphical and symbolic language.

This thesis focuses on ways to further improve the visual design of narrative open-ended questions. The results of three experiments will help explore the influence of different answer-box designs on data quality and the willingness of respondents to answer open-ended narrative questions. We will put four visual design variations to a test and expect every one of them to affect data quality in a positive way, when compared to a plain, standard open-ended answer-box design. We will first take a look at answer-box sizes and how they affect responses. Next, we will use a counter that indicates the number of characters left to improve responses. Third, dynamic answer-boxes that automatically grow with the respondent typing the answer will be tested by the respondent, in size adjusted answer-boxes, as a fourth design.

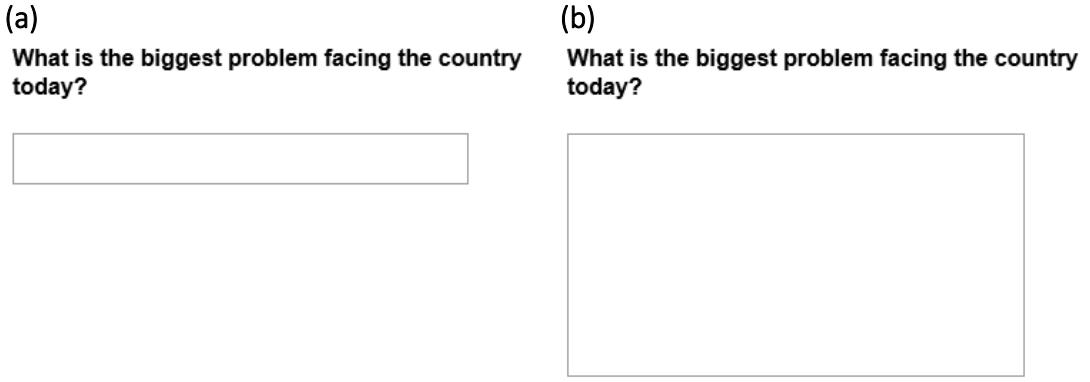
7.1 Answer-box size

Christian and Dillman (2004) as well as Israel (2010) altered the answer-box size in paper-based self-administered surveys and found longer

answers and more reported topics with the use of large answer-boxes in comparison to smaller answer-box sizes. Smyth, Dillman and Christian (2007) as well as Smyth, Dillman, Christian and Mcbride (2009) demonstrated the same pattern in a Web survey setting, where at least late respondents, who were assumed to be less motivated, provided more extensive responses to larger answer-boxes.

Choosing the best answer-box size for a narrative open-ended question is very important since respondents always try to interpret the researcher's intention; a large answer-box can create the impression that a lot of information is required while a small answer-box might convey that a short, less detailed answer is expected (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007).

Figure 9: Visual appearance of different answer-box sizes: (a) small single row input field, (b) large aswer-box



In addition, response length is likely to vary between different topics and questions. Generally it is assumed that respondents construe the answer-box size as auxiliary information when interpreting the question and generating their answer. If a large box is displayed, respondents are assumed to perceive the task at hand as broader, compared to a small box. The scope of the concept addressed in the question is assumed to be more extensive and that respondents will get the impression that they have to provide a longer response. By contrast, when providing a small box the scope of the question is limited and respondents are encouraged to shorten their answer. Consequently, small boxes seem to pose a lower

response burden while large answer-boxes increase this perceived burden, which could provoke increasing item-nonresponse. Choosing the ideal box size for an open-ended narrative question seems to require a trade-off balancing item-nonresponse with the extent of the answers provided. According to this reasoning, we expected an influence of the answer-box size on the information provided by respondents. Larger answer-boxes should lead to more characters and topics reported whereas responses to smaller answer-boxes would yield fewer characters and topics in comparison. While a small answer-box is supposed to pose a lower response burden and therefore produce less item-nonresponse, larger boxes pose a higher burden to answer. Therefore we expected lower item-nonresponse rates to smaller box sizes in comparison to larger boxes. At the same time, larger answer-boxes indicate that a more detailed answer is expected encouraging respondents to type longer responses and report more topics. Thus, response length and reported topics were expected to increase with larger box sizes.

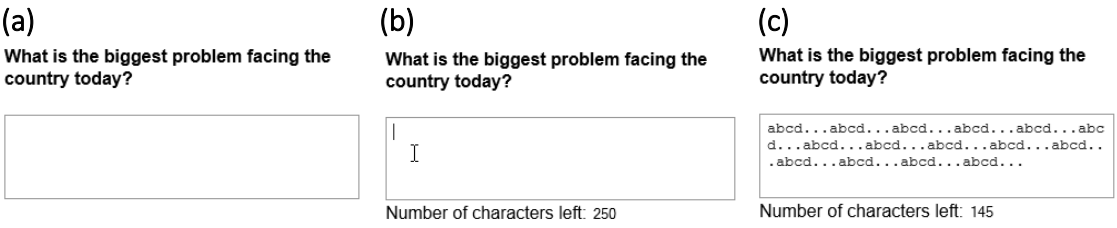
7.2 Counter

Large answer-box sizes are one way to enhance the response length and elaboration of answers to narrative open-ended questions. The downside of using large answer spaces is the higher response burden on the one hand and the fact that respondents tend to match the researcher's expectations as indicated by the answer space. Therefore, respondents are more likely to include irrelevant information in order to fill up the answer-box (Fuchs, 2009b).

One possible way to avoid this tradeoff of balancing response length and response burden is the use of counters indicating the number of characters left while the respondent types the answer. Just like the box size, the counter is important for the respondent as auxiliary information both when interpreting the question and generating the answer. Combining a small answer-box with a counter can help to reduce the burden to respond and at the same time strengthen the motivation to provide more detailed responses. This approach resembles in some way

the foot-in-the-door request strategy (e.g. Dillard, 1991). Asking for a smaller favor by using a small answer-box, more respondents will agree to this request and start to answer a question. When adding a counter, the requested task becomes bigger but respondents will be more likely to answer this broader task since they already agreed to answer the question in the first place by clicking the answer-box. The counter used in the later experiments was displayed after the respondent clicked the answer-box in order to isolate visual cues of the answer-box and the displayed counter (see Figure 10).

Figure 10: Visual appearance of a counter indicating the number of characters left: (a) initial appearance, (b) after clicking into the answer-box, (c) while typing

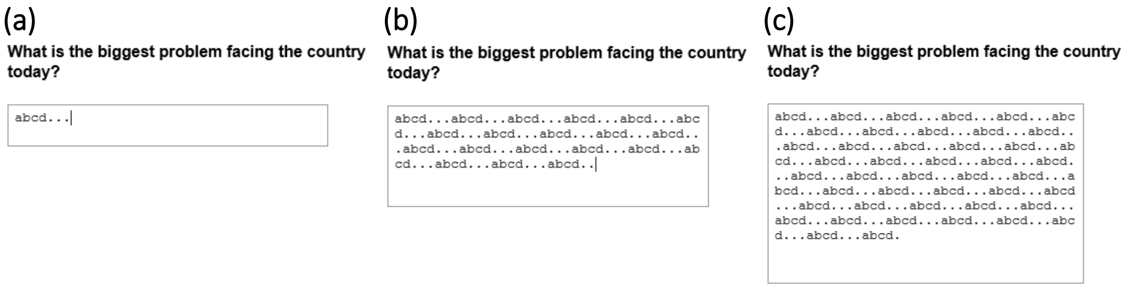


Accordingly, we expected the counter to work as a permanent stimulus that continuously motivates respondents to match their answers with the expected response length and provide more topics, more characters overall, and more characters per topic when exposed to an open-ended narrative question with a counter. We further expected the default counter value to affect the response length and the number of topics reported. A default counter set at a very low value as well as a very high default value might confuse respondents. The number of characters left as indicated by the counter should convey to some extent the impression of the amount of characters respondents infer from the answer-box size. We expect higher counter values to produce longer, more detailed answers, while lower default values shorten the responses provided.

7.3 Dynamic answer-boxes

Even though the rapid development of Internet data collection methods offers various graphic design features, previous studies mostly rely on a static Web survey design using pre-defined input fields. Adding a counter to an answer-box is a first step in making use of Web technologies. It is possible to further support and motivate respondents by dynamic, growing answer-boxes. By adding a dynamic answer-box to a narrative open-ended question, it is possible to start with a small answer-box implying a low response burden, while motivating respondents to provide a longer answer by continuously increasing the box size as the respondent types the answer. Each time the respondent completes a line of text in the answer-box, an additional empty line is added at the end of the answer-box (see Figure 11).

Figure 11. Visual appearance of a dynamic growing answer-box: (a) initial appearance, (b) after typing 4 rows, (c) continuously growing after 9 rows typed



As the counter is supposed to combine a low response burden implied by a small answer-box with the motivating factor, the dynamically growing box should work in a similar way. Once the respondent has agreed to answer the question, he will also consider providing more details as the answer-box grows. When using a dynamic, growing answer-box we expect less item-nonresponse in comparison to a large answer-box, and more information in comparison to a non-dynamic answer-box of the same size. However, continuously enlarging the answer-box space might also motivate respondents to provide too much or unimportant information. In addition, the initial size of the answer-box still influences

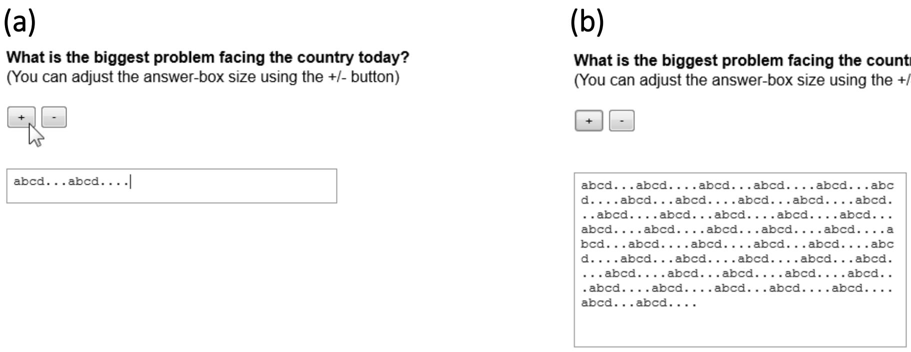
responses since it is the visual information respondents rely on when generating their answer. The expanding answer space appears later when respondents have already typed in the first words of their answer. Therefore, respondents might have to retrieve additional information to fill the additional space, or reformat their response. The burden of providing a response will change considerably as the respondent types the response, but we expect respondents to continue and complete the response process once they have begun it.

7.4 Respondent adjusted answer-boxes

Adding a counter or using dynamic growing answer-boxes can be a sensible way to improve responses to narrative open-ended questions, but that might also evoke problems. First of all, finding the ideal counter default value is difficult. Also, the dynamic growing answer-box has its flaws, since the adding of lines to the box occurs when the respondent has already formatted the answer to correspond to the starting answer-box size. Based on a foot-in-the-door-like principle (Deutsch & Deutsch, 1963), the counter as well as the dynamic design try to expand the task addressed to a respondent stepwise. On the one hand, this strategy works under many circumstances; on the other hand, respondents may feel misled by this spontaneous task enlargement. In addition, the counter and dynamic design somehow seem to patronize respondents in order to get the most information possible.

Instead of forcing respondents to provide longer and more detailed answers, widening their scope of action might be a better way to improve answers to narrative open-ended questions. In order to accomplish that, the answer-box sizes in our fourth visual design manipulation is set by the respondent. By using a “+” and “–” button above the answer-box, respondents are instructed to set their own size for the answer-box in order to give more space to those who need it, or fewer lines for those who have only a small amount of information to share on the question (see Figure 12).

Figure 12. Visual appearance of a size-adjusted answer-box: (a) initial appearance with buttons to increase and decrease the answer space, (b) after the answer-box size is set to 11 rows



The additional buttons, as well as the instructions, add to the overall burden, making this "self-adjusting" answer-box design even more demanding in comparison to a standard answer-box. On the other hand, providing the "self-adjustment" function for respondents grants them more freedom, which may make them more committed to the question, and the survey overall. Compared to the dynamic growing answer-boxes, the size-adjustment is placed at the beginning of the question answer-process, before the response has been formatted. In providing an initial box size, respondents still get an idea of the response length expected by the researcher, without unduly restricting their personal view on an adequate response length. Therefore, we expect self-adjusting answer-boxes to yield longer and more detailed responses in comparison to standard answer-boxes of the same size.

8 EXPERIMENTS TESTING VISUAL DESIGN IN NARRATIVE OPEN-ENDED QUESTIONS

8.1 Experiment I: Answer-box size, counter and dynamic answer-boxes

The first experiment is aimed at testing the influence of two different answer-box sizes, one, a counter that indicates the number of characters left and the other, a dynamic growing answer-box. This experiment will help to better understand the influence of visual design manipulations on data quality and how the different designs compare. We expected that larger answer-boxes would increase the number of characters and topics in responses, in comparison to small answer-boxes. Adding a counter to an answer space as well as using a dynamic growing answer-box design was also anticipated to help increase the response length.

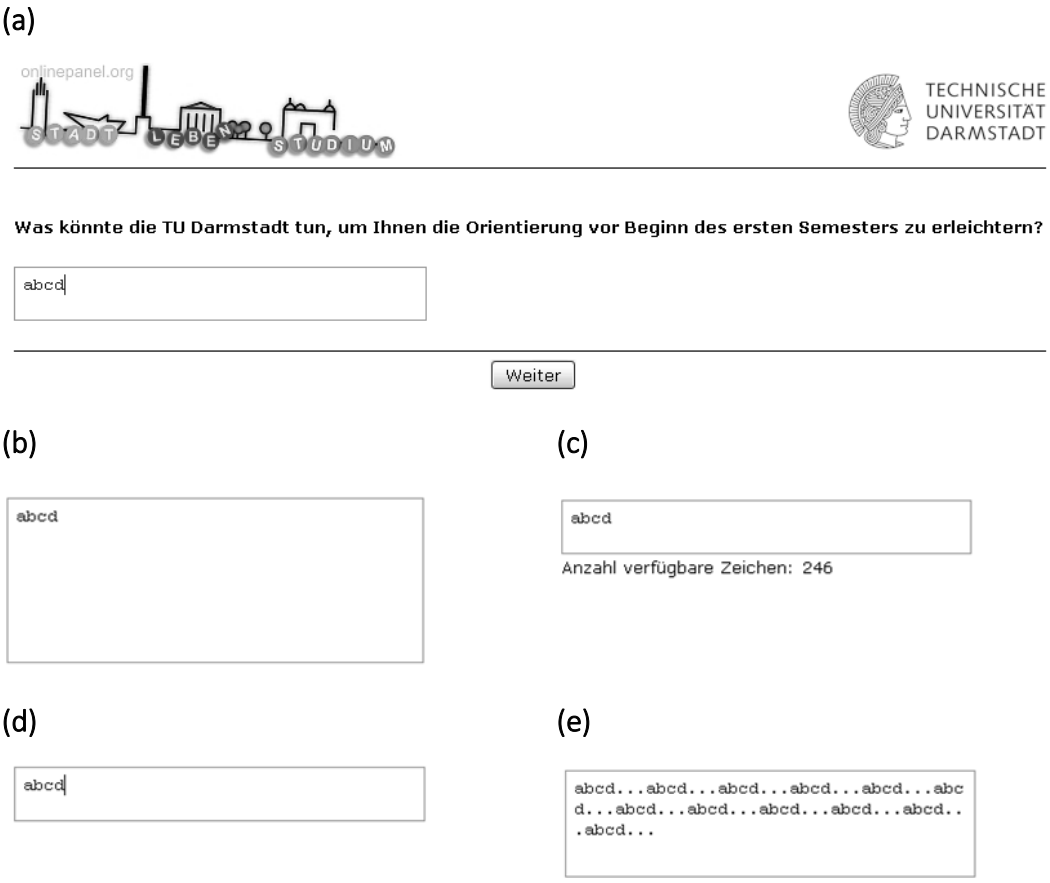
8.1.1 Experimental design

In October 2011, 4.342 university freshmen were invited by e-mail to join the online access panel at Darmstadt University of Technology. As an incentive, a lottery drawing for five book vouchers was offered and a reminder was sent after seven days of fieldwork. Overall, 673 students registered at the panel Web site and took the Web survey on the topic of being a freshman. No previous panel members were invited to the survey. The response rate amounted to 16 percent (AAPOR RR2).

Respondents were randomly assigned to one of four different versions of the same narrative open-ended question in a between-subjects design: *"In what ways could the university assist new students?"* The design of the first experiment is displayed in Figure 13. Respondents were randomly assigned to either a small (a) or large (b) answer-box. In the third experimental condition we added a counter (c) to an answer-box of the same size as the small box in experimental group (a). The counter was displayed after the respondent clicked into the field, in order to isolate visual cues of the answer-box and the displayed counter. The counter was displayed under the box, using an initial start value of 250

characters. The initial value of the counter was about three times higher than the number of characters that fit into the visible answer-box (2 lines of 42 characters each). The experimental condition (d) included a dynamic growing box. Starting with the same size as the small box, an additional empty line was added to the box for every row of text the respondent completed. Overall, answers were not limited by the box sizes and respondents were able to write up to 5,000 characters in every condition.

Figure 13. Answer-box designs in Experiment 1: (a) small answer-box; (b) large box; (c) small box with counter; (d) dynamic box after 4 typed characters; (e) dynamic box after 84 typed characters.



Notes. small box = 2 rows with 42 characters each; large box = 8 rows with 42 characters each, small box with counter = 2 rows with 42 characters each; dynamicl box = starting with 2 rows with 42 characters each

The validity of narrative open-ended question is difficult to measure. We gauge response quality of narrative open-ended questions by item-nonresponse, the number of characters, and the number of topics reported. A topic was defined (see Smyth et al., 2009) as a subject that answered the question and was independent of all other subjects mentioned within the same response. All topics were coded by one person. To measure elaboration, we use the characters per topic ratio, indicating how many characters were used by the respondents to account for every single topic.

8.1.2 Results

In Experiment 1, respondents were randomly assigned to one of four designs: a small answer-box, a large answer-box, a small answer-box combined with a counter, and a small dynamic growing answer-box. We assumed that the reduced response burden associated with a small input box would apply to all three designs starting with the small box, while the large box would produce a higher item-nonresponse.

Table 2. Item-nonresponse, characters, topics, and characters per topic, by answer-box type

	(a) Small box	(b) Large box	(c) Small box & counter	(d) Dynamic box
No response	33%	33%	36%	33%
Response	66%	66%	63%	66%
Characters	92	107 [d]	101 [d]	80 [b,c]
Topics	1,3 [d]	1,3 [(c),d]	1,2 [(b)]	1,1 [a,b]
Characters per topic	74 [(c)]	84 [(d)]	89 [(a),d]	73 [(b),c]
N	162	174	161	174

Notes. Pairwise Chi^2-tests: a, b, c, d = p < .05; (a), (b), (c), (d) = p < .1. Letters denote significant differences to the indicated columns. Outliers were excluded at two standard deviations above the group mean for characters, topics and characters per topic.

In comparison to the small answer-box, we expected that the large box, the small box with the counter, and the small dynamic box would motivate respondents to provide longer answers. Results on the

proportion of item-nonresponses and the number of characters respondents typed into the open-ended questions are summarized in Table 2. Contrary to expectations, the answer-box size did not significantly influence item-nonresponse. For the small answer-box as well as for the large box, two out of three respondents (66 percent) answered the open-ended question; the larger response box did not seem to pose a higher response burden in our study. Similar results were found for the small box with a counter (63 percent), as well as the dynamic growing box (66 percent).

Results concerning the length of the responses to the narrative open-ended question were inconclusive. While the small box yielded 92 characters on average (outliers were excluded at two standard deviations above the group mean), the large answer-box version yielded an average of 107 characters. However, this difference was not statistically significant. The average length of the answers for the version with the counter was higher as well (101 characters) compared to the small box (again, not statistically significant). By contrast, the dynamic growing box produced significantly shorter responses (80 characters) in comparison to the three other conditions. This difference was statistically significant in comparison to the large answer-box and to the box with a counter ($p < .05$).

Slightly varying results were noticeable for the number of topics reported by the respondents. The number of reported topics in response to the large box did not increase as compared to the small answer-box; the number of reported topics for the counter version (1.2 topics, $p < .1$) differed at a marginally significant level; the dynamic answer-box (1.1 topics, $p < .05$) differed significantly.

There was a more conclusive result for the number of characters used per topic. Respondents to the large answer-box used more characters per topic (84)compared to the dynamic box (73 characters per topic, $p < .1$). The box with the counter produced 89 characters per topic and therefore significantly more characters per topic as compared both to

the small box (74 characters per topic, $p < .1$) and the dynamic box (73 characters per topic, $p < .05$).

8.1.3 Summary

The first experiment tested different visual design approaches to lowering the response burden and motivating respondents to provide long and rich responses. Previous studies have shown an influence of the answer-box size on data quality in narrative open-ended questions (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007). Indeed, we could demonstrate that different box sizes affect the number of characters, the number of topics, and number of characters per topic that respondents provided. Overall, in this experiment, responses to the large box yielded more characters and characters per topic that for the small answer-box. Contrary to expectations, however, no differences in item-nonresponse between small and large boxes occurred, suggesting that a large box-size did not pose a higher response burden in our study. The relatively small differences in the box size in our experiments might be an explanation for that. The fact that we experimented with a homogenous highly educated and motivated student sample might be another explanation.

In order to motivate respondents to provide high-quality answers, we examined what happened when we added a counter to a small answer-box. With a small answer-box, the response burden was assumed to be rather low, although respondents should be encouraged to provide extensive responses, given the counter continuously indicating the number of characters left. The answer-box with the counter produced slightly more characters and significantly more characters per topic in comparison to the answer-box of the same size without a counter, while no differences in item-nonresponse appeared. Therefore we assume the counter to be a valuable enhancement to any open-ended question.

The results for the dynamic growing answer-box revealed a shorter response length and fewer reported topics. This finding is especially unexpected since the dynamic answer-box should accomplish results at least comparable to the small box. We hypothesize that the resizing of the dynamic answer-box while the respondent is typing might confuse

them, and even lower their motivation to answer as they might feel misled. Overall, the dynamic box in this experiment seems to have a negative influence on data quality. Regarding the question-answer process, the dynamic box has the disadvantage that the formatting of a response has already been finished before the respondent becomes aware of additional rows in the dynamic design. Therefore, altering the box size during the typing process may not be ideal.

8.2 Experiment II: Answer-box size, counter and counter default values

The second experiment takes a closer look at the influence of the answer-box sizes and the use of counters in order to improve responses to narrative open-ended questions. Since the answer-box size in Experiment I had a rather distinctive influence on response length, we added three different answer-box sizes in this second experiment to better understand the way the answer-box size affects responses. We expected larger answer-boxes would improve the response length while at the same time increasing the burden to respond. In addition, we varied the presence and default value of the counter displayed with the answer-box. When setting different starting values for the counter, we expected smaller counter values to restrict responses, and large ones to help expand responses further, in comparison to counter values that matched the answer-box size capacity. We assume that the counter allows the researcher to limit the size of the answer-box in order to reduce item-nonresponse without curtailing the length of the responses provided.

8.2.1 Experimental design

The data were collected in August 2012, when all 2012 applicants for admission to study at Darmstadt University of Technology were invited to participate in a Web survey (before they were informed if their application was successful or not). Overall, 18.679 applicants were invited to the survey, and ultimately 5.997 respondents answered at least 50

percent of the survey questions, a response rate of 32 percent (AAPOR RR2).

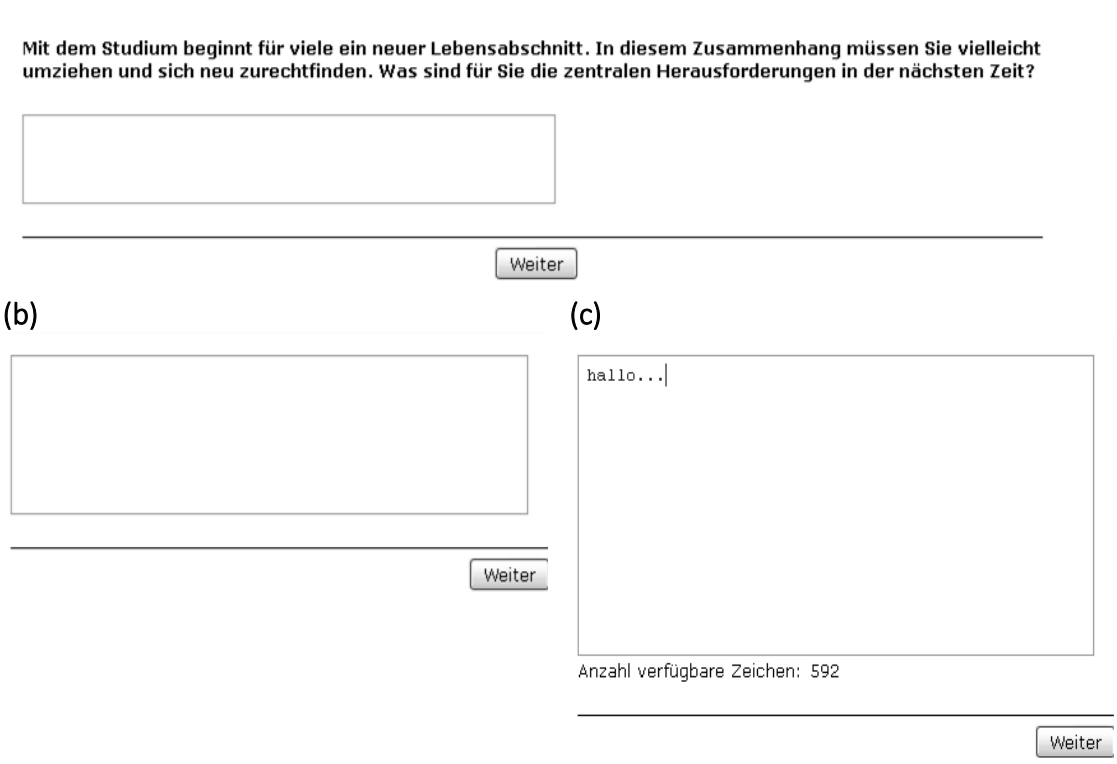
The narrative open-ended question "*When beginning your studies, you might have to relocate and find a way to manage this new situation. What do you think are the challenges for you in the near future?"* in the Web survey varied two visual design factors. The first was the answer-box size. Respondents were assigned to either a small answer-box with a capacity of 150 characters (3 rows of 50 characters), a medium answer-box measuring 300 characters (6 rows of 50 characters), and a large answer-box measuring 600 characters (12 rows of 50 characters). Secondly, the design of the counter varied as well. Respondents were assigned to an answer-box with either: no counter at all; a smaller counter default value measuring 66 percent of the number of characters of the answer-box size; a 100 percent default counter value matching the exact size of the answer-box; or a 133 percent default counter value implying a higher number of characters left than the answer-box size actually had. Overall this yielded a 3 x 4 design (see Figure 14) and each respondent was randomly assigned to one of these twelve differently designed open-ended questions.

Figure 14. Experimental 3 x 4 design (3 answer-box sizes, 4 counter designs)

	Counter design			
Answer-box size	no counter	66% counter	100% counter	133% counter
Small (150 characters)	-	100 characters counter	150 characters counter	200 characters counter
Medium (300 characters)	-	200 characters counter	300 characters counter	400 characters counter
Large (600 characters)	-	400 characters counter	600 characters counter	800 characters counter

As in the previous study, in order to isolate visual cues of the answer-box and the displayed counter, the counter was displayed after the respondent clicked into the input field. Therefore, the respondent's willingness to answer should not have been affected by the counter design. However, answer-box sizes were expected to have an influence on item-nonresponse. The actual design used within the open-ended question is displayed in Figure 15 for all three answer-box sizes, and the visual appearance of the 100 percent counter when added to a large answer-box.

Figure 15. Answer-box sizes and counter appearance: (a) small answer-box, (b) medium answer-box and (c) large answer-box and counter while typing



8.2.2 Results

Item-nonresponse

Overall open-ended questions, and especially narrative open-ended questions, suffered from higher rates of item-nonresponse (e.g. Galesic, 2006; Reja, et al., 2003; Scholz & Zuell, 2012; Smyth, et al., 2012). With enlarging the answer-box, we expected to increase the burden to respond, while at the same time the counter was expected to affect responses as it was displayed later with the respondent clicking into the answer-box in order to type in the response. The results on item-nonresponse are presented in Table 3. Overall we found no differences in item-nonresponse – whether a counter was displayed or not (counter appeared with the respondent clicking into the answer-box), nor if the answer-box varies in size. A slight tendency to less item-nonresponse in smaller answer-boxes was not significant, nor were the slightly higher item-nonresponse rates to the 66 percent and the 100 percent counter in comparison to the 133 percent and no-counter condition. Therefore, neither the answer-box size nor the counter has a significant influence on item-nonresponse in the second experiment.

Table 3. Item-nonresponse for the four counter designs

Answer-Box Size	(a) no counter	(b) 66% counter	(c) 100% counter	(d) 133% counter	Total
Small	16.4% 82	20.0% 101	18.3% 85	17.0% 81	17.9% 349
Medium	20.4% 88	21.0% 107	17.6% 86	16.1% 79	18.7% 360
Large	19.7% 97	20.2% 95	21.6% 109	19.4% 98	20.2% 399
Total	18.8% 267	20.4% 303	19.2% 280	17.5% 258	19.0% 1108

Notes. No differences overall; only slightly higher rates of item-nonresponse to larger – n.s. (p = 0.181)

Response quality

Again, we gauge response quality by the response length measured in characters used by a respondent, the topics reported, and the characters used to elaborate on each topic in Table 4 (outliers are excluded two standard deviations above the mean).

Based on previous findings (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007) we anticipated longer responses to large answer-boxes, while small answer-boxes would shorten the response length. Overall we found a significant increase in the number of characters in larger answer-boxes. This difference appears significant when no counter was used as well as for all three designs including a counter.

Despite the overall significance of an increase in the number of characters reported, the post-hoc tests revealed that for the answer-boxes with no counter, only the large answer-box (125 characters) differs significantly from the small answer-box (103 characters). With the addition of any of our three counter designs (66, 100 or 133 percent), the number of characters increases significantly from the small to the medium as well as from the medium to the large answer-box in the Scheffé post-hoc tests. Results on response length and the answer-box size are therefore consistent with existing findings and even more pronounced when adding a counter to a narrative open-ended question. Despite the influence of the answer-box size, the number of characters and number of topics is significantly different when comparing the 66, 100, 133 percent counter and the design that included no counter. Overall the number of characters and characters per topic are influenced by the varying counter designs when a small answer-box is displayed, while for the medium answer-box size the number of characters and the number of topics varied significantly. For the large answer-box, no significant differences appear between the 66, 100, 133 percent sizes and no-counter condition.

Table 4. Number of characters, topic, and characters per topic reported

Answer-Box Size		(a) no counter	(b) 66% counter	(c) 100% counter	(d) 133% counter	Total
1 Small	characters	102.9 b,(3)	72.1 a,c,d,2,3	94.2 b,(2),3	98.7 b,(3)	91.9 ***
	topics	2.1 (3)	2.0 (3)	2.2	2.1 (3)	2.1
	characters per topic	56.3 b	41.9 a,(c),d,2,3	52.0 (b),3	53.9 b,(3)	51.0 ***
	N	405	403	379	396	1583
2 Medium	characters	107.2	104.2 (d),1,3	112.4 (1),3	120.5 (b),(1),(3)	111.1 *
	topics	2.1 (c),(d),(3)	2.1 (c),(d),(3)	2.4 (b)	2.3 3, (a), (b)	2.2 ***
	characters per topic	57.5	59.3 1,(3)	53.3 3	59.7	57.5
	N	323	402	372	384	1481
3 Large	characters	124.9 (1)	135.7 1,2	141.3 1,2	138.9 1,(2)	135.2
	topics	2.4 (1),(2)	2.3 (1),(2)	2.3	2.4 (1)	2.4
	characters per topic	60.3	69.4 1,(3)	69.7 1,2	67.4 (1)	66.7
	N	377	361	375	384	1497
4 Total	characters	114.1***	102.0 ***	116.5 ***	119.1 ***	113.0
	topics	2.2***	2.1 *	2.3	2.3 *	2.2
	characters per topic	56.6	54.8 ***	56.5 ***	57.8 *	56.4
	N	1105	1166	1126	1164	4561

Notes. F-Test: Size: Characters and characters per topic p < .001; topics n.s. Design: Characters and topics p < .001; characters per topic n.s.
Scheffé Post-Hoc Tests comparing counter designs in-between answer-box sizes :
a, b, c, d = p < .001; (a),(b), (c), (d) = p < .05

Based on previous findings (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007) we anticipated longer responses to large answer-boxes, while small answer-boxes would shorten the response length. Overall we found a significant increase in the number of characters in larger answer-boxes. This difference appears significant when no counter was used as well as for all three designs including a counter. Despite the overall significance of an increase in the number of characters reported, the post-hoc tests revealed that for the answer-boxes with no counter, only the large answer-box (125 characters) differs significantly from the small answer-

box (103 characters). With the addition of any of our three counter designs (66, 100 or 133 percent), the number of characters increases significantly from the small to the medium as well as from the medium to the large answer-box in the Scheffé post-hoc tests. Results on response length and the answer-box size are therefore consistent with existing findings and even more pronounced when adding a counter to a narrative open-ended question. Despite the influence of the answer-box size, the number of characters and number of topics is significantly different when comparing the 66, 100, 133 percent counter and the design that included no counter. Overall the number of characters and characters per topic are influenced by the varying counter designs when a small answer-box is displayed, while for the medium answer-box size the number of characters and the number of topics varied significantly. For the large answer-box, no significant differences appear between the 66, 100, 133 percent sizes and no-counter condition.

When a small answer-box is combined with no counter, respondents provide longer answers, in comparison to each of the counter designs. The 66 percent counter particularly constrains significantly the response length and characters per topic when comparing it to the answer-box with no counter, the 100 and the 133 percent counter. Although the number of characters varies widely depending on the counter design added to a small answer-box, the counter does not affect the number of topics reported, which is rather constant at about 2.1 topics. Because the number of reported topics is constant, the number of characters used to elaborate each topic rises along with the number of characters provided, with respondents assigned to higher default counter values. The distinct influence of the answer-box size on the number of characters becomes more obvious when comparing the small answer-box in the 133 percent condition (98 characters) with the medium answer-box in the 66 percent condition (107 characters), in which the counter default values are identical at 200 characters.

The restricting nature of the different counter designs diminishes when comparing the response length provided for the medium answer-box. Only the 66 percent counter condition (104 characters) results in a

shorter response length than the medium box with no counter (107 characters), while the 100 percent (112 characters) and 133 percent counter (120 characters) both lead to longer responses. However, the number of characters reported for the medium answer-box only varies significantly between the 66 percent and 133 percent counter in the Scheffé post-hoc tests. Further, the number of topics reported for the medium answer-box varies significantly between the no-counter and the 66 percent counter condition (both 2.1 topics), as compared to the 100 percent (2.4 topics) and 133 percent (2.3 topics) counter condition. The identical default counter value of 600 characters in the 133 percent small and 66 percent large answer-boxes again reveals the box-size influence, with 120 characters for the medium and 135 characters for the large answer-box.

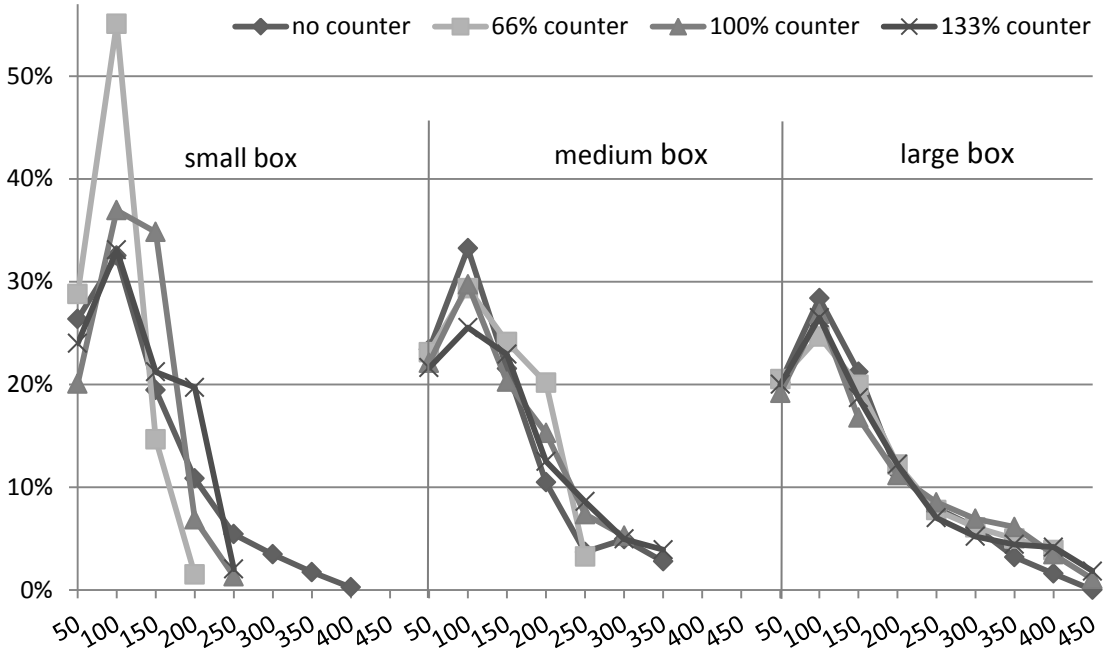
Answers for the large answer-box do not differ in the number of reported topics, with 2.3 topics reported for the 66 percent and 100 percent counters, and 2.4 topics reported for the no-counter and 133 percent counter. The number of reported topics to the large answer-box increases with the default counter value. All counter versions lead to longer responses than the no-counter design does, even though this difference is not significant. In the 66 percent counter condition, the default counter value indicates 400 characters left at the beginning, so no respondent seems to be restricted with regard to intended answer length.

Overall, counter designs shorten responses for the small answer-box, but lengthens them for the large answer-box, suggesting an interaction between answer-box sizes and the counter design. The graph in Figure 16 shows the number of characters reported recapped into blocks of 50 characters for the small, medium or large answer-box size, depending on the design of the counter.

Even though respondents were not limited by the counter and could write even more if they wanted to, whenever the default value of a counter is accomplished, the percentage of respondents continuing their answer drops dramatically. Therefore, all three counter designs added to

the small answer-box cannot resemble the response length of the small answer-box with no counter added.

Figure 16. Number of characters reported



The restrictive character of the counter indicates that the provided response length is not on par with the intended length. For the medium answer-box, the limiting factor of the counter manifests, especially for the 66 percent, and less influential for the 100 percent counter. Overall the differences between the counter designs are smaller when comparing the response length provided to the medium answer-box. The default counter values added to the large answer-box no longer have a negative restricting influence on the response length, and in contrast they even help to improve the response length. The percentage of respondents providing more than 250 characters for the large answer-box is higher, in all three counter conditions, compared to the no-counter condition.

Personal characteristics

The closer look at information processing and the factors contributing to satisficing described in chapter 4 revealed the importance of question difficulty and respondents' motivation. While the visual design alterations tested in this thesis attempt to address these two components of the satisficing theory (Krosnick, 1991; Krosnick, et al., 1996), personal abilities cannot be affected by a researcher. In addition, the respondents' skills and abilities to formulate and type a response vary, and narrative-open-ended question will at least to some extent reflect and measure these abilities (Geer, 1988, 1991). That the visual design affects responses has been demonstrated in the preceding analysis, but we also expect personal characteristics to have an influence on the willingness to respond to a narrative open-ended question, as well as on the response length (e.g. Denscombe, 2008; Oudejans & Christian, 2010; Smyth, et al., 2012; Stern, et al., 2007). Since we used a highly educated student sample in this study, the influence of personal characteristics and abilities should be minor in comparison to a survey of the general population, where the variance in personal characteristics is presumed to be higher. A closer look at the influence of personal characteristics on item-nonresponse is analyzed via logistic regression analysis in Table 5.

While we found no differences in item-nonresponse in the prior analysis, the regression Model 1 reveals an influence by the answer-box size on item-nonresponse. The likelihood of a respondent answering the large answer-box is significantly lower than answering the small answer-box (odds ratio = .83, $p < .05$). The medium answer-box shows no differences in comparison to the small answer-box as well as the use of a counter. Overall the visual designs do not have a major influence, which is emphasized by a very small Nagelkerkes R^2 in Model 1.

With controlling for personal characteristics on the willingness to respond, in Model 2, personal characteristics do show an influence on item-nonresponse. Since grades in German schools are rated from 1 (very good) to 6 (failing), the math grade has a positive effect on item-nonresponse so that respondents with poorer algebra skills are more

likely to answer the open-ended question in the experiment (odds ratio = 1.26, p < .01). In contrast to this result, the Final grade has the exact opposite effect, with poorer Final grades going along with fewer responses, while the grade in German has no influence on item-nonresponse (odds ratio = .71, p < .01).

Table 5: Logistic regression of visual design manipulation and personal characteristics on the likelihood to respond

	Model 1		Model 3		Model 3	
	Odds	Sig.	Beta	Sig.	Beta	Sig.
Visual Design						
Medium (small)	0.93	0.38			0.95	0.61
Large (small)	0.83*	0.02			0.86	0.10
Counter (no counter)	0.99	0.94				
Personal characteristics						
Math grade			1.26**	0.00	1.28**	0.00
German grade			0.96	0.54		
Final grade			0.71**	0.00	0.68**	0.00
Gender (male)			0.82*	0.01	0.82**	0.00
Number of surveys			1.01**	0.00	1.01**	0.00
Computer skills (low)			0.71*	0.01	0.83	0.07
Internet skills (low)			1.24	0.07		
Constant	4.81**	0.00	7.59**	0.00	8.13**	0.00
Nagelkerkes R-squared	.001		0.027		0.029	

*Notes. **.01 *.05 Wald-Test*
N=5454
Reference categories in brackets
Grades 1 = very good, 2 = good, 3 = satisfactory, 4 = sufficient, 5= poor, 6 = fail

Female respondents leave the open-ended question unanswered more often in the study (odds ratio = .82, p < .05), while the number of surveys taken previously significantly heightens the chance of respondents providing a response (odds ratio = 1.02, p < .01). Respondents' self-reported proficiency in computer skills lowers the chance of response (odds ratio = .71, p < .05), while Internet skills do not affect whether

respondents answer the open-ended question or not. Age as a control variable was excluded from the analysis based on the fact that the student samples do not vary widely in age. Nagelkerkes R^2 in Model 2 is somewhat higher, suggesting a broader influence of personal characteristics on item-nonresponse, in comparison to the visual design manipulations. This hierarchy becomes even more apparent in the combined Model 3, where all personal characteristics variables continue to be significant in the same way as in Model 2, but with the answer-box size no longer significantly affecting item-nonresponse.

Despite the influence of personal characteristics on item-nonresponse, these characteristics are supposed to influence the responses as well (e.g. Denscombe, 2008; Oudejans & Christian, 2010; Smyth, et al., 2012; Stern, et al., 2007). The regression analysis on the number of characters reported in Table 6 shows the influence of visual design manipulations and personal characteristics on the number of characters reported.

In Model 1 the influence of the answer-box size (small answer-box with no counter is the reference category) and the presence of a counter are displayed. While the medium (Beta =19.02, $p < .05$) and the large answer-box (Beta =43.79, $p < .05$) have a positive influence on the number of characters reported in comparison to the small answer-box, the counter does not have a significant influence on the response length. Based on previous findings indicating an interaction between the answer-box size and the presence of a counter, an interaction term is integrated into our regression analysis in Model 2. Those respondents assigned to a small answer-box with a counter (small*counter) provide shorter responses, while the assignment to a large answer-box without a counter does not significantly influence the response length. Overall the visual design manipulations included in Model 3 explain 5.7 percent of the variance (R^2 = 0.057) of the response length, while the personal characteristics included in Model 4 only account for 1.2 percent of the variance (R^2 = 0.012) of the response length.

Table 6: Linear regression of visual design manipulation and personal characteristics on the number of characters provided (outlier excluded)

	Model 1		Model 2		Model 3		Model 4	
	Beta	Sig.	Beta	Sig.	Beta	Sig.	Beta	Sig.
Visual Design								
Medium (small)	19.02 **	0.00	3.18	0.56			3.57	0.53
Large (small)	43.79 **	0.00	23.00 **	0.00			21.46 **	0.00
Counter	0.68	0.79	5.99	0.20			4.82	0.32
Small*counter			-21.00 **	0.00			-19.97 **	0.00
Large*counter			6.82	0.29			9.59	0.15
Personal characteristics								
Math grade					4.24 **	0.00	3.47 **	0.00
German grade					-3.23 **	0.07	-3.97 **	0.00
Final grade					-2.92	0.33		
Gender (male)					9.65 **	0.00	-9.01 **	0.00
Number of surveys					0.15	0.18		
Computer skills (low)					-9.11 *	0.04	-9.10 **	0.00
Internet skills (low)					-1.06	0.77		
Constant	89.28 **	0.00	100.97 **	0.00	112.27 **	0.00	100.38 **	0.00
R-squared	.053		.057		0.013		0.069	

*Notes. **.01 *.05 T-Test*
N=4341
Reference categories in brackets
Grades 1 = very good, 2 = good, 3 = satisfactory, 4 = sufficient, 5 = poor, 6 = fail

The math grade (1 = very good to 6 = very poor) has a significant positive influence on the response length, indicating that respondents with poor algebra skills report longer responses (Beta = 4.24, $p < .01$), whereas German language skills (German grade) improve the response length significantly (Beta = -3.97, $p < .01$). No influence of the Final grade appears to be significant, nor does the number of surveys taken and the self-reported proficiency in Internet skills. However, computer skills significantly decrease the response length (Beta = -9.1, $p < .01$) as well as

the gender implying shorter answers by male respondents (Beta = -9.97, $p < .01$). Finally, Model 4 combines visual design manipulations and personal characteristics with only minor differences in comparison to the prior separated models, resulting in an overall variance explanation of 6.9 percent ($R^2 = 0.069$). Thus large answer-boxes produce longer responses than small answer-boxes and whatever counter is added to a small answer-box shortens responses. Even though personal characteristics such as math and German grades, gender, and computer skills contribute to the response length, the variance explanation of the visual design manipulations is somewhat higher.

8.2.3 Summary experiment II

The second experiment varied the size of the answer-box, the presence of a counter, and the default value of a counter. While narrative open-ended questions are prone to higher rates of item-nonresponse (e.g. Galesic, 2006; Reja, et al., 2003; Scholz & Zuell, 2012; Smyth, et al., 2012), we found no differences in item-nonresponse. This result holds true whether or not a counter is displayed with the answer-box, or if the answer-box varies in size. However, corresponding to prior findings (Experiment I) and the literature (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007), large answer-boxes yield longer responses in comparison to medium or small answer-boxes. At the same time, respondents used the additional space to elaborate on their response, rather than to report more topics.

Differences in the number of topics between the counter designs were also minor and the number of characters per topic only varied between the counter designs added to the small answer-box. Our analysis revealed an interaction between the answer-box size and the default value of the counter. Especially when combined with a small answer-box, the counter limited the response length in comparison to an answer-box without a counter. For the medium answer-box, the counter helped to increase the number of characters and topics reported only slightly in comparison to the large answer-box, where every counter design exceeds the standard design in the response length.

One interesting finding is that this restricting counter limits the number of words used by respondents, but not the number of topics reported. Even though the number of topics was not compromised by a low counter starting value, we expected the formatting process to be more burdensome for respondents since they have to shorten single words and use abbreviations in order to provide their response. In addition, decoding these often-cryptic formatted responses is not easy, and sometimes even impossible. Therefore, counter values should at least match, if not exceed, the answer-box size.

Based on these results, the provided answer-box should not be too small, and a counter should always at least match the number of characters fitting in the answer-space. Setting the default counter value higher than the box size will increase the response length even further.

The question–answer process and likewise the answer to a narrative open-ended question is driven by the difficulty of the task, the respondent's motivation, and the respondent's abilities. Visual design should help improve responses especially by motivating respondents to provide long and rich responses. But despite the influence of visual design and motivation, narrative open-ended questions remain burdensome to answer. Personal characteristics and abilities have a strong influence on item-nonresponse as well as on the response length, even though a highly-educated sample was used for this experiment. In general population surveys, personal characteristics will influence narrative open-ended questions even more.

Whatever differences occur between the responses provided to a narrative open-ended question, it is important to point out that every single open-ended question is unique and that the visual design can have different implications depending on the question it is applied to. The limiting factor of the counter depends heavily on the intended response length evoked by the question itself. Therefore, if a counter has a limiting influence on one question, it doesn't necessarily have the same influence on another open-ended question as well. To put the question used in the second experiment into perspective, all applied narrative open-ended

questions asked in the Web survey are displayed in order of appearance in Table 7.

Table 7: Narrative open-ended questions in the questionnaire

Rank	Question wording	Nonresponse rate	Number of characters
1	What other reasons were important to you in applying for studying at a university instead of vocational training or starting a career?	14%	151.8
2	In what situations have you felt overwhelmed or unable to cope with a situation within the last 6 months?	20%	114.8
3	Please name the reasons you apply to study at Darmstadt University of Technology.	19%	113.0
4	When beginning your studies, you might have to relocate and find a way to manage your new situation. What do you think are the challenges for you in the near future?	12%	108.2
5	What personal experiences of success have you already achieved in your educational and professional career?	20%	160.4

Notes. Outliers were excluded at two standard deviations above the group mean for characters.

Overall item-nonresponse rates vary between 12 and 20 percent, and demonstrate neither effects from response fatigue, nor any order effect based on the narrative open-ended questions' positions in the questionnaire. Despite the varying willingness to respond to questions, the number of characters used for the responses differs as well. The narrative question “When beginning your studies, you might have to relocate and find a way to manage your new situation. What do you think are the challenges for you in the near future?” used in the second experiment had the highest item-response rate, but at the same time respondents typed in fewer characters. The third question, asking for the reason a respondent applied to study at Darmstadt University of Technology, had an item-nonresponse rate of 19 percent, and respondents typed in 113 characters on average. This third question is

used in the following third experiment, and used the same answer-box sizes as the second experiment, while all other narrative open-ended questions used a larger answer-box (6 rows of 73 characters). For that reason, the longer responses to the other narrative open-ended question in the survey are to some extent a result of the larger answer-box sizes. The differences in item-nonresponse and the number of characters reported are minor, and the findings of this study will most likely apply to other narrative open-ended questions as well.

8.3 Experiment III: Answer-box size, dynamic, and respondent-adjusted answer-boxes

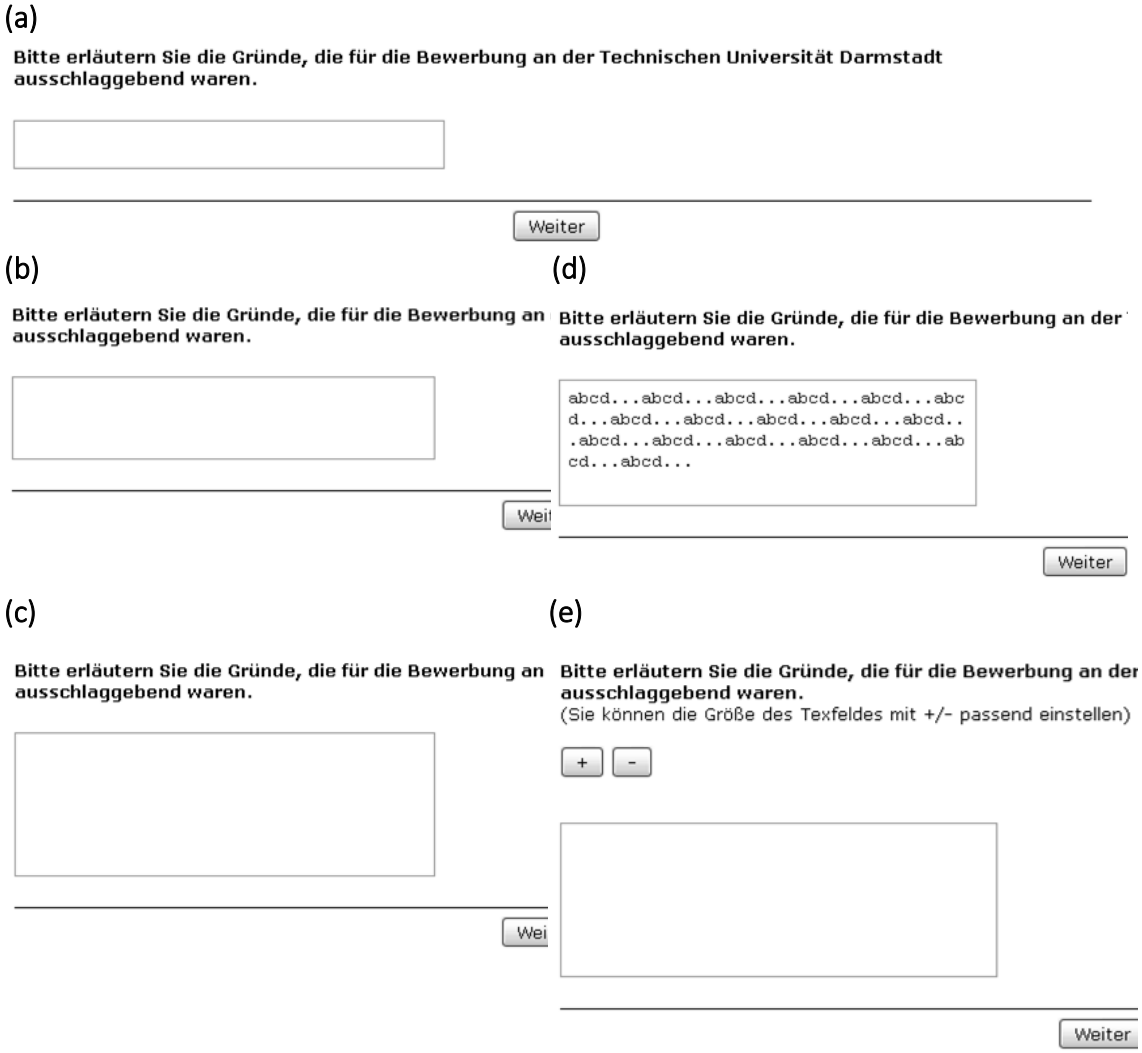
The third experiment again tested the influence of answer-box sizes on the responses to narrative open-ended questions using the same answer-box sizes as in the second experiment. But instead of varying the presence and the default value of a counter, respondents were randomly assigned to a static answer-box, a dynamic auto-adjusting answer-box, or a self-adjusting answer-box. In the auto-adjusting box design, an additional empty line was added at the end of the answer-box each time the respondent completed a line of text. In the self-adjusting design, the respondents themselves could adjust the box size using a plus and a minus button. Both adjustable box designs allowed smaller initial box sizes, implying a lower response burden while at the same time motivating respondents to provide longer answers.

8.3.1 Experimental design

The third experiment was also embedded in the survey of 2012 applicants to Darmstadt University of Technology Web, like the previous experiment on the counter design. Again, the response rate was 32 percent, and 5.997 respondents completed at least 50 percent of the questionnaire's questions. Respondents were randomly assigned to three different versions of the same open-ended question in a between-subjects design asking: *"Please name the reasons you apply to study at Darmstadt University of Technology?"* Again, the answer-box size was

varied, and respondents were assigned to either a small answer-box measuring 150 characters (3 rows of 50 characters), a medium answer-box measuring 300 characters (6 rows of 50 characters) and a large answer-box measuring 600 characters (12 rows of 50 characters). Instead adding a counter and varying its default values, this experiment varied the design of the answer-box itself. Respondents were either assigned to standard answer-box (a); an auto-adjusting answer-box where after every completed line an additional row was added to answer-box (b); or an answer-box where respondents could alter the box sizes themselves by pressing a plus or minus button (c) (see Figure 17).

Figure 17. Experimental answer-box designs: (a) small standard design, (b) medium standard design, (c) large standard design, (d) dynamic auto-adjusting when typed (medium size), (e) respondent-adjusted design (default large size).



8.3.2 Results

We carried out the same analysis as in the prior experiments, examining the influence of the visual design on item-nonresponse first. Again, no differences in item-nonresponse appeared between the different visual designs. The large answer-box again showed higher item-nonresponse rates in comparison to the small and the medium answer-boxes (see Table 8). However, this difference is not significant like the difference between the slightly higher item-nonresponse of the respondent-adjusted answer-box, in comparison to the dynamic and standard answer-box design.

Table 8. Item-nonresponse for the standard, dynamic, and respondent-adjusted answer-box design for small, medium and large answer-boxes

Answer-Box Size	(a) standard	(b) dynamic	(c) respondent-adjusted	Total
Small	9.7%	13.1%	12.5%	11.8%
	65	86	80	231
Medium	10.7%	9.9%	13.0%	11.2%
	68	65	83	216
Large	13.2%	12.2%	14.0%	13.1%
	94	83	97	274
Total	11.3%	11.7%	13.2%	12.1%
	227	234	260	721

Notes. n.s.

Results for the number of characters, topics, and characters per topic are displayed in Table 9. As in the prior experiments, the answer-box size influenced the response length. Overall, large answer-boxes led to more reported characters. In the standard design, the large answer-box (108 characters) differs only significantly from the small answer-box (95 characters), while in the dynamic box design, the small answer-box (96 characters) differs significantly from the medium (114 characters) and large answer-boxes (113 characters). In the design, where respondents adjusted the answer-box size themselves, the small (103 characters) and medium (108 characters) answer-box sizes differ significantly as compared to the large answer-box (132 characters). Even though there is

a tendency for the dynamic answer-box design to evoke longer responses in comparison to the standard answer-box, these differences are not significant. The only significant difference appears when comparing the answers provided in the large answer-box, where the respondent-adjusted design (132 characters) exceeds the response length of both other designs significantly.

Table 9. Characters, topics, and characters per topic reported to the varying answer-box sizes and visual designs

Answer-Box Size		(a) Standard	(b) Dynamic	(c) Respondent-adjusted	Total
1 Small	characters	94.7 (3)	96.4 (2),(3)	103.2 3	98.0
	topics	2.1 c	2.1 c	2.4 a,b	2.2 ***
	characters per topic	50.2 (3)	50.2 (2),(3)	47.8	49.4
	N	561	543	522	1626
2 Medium	characters	102.9	113.7 (1)	108.3 3	108.4
	topics	2.1 c	2.1 (c)	2.4	2.2 ***
	characters per topic	52.5 (3)	58.3 (1),(c)	50.6 3,(b)	53.9 **
	N	532	564	519	1615
3 Large	characters	108.2 (1), c	113.3 (1),(2),(c)	131.6 1,2,a,(b)	117.7 ***
	topics	2.1 c	2.2 a, (b)	2.4 a,b	2.2 ***
	characters per topic	58.9 (2),(3)	58.8 (1),(3)	62.2 (1),(2)	60.7
	N	578	563	577	1718
4 Total	characters	102.0 **	108.0 ***	114.9 ***	108.2
	topics	2.1	2.1 *	2.4	2.2
	characters per topic	53.9 **	55.8 **	53.8 ***	56.4
	N	1671	1670	1618	4959

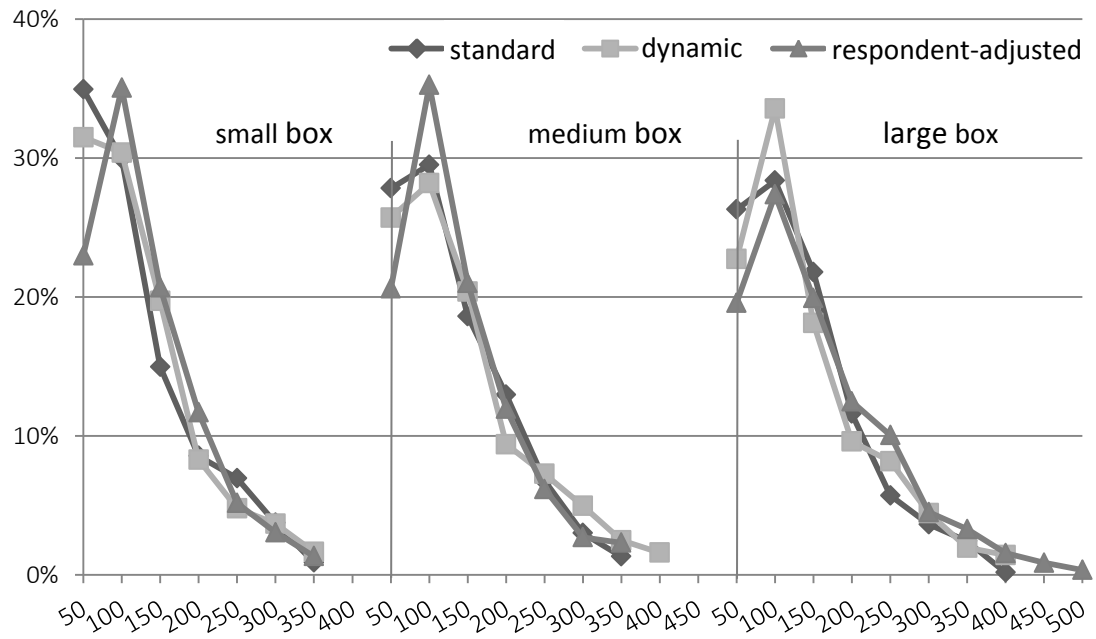
F-Test: Size: Characters and characters per topic p < .001; topics n.s. Design: Characters and topics p < .001; characters per topic n.s.
Scheffé Post-Hoc Tests comparing counter designs in-between answer-box sizes :
a, b, c, d = p < .001; (a),(b), (c), (d) = p < .05

The number of topics reported is stable across the varied answer-box sizes. Only within the dynamic box design was the number of topics reported significantly higher when comparing the small (2.1 topics) and

medium (2.1 topics) box sizes to the large answer-box (2.2 topics). When comparing the three designs for all three different answer-box sizes, the self-adjusting design led to significantly more reported topics than the standard and the dynamic answer-box designs.

Elaboration, measured by characters per topic, varied significantly among the answer-box sizes. While characters per topic do not vary between the small answer-box and the medium answer-box, the dynamic design (58 characters per topic) resulted in significantly more elaborated answers in comparison to the respondent-adjusted design (51 characters per topic).As with the small answer-box, the number of characters to elaborate each topic in the large answer-box did not vary significantly between the standard, dynamic, and respondent-adjusted designs.

Figure 18. Number of characters reported



When looking at the number of characters reported, recapped into blocks of 50 characters (Figure 18), the differences between the three designs appear to be less prominent and influential in comparison to the results of the previous counter experiment. The small respondent-adjusted answer-box leads to fewer short responses in comparison to the standard and the dynamic answer-box designs. The same pattern

appears for the answers provided to the medium and large answer-boxes, with the respondent-adjusted design leading to fewer short responses (50 characters or less). While the dynamic box design led to slightly longer responses for the medium answer-boxes, the respondent-adjusted design shows this tendency when combined with a large answer-box. Even though outliers were excluded two standard deviations above the mean, several respondents in the large adjustable answer-box setting continued writing more than in the standard and dynamic answer-box design.

Personal characteristics

Despite the rather strong influence of the respondent-adjusted resizing answer-box design, we again expected the influence of personal characteristics on the responses provided. The willingness to respond is again analyzed via logistic regression in Table 10. Visual design manipulations included in the first Model 1 do not affect item-nonresponse.

Even though offering buttons with the resize design seem to provoke more respondents to leave the open-ended question unanswered, this tendency is not significant (p=0.06). Overall, the visual design alterations have no positive or negative effect on item-nonresponse. The personal characteristics displayed in Model 2 are related to the willingness to respond. Good algebra skills correlate with lowering the chance of responding to the narrative open-ended questions (odds ratio = 1.15, $p < .05$), while good German language skills correlate with heightening the chance of responding (odds ratio = 0.84, $p < .01$). Being a female respondent (odds ratio = .68, $p < .01$) affects response rates negatively, while the number of surveys taken had a positive effect (odds ratio = 1.02, $p < .01$). The findings are congruent with the prior Experiment II. The only difference here is the significant influence of the German language grade, while in the prior experiment the final grade influenced the willingness to respond.

Table 10. Logistic regression analysis on the willingness to respond

	Model 1		Model 2		Model 3	
	Odds	Sig.	Odds	Sig.	Odds	Sig.
Visual Design						
Med(small)	1.06	.58			0.98	.89
Large(small)	0.88	.20			0.86	.15
Dynamic	0.95	.62			0.90	.38
Respondent-adjusted	0.83	.06			0.79*	.03
Personal characteristics						
Math grade			1.15*	.01	1.13**	.00
German grade			0.84**	.00	0.83**	.00
Final grade			0.95	.63		
Gender (male)			0.68**	.00	0.69**	.00
Number of surveys			1.02**	.00	1.02**	.00
Computer skills (low)			1.10	.55		
Internet skills (low)			0.89	.42		
Constant	8.10**	.00	13.03**	.00	14.24**	.00
Nagelkerke R-squared	0.002		0.026		0.027	

*Notes. <**.01 <*.05 Wald-Test*
N=5474
Reference categories in brackets
Grades 1 = very good, 2 = good, 3 = satisfactory, 4 = sufficient, 5 = poor, 6 = fail

In combining visual design variables and personal characteristics in Model 3, the negative influence of the respondent adjusted-answer-box design now appears to be significant. However, whether respondents answered narrative open-ended questions predominantly depended on their personal characteristics, rather than the visual presentation of the question (also indicated by differences in Nagelkerkes R-squared).

However, the way that respondents answer a narrative open-ended question depends on visual design. A linear regression analysis carried out (shown in Table 11) on the number of characters reported reveal the

influence of visual design manipulations and personal characteristics of those respondents who answered the narrative open-ended question. In Model 1, the influence of the answer-box size, the auto adjusting dynamic, and the respondent-adjusted design are analyzed (small standard design is the reference category).

Table 11. Linear regression on the number of characters reported

	Model 1		Model 2		Model 3	
	Beta	Sig.	Beta	Sig.	Beta	Sig.
Visual Design						
Med (small)	10.48**	.00			10.37**	.00
Large (small)	19.68**	.00			19.40**	.00
Dynamic	6.56*	.01			7.05*	.01
Respondent-adjusted	13.53**	.00			13.14**	.00
Personal characteristics						
Math grade			3.79**	.00	3.37**	.00
German grade			-1.88	.27		
Final grade			-0.06	.98		
Gender (male)			3.48	.14		
Number of surveys			0.02	.83		
Computer skills (low)			0.31	.94		
Internet skills (low)			-1.17	.74		
Constant	90.23**	.00	101.90	.00	82.71**	.00
R-squared	0.016		0.004		0.018	

*Notes. **.01 *.05 T-Test*
N=4570
Reference categories in brackets
Grades 1 =very good, 2=good, 3=satisfactory, 4=sufficient, 5=poor, 6=fail

The positive influence of the answer-box size on the response length is again significant for the medium answer-box, eliciting 10 characters more (Beta = 10.48, $p < .01$), and the large answer-box, which elicits 20 characters more (Beta = 19.68, $p < .01$) than the small standard answer-

box design. Despite the answer-box size, the two tested visual design manipulations also enhanced the number of characters reported. Overall the dynamic answer-box design produces longer responses in comparison to the standard design (Beta = 6.56, $p < .05$), while the respondent-adjusting answer-box design enhances the response length by 14 characters (Beta = 13.53, $p < .01$). Taking a closer look at the personal characteristics in Model 2, only the math grade significantly influences the number of reported characters (Beta = 3.79, $p < .01$), and the personal characteristics only account for 0.4 percent of the variance (R^2=0.004), while the different visual designs included in the first model accommodate 1.6 percent of the variance of the number of characters reported (R^2=0.016). Again, the third model combines visual design variables as well as the significant math grade, explaining 1.8 percent of the variance in the response length (R^2=0.018). Overall, the response length is affected by the visual design, while personal characteristics do not seem to be as important.

8.3.3 Summary experiment III

In this third experiment the answer-box size again did not influence item-nonresponse. The other designs tested also had no significant influence on item-nonresponse, except a slight tendency to higher item-nonresponse when the respondents could set the size of an answer-box themselves.

As in the previous experiments, the answer-box size affected the response length, and responses were more elaborate since the number of topics remained constant across the varying answer-box sizes. While in the first experiment, the dynamically growing answer-box shortened responses, in this experiment, the dynamic growing answer-boxes at least showed a tendency to improve the response length in comparison to the standard answer-box design. Considering results from the first experiment, this study found no evidence to support the use of dynamic answer-boxes. As already mentioned, the resizing of the dynamic answer-box may be confusing to respondents and even lower their motivation to answer. The dynamic design also does not correspond to

the question-answer process, since the answer-box size is increased after a respondent has already formatted their response.

The respondent-adjusted answer-box design lets respondents decide on the answer-box size. In comparison to the dynamic design, this approach is more in line with the question-answer process. And indeed, the respondent-adjusted design improved responses as compared to both other designs tested. Respondents reported more topics and produced longer responses. The only weakness of the respondent-adjusted design is that the answer-box adjustment adds to the overall burden to respond, since respondents have to read the instructions on the box-size adjustment and also carry it out. The not-significant but slightly higher item-nonresponses on the respondent-adjusted box design indicate the added burden. Overall, the respondent adjusted answer-box seems to be useful tool to integrate into narrative open-ended questions in Web surveys.

The closer look at personal characteristics in this third experiment illustrated that personal characteristics influence responses to narrative open-ended questions. Again, we found a strong effect of personal characteristics on item-nonresponse. School grades and the respondent's gender influenced the likelihood to respond, while the effects of these traits on response length was minor. In this experiment, personal characteristics seemed to influence whether an open-ended questions is answered, but not how it is answered. However, the highly-educated sample used for this experiment suggests that in a general population survey these differences will be far more pronounced.

8.4 Discussion of the visual design experiments

The first three experiments in this thesis focused on ways to further improve the visual design of narrative open-ended questions. Four visual design variations were put to a test. First we increased the answer-box sizes to improve responses. All three experiments confirmed the results of prior studies (e.g. Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007), finding that increasing the answer space will result in longer and richer

responses. In our experiments, respondents used the additional space to elaborate on their responses rather than to report more topics. While we expected larger answer-boxes to pose an additional burden, we found no influence of the answer-box size on item-nonresponse. Thus, when designing a narrative open-ended question, answer-boxes should be sized larger than smaller.

Second, we tested the use of a counter that indicated the number of characters left. When combined with a small answer-box, the counter limited response length. With larger answer-boxes, the counter helped improve the number of characters and topics reported. Based on these results, a counter should always at least match the number of characters fitting to an answer-space. Setting the default counter value higher than the box size will increase the response length even more. One interesting finding is that a restricting counter only limits the number of words used by respondents, but not the number of topics reported. Respondents, therefore, always seem to report what they intended to report. This leads us to consider the counter a valuable enhancement to any open-ended question.

Third, we tested dynamic answer-boxes that automatically grow as the respondent types an answer. In the first experiment, the dynamic answer-box design curtailed responses, while in the third experiment the dynamic growing answer-boxes at least showed a tendency to enhance response length in comparison to the standard answer-box design. Overall, the appearance of additional rows in the dynamic design seems to occur too late in the response process and may confuse respondents. The dynamic design also does not correspond to the question-answer process since the answer-box size is increased after the respondent has already formatted the response. Based on the experiments carried out, the dynamic answer-box design does not improve responses and data quality and should not be considered as a viable answer-box design for a Web survey.

The fourth visual design tested permits the respondent to alter the box sizes themselves by pressing a plus or minus button. In the respondent-

adjusted design, respondents were able to set the answer space size upfront, according to the question–answer process, before a response is formatted. With this design, respondents addressed more topics and produced longer responses. But the answer-box adjustment added to the overall response-burden, resulting in slightly higher item-nonresponses. However, since this difference in item-nonresponse was not of significant effect, the respondent-adjusted answer-box design seems to be reasonable for use in narrative open-ended questions in Web surveys.

Open-ended questions come under criticism for their higher response-burden. The visual design cannot reduce this in the first place. None of the designs tested get more respondents to answer narrative open-ended questions. When looking at the influence of personal characteristics on willingness to respond, the analysis showed an influence of personal characteristics across both experiments (II and III). Female respondents as well as respondents with low algebra skills answered narrative open-ended questions less frequently. While the number of characters reported was affected by personal characteristics (math grade, gender, computer skills) in the second experiment, only the math grade had a significant influence on response length in both experiments, with respondents writing longer responses if they had received a bad math grade. Overall, personal characteristics seem to influence the willingness to respond, while their influence on the response length is less pronounced. The use of a student sample might even lessen the influence of personal characteristics. Students are a highly educated population and may be more able and willing to answer narrative open-ended questions. Whether these results will hold for Web surveys in the general population remains to be tested.

9 USING ADAPTIVE DESIGN TO ENHANCE NARRATIVE OPEN-ENDED QUESTIONS

The prior experiments varied the visual appearance of answer-boxes in order to obtain more detailed, elaborate, and rich responses in narrative open-ended questions. Visual appearance, especially the answer-box size, had a strong influence on the number of characters and topics reported in narrative open-ended questions. The other designs tested also affected responses to open-ended questions. Results for the dynamic growing answer-box in the first experiment, especially, suggested that varying the visual appearance of answer-spaces during the question–answer process is not ideal. Even though the visual design improved responses to open-ended questions, its influence on item-nonresponse was only minor. Thus, the higher response burden remains one of the largest problems when using narrative open-ended questions, especially when these missing responses induce a bias, which is suggested, to some extent, by the influence of personal characteristics on willingness to respond.

The second area investigated for improving narrative open-ended questions in this thesis is adaptive design. For now, Web surveys use answers provided predominantly to make skip patterns easy. A new approach would be the use of prior responses in order to adapt the questionnaire to different groups of respondents. The following two experiments make use of this approach by using narrative open-ended questions in the questionnaire to assign respondents individually to a specific answer-box size in Experiment IV, or a closed probe in Experiment V.

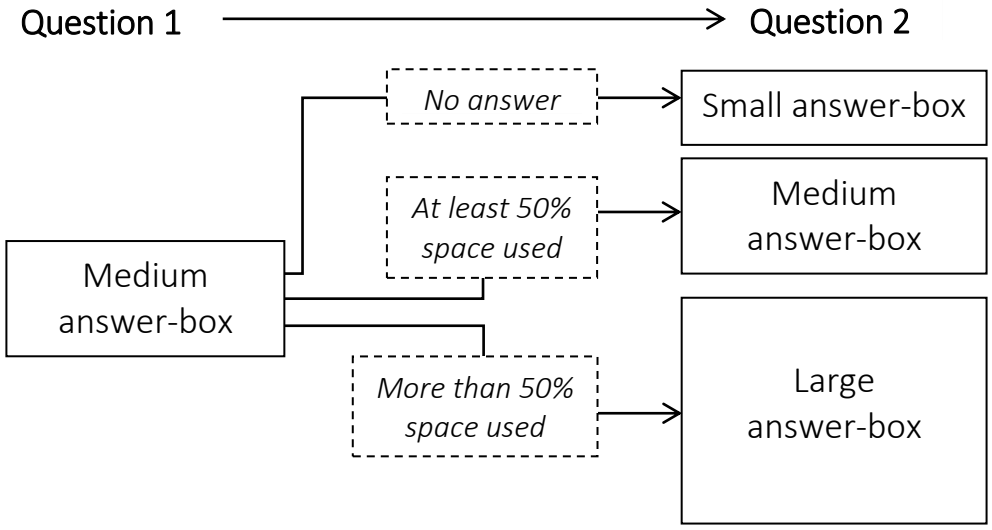
9.1 Adapting individual answer-box sizes

The answer-box size has an extensive influence on the responses provided in the first three experiments reported here, as well as in the literature (e.g. Christian & Dillman, 2004; Israel, 2010; Smyth, et al., 2009; Stern, et al., 2007). As noted earlier, respondents use answer-box size as auxiliary information when they interpret the question. Small

answer-boxes suggest a short response, and lower the response burden in comparison to a large answer-box, where the ideal response is expected to be longer and therefore the burden to respond somewhat higher. Even though prior experiments showed only a slight tendency of larger answer-boxes to result in higher item-nonresponse, the higher burden of large answer spaces should remain apparent. Finding the ideal answer-box sizes always seems like a trade-off requiring a balance between item-nonresponse and response quality.

Instead of a one-size-fits-all perspective, answer-boxes can be individually adapted based on the information provided to preceding open-ended questions (see experimental design in Figure 19). Respondents who did not answer prior open-ended questions would be assigned to a small answer-box in the adaptive design, to lower the response burden and to encourage more former non-respondents to answer the question. The additional responses captured by using adaptive design might be short, but at least some information would be gathered.

Figure 19. Adaptive answer-box design



Respondents who used almost the whole answer space and provided extensive information in preceding open-ended questions would be assigned to a large answer-box in the adaptive design, to give those respondents even more space to elaborate their point of view. Respondents who used up to the half of the space provided for the first answer-box are considered neither non-respondents nor prolific writers, and would be assigned to the same box size again for the second question. By using an adaptive assignment of answer-box sizes for groups of respondents, we expected to reduce item-nonresponse for those respondents who typically would not answer the question, and at the same time encourage prolific writers to provide even more detailed answers. Overall, the adaptive box design with customized answer-box sizes is presumed to reduce item-nonresponse, and increase the quantity and quality of responses to narrative open-ended questions.

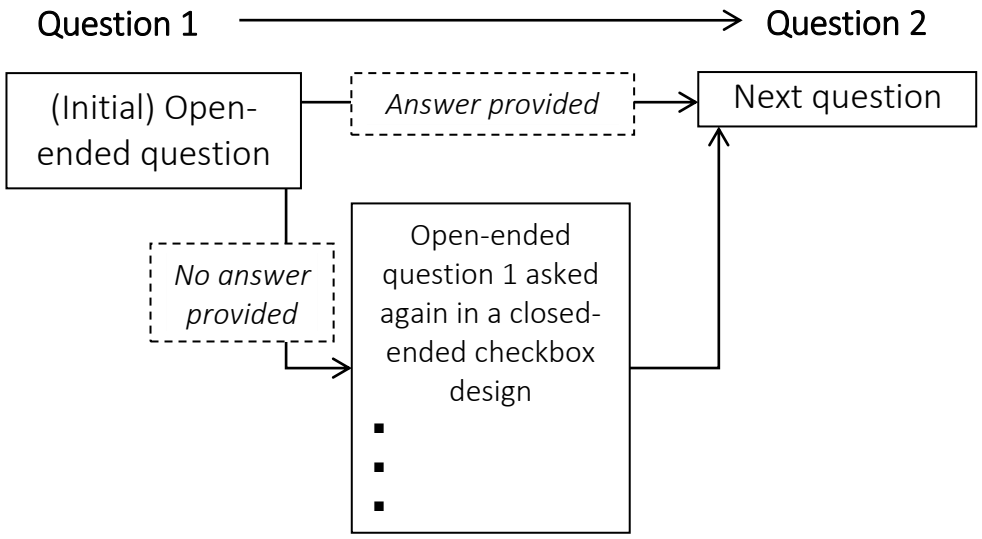
9.2 Closed probes to narrative open-ended questions

Despite using varying box-sizes, counters, self- and auto-adapting answer-boxes in order to improve response quality, many respondents still leave narrative open-ended questions unanswered (e.g. Galesic, 2006; Reja, et al., 2003; Scholz & Zuell, 2012; Smyth, et al., 2012). The response burden remains one of the major disadvantages of narrative open-ended questions. Even though prior experiments used highly-educated student samples, item-nonresponse in narrative open-ended questions still accounts for 12 to 20 percent of the total (see Table 7 in chapter 8.2.3). Using prior response to open-ended questions to assign respondents to custom-sized answer spaces is one approach to getting more respondents to answer a narrative question.

Motivational statements can help improve response rates to narrative open-ended questions just by quoting the importance of an open-ended question for the survey, as well as its purpose (Israel, 2013; Smyth, et al., 2009). Follow-up probes can be a valuable tool for improving responses to narrative open-ended questions. Web surveys allow interaction with respondents, and probing could be used similarly to the way that

interviewers probe for additional information in face-to-face or telephone surveys (Holland & Christian, 2009). Oudejans and Christian (2010) used the interactive nature of the Internet in combination with motivational statements and follow-up probes, to improve the data quality of narrative open-ended questions. The motivational statement again stressed the importance of the questions; it increased response rates and the number of topics in some questions. The follow-up probe was asked immediately after the narrative open-ended question. The probe displayed the respondents' answer to the initial question and asked if the subject would like to add anything to their previous response (which was displayed along the answer-box for their response to the probe). Only a few respondents provided a response to the follow-up probe, but overall probing increased the number of words and topics reported. Motivational prompts and follow-up probes might therefore be a viable tactic for improving responses.

Figure 20. Closed-ended follow-up probe design



Item-nonresponse will remain a major concern with regard to narrative open-ended questions. Additionally, probing a narrative open-ended question with another open-ended question may have a negative influence on respondent cooperation. Even variations in visual questionnaire design could not improve response rates to open-ended

questions (see prior experiments). Switching to a closed-ended question might be a solution for nonresponse issues, but will compromise the accuracy of survey results and evoke other difficulties associated with closed-ended questions (see chapter 5). Therefore, replacing open-ended questions by closed-ended ones is not ideal in most situations, but combining the two types of questions might be a solution. The closed-ended probing approach is an attempt to combine the advantages of both open- and closed-ended questions. To make use of the rich and detailed measurement of open-ended questions, respondents are asked a question in an open format. If they answer the question, respondents just continue to the next question in the survey; if not, they are asked the exact same question again in a closed-ended format (see Figure 20). Since closed-ended questions have a distinct lower response burden, we expect that many respondents who left the initial open-ended question unanswered will answer the closed-ended probe. This probing approach will therefore combine the accuracy of open-ended questions with the efficiency of closed-ended ones in order to reduce the overall measurement error.

10 Experiments testing adaptive design in narrative open-ended questions

10.1 Experiment IV: Adapting individual answer-box sizes

Experiment four is aimed at the utility of adapting answer-boxes, in order to offer a survey design that best fits the motivation and behavior of groups of respondents. By using an adaptive assignment of answer-box sizes for groups of respondents, we expected to reduce item-nonresponse for those respondents who usually would not answer a narrative open-ended question, and also encourage prolific writers to write even more detailed and rich responses. Overall, the adaptive box design with customized answer-box sizes is assumed to reduce item-nonresponse and increase the quantity and quality of responses to narrative open-ended questions.

10.1.1 Experimental design

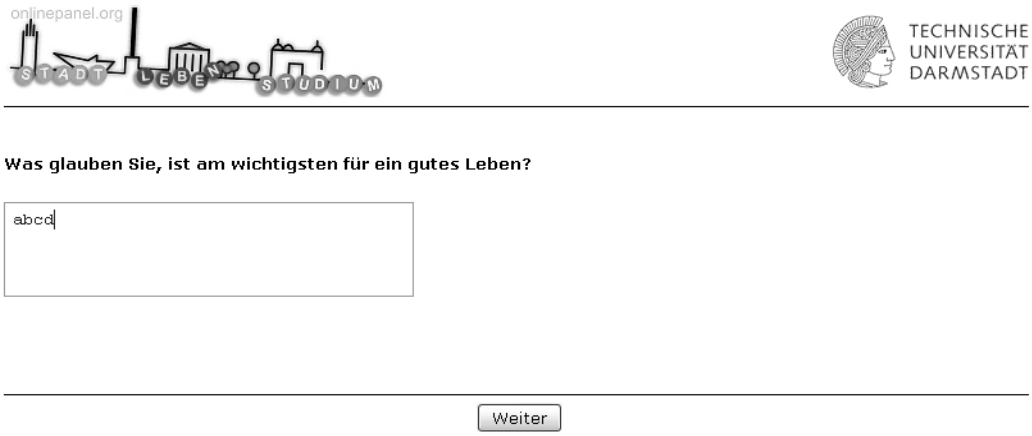
The fourth experiment was carried out in the Darmstadt University of Technology online access panel in November 2011 (the same study as experiment I). For this study, 2446 panel members were invited to a Web survey on student life satisfaction. Of those, 907 panelists took the survey, resulting in a response rate of 41 percent (AAPOR RR2). In the Web questionnaire, respondents were presented with an open-ended narrative question: *"In your opinion, what is most important to lead a good life?"* Later in the questionnaire, respondents were again asked an open-ended narrative question: *"In your opinion, what is essential to succeed in academic studies?"* After the first question, 25 percent of the respondents were randomly assigned to a control group in a between-subjects design (see Figure 21).

Respondents in the control group got the same answer-box size for the second questions (b) as they had for the first (4 rows with 42 characters each). The remaining 75 percent of respondents composed the experimental group for the adaptive design.

Figure 21. Adaptive answer-box design

First question

(a) Initial box (mid-size)



Second question

(b) Control group box (mid-size)

(c) Small box: no characters were typed into the initial box

(d) Mid-size box: up to 84 characters were typed into the initial field

(e) Large box: 85 or more characters were typed into the initial field

Notes. Small box = 2 rows with 42 characters each; Mid-size box = 4 rows with 42 characters each; Large box = 8 rows with 42 characters each

In the experimental group, respondents who did not respond to the first open-ended question were assigned to a small answer-box (c) in the second open-ended question (2 rows with 42 characters each).

Respondents who filled half or less of the first answer-box (up to 84 characters) were assigned to the same box size (d) for the second open-ended question (4 rows with 42 characters each), while respondents who filled more than half of the first question's answer-box (85 characters or more) were assigned to an answer-box twice the size (e) of the first question (8 rows with 42 characters each). Response quality was measured as in Experiment I, considering item-nonresponse, the number of characters, the number of topics, and the number of characters per topic.

10.1.2 Results

Overall, we found no differences in item-nonresponse in this fifth experiment (Table 12) when comparing responses to the initial box size and responses in the control group, nor when comparing the control group to the experimental group. As in Experiment I, the box size did not affect item-nonresponse in the adaptive design setting.

Table 12. Item-nonresponse, characters, topics, and characters per topic by first initial and second/adaptive question

	(a) First/initial question	Second/adaptive question	
		(b) Control group	(c) Experimental group
No response	13%	13%	15%
Response	87%	87%	85%
Characters	48	46[c]	53[b]
Topics	2.4[b,c]	2.1[a]	2.0[a]
Characters per topic	21[c]	22[c]	29[a,b]
N	800	209	591

Notes. Comparing initial question, control group, and experimental group: one-way ANOVA, F-test; a,b,c = p < .05. Letters denote significant differences to the indicated columns. Outliers were excluded at two standard deviations above the group mean for characters, topics and characters per topic

The overall length of the answer did not vary between the initial (48 characters) and the control (46 characters) groups, indicating that both questions are comparable in terms of the length of the answers they

elicit. Answers to the adaptive box design in the experimental group were significantly longer (53 characters) in comparison to the control group ($p < .05$). Thus, in total we accomplished on average more characters using the adaptive box design. When comparing the number of topics for both questions, a significant decrease ($p < .05$) between the initial question (2.4 topics) and the control group (2.1 topics) became evident. Also, there was a significant difference to the experimental group (2.0 topics, $p < .05$). The differences between the answers to the initial question and the control group in reported topics might have been caused by the question itself and not necessarily by the differences in the answer-box design. Therefore, the characters per topic are a better measure for elaboration and data quality, as it relates to the answer differences affected by visual design and not by question wording. Even though the adaptive design has no effect on item-nonresponse, the overall number of characters and the proportion of characters per topic were significantly higher ($p < .05$) for the adaptive answer-box design in the experimental group, as well as compared to the initial question.

Table 13 summarizes results of a simulation. Results for the first/initial question are reported separately for the groups of respondents who were assigned to the three answer-box designs for the second/adaptive question. Thus, we are able to compare response behavior for the first/initial and the second/adaptive question for non-respondents to the first/initial question (small box), for respondents who provided a condensed answer (mid-size box), and for high-volume writers (large box). This enables an assessment of the impact of the assigned box size on the response behavior of respondents in the respective groups (control, small box, mid-size box, and large box). In addition, responses by members of the control group are displayed. In the control group, a slight decrease in the number of characters occurred when comparing the responses to the first open-ended question (50 characters) and second open-ended question (46 characters, not statistically significant). While the decrease in the number of topics (2.6 vs. 2.1 for the second question) is significant ($p < .001$), the number of characters per topic did not differ significantly (20 vs. 22 characters per topic).

Table 13. Number of characters, topics, and characters per topic by answer-box type in the second adaptive question

	Control group (mid-size box)	Small box	Mid-size box	Large box
1st initial question				
Characters initial question	50	-	33	120
Topics initial question	2.6	-	2.1	3.6
Characters per topic initial question	20	-	20	40
2nd adaptive question				
Characters adaptive answer-box	46	28	47 (I)	114
Topics adaptive answer-box	2.1 II	1.8	1.9 II	2.7 II
Characters per topic adaptive answer-box	22	18	24 III	50 (III)
N	165	17	374	74

Notes. Pairwise Chi²-tests comparing the adaptive answer-box with the answers given to the first initial question: I, II, III= p < .001; (I), (II), (III) = p < .05. Roman numbers denote significant differences to the indicated rows in the upper part of the table. Outliers were excluded at two standard deviations above the group mean for characters, topics and characters per topic and the number of cases (N) including all respondents who answered the second open-ended question

The small answer-box was applied to all non-respondents to the first question. Compared to respondents who answered the initial question (mid-size box and large box), respondents in this group typed in fewer characters (28) and topics (1.8); characters per topic proportion (18) is also lowest for this group. However, we were able to motivate 17 former non-respondents to answer the second open-ended question, while item-nonresponse to the second question rarely occurred.

The mid-size box was assigned to respondents who filled up to one half of the answer-box in the first/initial question. In comparison to their answers to the first question, respondents provided significantly more characters to the second adaptive questions. The number of characters increased significantly ($p < .05$) from 33 to 47, and the number of characters per topic rose from 20 to 24 ($p < .001$). Only the number of

topics decreased, from 2.1 to 1.9 ($p < .001$) for the mid-size answer-box group.

The large answer-box was presented with the second question if respondents provided a long response to the first open-ended question (85 or more characters). In this group, the number of topics reported also decreased compared to the first/initial question (from 3.6 to 2.7, $p < .001$), likely due to the subject of the second open-ended question. The number of characters provided in response to the large box (114 characters) in the adaptive box design setting did not change significantly in comparison to the initial question (120 characters). Therefore, prolific writers still provided extensive information in the adaptive answer-box design. However, the number of characters per topic for this group (50 characters per topic) was significantly higher ($p < .05$) than in the initial question (40 characters per topic), indicating that the large box increased elaboration for this group (plus 25 percent more characters per topic). This increase was more pronounced in comparison to the control group (plus 10 percent) and the mid-size box group (plus 20 percent).

10.1.3 Summary experiment IV

In this fourth experiment we tested an adaptive answer-box assignment. Even though we are able to gain a few respondents by using the adaptive design, there is no substantial influence on item-nonresponse when using adaptive box sizes. As the analysis of personal characteristics on the willingness to respond revealed, non-respondents differed from those who provided a response, suggesting a possible bias in the survey data. However, the use of adaptive answer-boxes produced more information with respect to the number of characters and the characters per topic, in comparison to the control group with a mid-size answer-box. Compared to the first initial question, the adaptive box design yielded more characters per topic. This result is even more remarkable as respondents tend to write shorter responses to narrative open-ended questions, when they appear later in the questionnaire (Galesic & Bosnjak, 2009; Oudejans & Christian, 2010).

Overall the adaptive answer-box design was meant to get the most out of the respondent, taking into account his or her motivation and capabilities. The prior three experiments using visual design to improve narrative open-ended questions failed to prove a higher response burden for larger answer-boxes. The highly-educated student sample might be a reason that this higher burden is not indicated by higher item-nonresponse in the prior experiments. Therefore, testing the adaptive design in a larger sample general-population survey would help to assess the usefulness of adaptive answer-box sizes. Even though the adaptive design did not improve item-nonresponse in our experiments, it increased the response length and elaboration of provided answers, making it an interesting approach to take advantage of the opportunities that Web surveys offer.

10.2 Experiment V: Adaptive checkbox

Adapting answer-boxes individually to groups of respondents can be one way to improve responses to narrative open-ended questions. But prior responses in Web surveys can also easily be used for probing. The adaptive checkbox design assigns respondents who did not answer a narrative open-ended question in the first place to the same question in a closed-ended format.

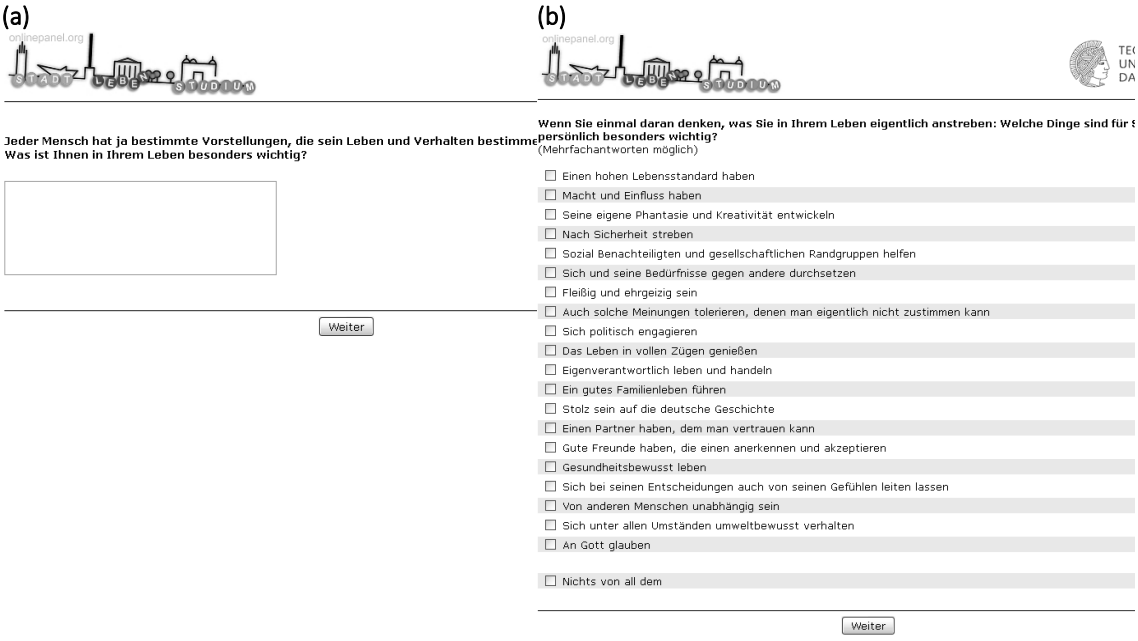
Item-nonresponse is one of the largest problems of narrative open-ended questions, based on their relatively high response burden (e.g. Galesic, 2006; Reja, et al., 2003; Scholz & Zuell, 2012; Smyth, et al., 2012). The adaptive checkbox design attempts to compensate for item-nonresponse by providing a follow-up probe. Respondents with lesser cognitive abilities or motivation are assigned to a less burdensome closed-ended question and will likely provide at least some information. This design is, of course, a trade-off, and one might easily suggest asking a closed-ended question right from the start. However, for questions that would best be asked open-ended, probing can be useful. Questions that can be asked best in a closed-ended design, however, should be asked using a closed-response format.

10.2.1 Experimental design

The experiment was carried out in July 2012 in the online access panel at Darmstadt University of Technology. More than 2,228 panelists were invited to a survey about marriage and divorce. At least 891 respondents answered 50 percent of the questions, resulting in a response rate of 39.99 percent (RR2).

Later in the questionnaire, respondents were assigned to the narrative open-ended question, “Everybody has some major things to accomplish in his life: What are yours?” The answer-box measured 12 rows with room for 50 characters in each (see Figure 22).

Figure 22. Adaptive checkbox design: (a) initial open-ended question and (b) closed follow-up probe



Respondents who typed anything into that answer-box and clicked the submit button continued the survey with the next survey question. Respondents who left the open-ended question unanswered (typed nothing into the open-ended question) and continued the survey were assigned to the same question about the “main things to accomplish” again, this time using a closed-ended “check all that apply” answer format. The question, as well as the answer categories, were taken from

the Shell youth survey (e.g. Shell, 2000) and displayed in a multiple checkbox format.

10.2.2 Results

The narrative open-ended question was answered by 72 percent of the respondents who then continued the survey (see Table 14). However, 254 respondents did not answer the open-ended question and were therefore assigned to the same question in a closed-ended format. Of those respondents, 213 (84 percent) answered the same question in the closed-ended format, with only 16 percent leaving this rather demanding question unanswered.

Table 14. Response and nonresponse rate for initial open-ended question and the following closed probe

Answer provided	(1) Open-ended	(2) Closed-ended	Combined
Response	71.5 %	83.9 %	95.4%
	637	213	678
No response	28.5 %	16.1 %	4.6%
	254	41	41
N	891	254	891

Overall, the item-response rate accomplished by the open-ended format (at 72 percent) is expanded by a further 23 percent using the closed-ended format, resulting in a combined response rate of 95 percent. The closed-ended format also included a “don’t know” option. This would have allowed respondents to note that they did not answer the question in the first place because they did not hold an attitude towards the topic. Interestingly, even though there was a “don’t know” option, not a single respondent used it in the closed-ended format.

Using a closed-ended probe may improve the overall response rate. In regard to their content, how responses of the initial open-ended question compare with the closed-ended probe is displayed in Table 15.

Open-ended responses were sorted according to topic by one coder, which is why inter- and intra-coder reliability cannot be highlighted here.

The open-ended question allowed for even more detailed coding, but the categories used were implied by the response options used in the closed-ended format. Responses were only coded if at least two respondents provided a response category (single entries were not coded).

Table 15. Frequency comparison open- vs. closed-ended question

	Open-ended (N=637)			Closed-ended (N=213)		
Items	Rank	N	Percent	Rank	N	Percent
friends	**2**	154	24.2	**1**	197	89.1
relationship	**4**	108	17.0	**2**	185	83.7
good family life	**1**	240	37.7	**3**	163	73.8
self-dependency	24	2	0.3	**4**	149	67.4
high standard of living	8	68	10.7	**5**	126	57.0
enjoy life to the full	12	46	7.2	6	123	55.7
seek safety	16	30	4.7	7	109	49.3
hardworking and ambitious	20	11	1.7	8	97	43.9
healthy living	7	69	10.8	8	97	43.9
independent	6	73	11.5	10	94	42.5
tolerant towards other assumptions	14	32	5.0	11	91	41.2
imagination and creativity	9	65	10.2	12	87	39.4
be intuitive	10	57	8.9	13	72	32.6
help others	17	24	3.8	14	45	20.4
environmental consciousness	20	11	1.7	15	44	19.9
political involvement	24	2	0.3	16	33	14.9
influence and power	24	2	0.3	17	32	14.5
believe in God	14	32	5.0	18	28	12.7
prevail own interests	22	8	1.3	19	16	7.2
pride in being German	27	0	0.0	19	16	7.2
don´t know	23	5	0.8			
Additional categories						
having a job	**3**	127	19.9			
being happy	**5**	105	16.5			
being calm and balanced	11	54	8.5			
succeed in studies/education	13	43	6.8			
traveling	19	18	2.8			
single entries not coded	18	20	3.1			
Total		1406			1804	

The different distributions and rankings of the open- and closed-ended question Table 15 reveal some major differences, even though they could not be directly compared (since respondents were not randomly assigned). While "self-dependency" and "hardworking and ambitious" ranked high in the closed-ended probe, they were only rarely highlighted in the initial narrative open-ended question. On the other hand, the closed-ended question showed a lack of response to the answer categories "having a job" and "being happy," which were among the most frequent responses in the closed-ended format. Overall, coding revealed at least five possible additional topics respective to answer categories that were missing from the closed-ended question. These additional topics were highlighted by more than 50 percent of the respondents who answered the open-ended question.

Another difference between the initial open-ended question and the closed-ended probe appears in the number of topics reported respective to answer categories checked. While the 637 respondents to the open-ended question reported 1,406 topics altogether, the 237 respondents in the closed-ended probe checked 1,804 response categories. Therefore, the number of topics reported amounted to 2.2 topics on average to the narrative open-ended question, while respondents to the closed-ended question reported 8.5 topics on average.

10.2.3 Summary experiment V

The closed probe tested in the fifth experiment is similar to the idea of an adaptive design, insofar as the following probe is based on prior responses and only displayed if the initial narrative open-ended question was not answered. This approach, using a closed-ended follow-up probe, was aimed at combining the strengths of closed- and open-ended questions: getting thick and rich responses from narrative open-ended questions that allow respondents to freely answer questions in great detail while at the same time increasing response rates by presenting a closed-ended probe. The follow-up probe assures exactly that.

The question used in the fifth experiment has been part of the Shell youth survey for several years (e.g. Shell, 2000), and the categories

provided were expected to have been sufficient. However, we considered that the question would be better asked in an open-ended format, since its topic is rather broad. The results confirm this view, and exposed five additional categories missing from the closed-ended question. These five categories accounted for over 50 percent of all responses, clearly indicating that the open-ended format is more suitable for this specific question.

This experiment also revealed more differences when using open- and closed-ended questions. The narrative open-ended question suffered from higher item-nonresponse, but measurement seems to be superior over the closed-ended probe. Respondents provided fewer topics but expanded on them and indicated their importance. In the follow-up probe, respondents just checked all answers that applied, even if they were of minor importance and would not have occurred to them when asked in an open format instead. Closed-ended answer categories might "give away" the answers too easily and bias responses. That becomes especially apparent when looking at the number of topics reported for the open-ended question (2.2 topics on average), in comparison to the number of answer categories checked in the closed-ended probe (8.5 topics on average).

The closed-ended probe was especially aimed at reducing item-nonresponse. While 29 percent of the respondents did not answer the initial open-ended question, only 5 percent of the respondents left the question unanswered when combining the results of the narrative open-ended question and the closed-ended follow-up probe. Personal characteristics affected item-nonresponse in open-ended questions (see Experiments II and III), suggesting a bias compromising data quality. Increasing response rates might therefore help to reduce the bias at work in narrative open-ended questions.

10.3 Discussion on adaptive design experiments

Experiments IV and V made use of the opportunities provided by Web surveys to adapt questionnaires individually to respondents based on

their prior responses. While the visual design experiments failed to reduce item-nonresponse in narrative open-ended question we expected the adaptive design to increase response rates.

Adapting individually-sized answer-boxes increased only the length of responses to narrative open-ended questions. The answer-box size does not seem to pose a higher burden to respond, at least not for the samples used to carry out this experiment. As in the prior visual design experiments, responses can be improved using the adaptive answer-box size assignment, but the willingness to respond is not affected by any of the designs tested. This problem becomes especially important when those participants who refused to respond differed from those who provided a response. If they do differ, survey results may be biased. Compared to a standard open-ended answer-box design, the adaptive assignment of answer-box sizes can increase the elaboration of responses and does not compromise data quality.

In order to improve response rates, the final experiment in this thesis used a closed-ended follow-up probe to combine the strengths of closed- and open-ended questions. Switching to a closed-ended question is not ideal, but the design accomplished the aim of getting at least some information from former non-respondents. In the initial open-ended question, respondents provided fewer topics but elaborated them. In the closed-ended follow-up probe the respondents checked more answer categories respective to topics, most likely because they were at hand. Overall, the probe succeeded in getting information from those respondents who neglected to answer the same question in an open format. Therefore, the probe is a valuable addition to any open-ended question, as long as reasonable closed-ended answer categories can be provided. In general population surveys, item-nonresponse to narrative open-ended questions might be more pronounced than in the highly-educated student sample we used, making closed-probing even more sensible.

11 SUMMARY AND CONCLUSIONS

This thesis was aimed at using visual and adaptive design to improve the data quality of narrative open-ended questions in Web surveys. Open-ended questions are powerful tools for collecting specific data on a topic, from large samples of respondents (Oudejans & Christian, 2010). Without constraining responses, open-ended questions allow respondents to freely answer and elaborate on their responses. They are especially useful when no suitable answer categories are at hand for a closed-ended question format, or if suitable response options might bias respondents. However, open-ended questions compromise data quality to some extent, especially because they are burdensome to answer. Question difficulty, and the respondent's motivation and abilities all contribute to the quality of responses and the likelihood of respondents providing satisficing, rather than optimizing, responses.

The best way to reduce task difficulty and improve responses to narrative open-ended question is by providing a good, and easily understandable, question. Further instructions or examples can compensate for the lack of closed-ended guiding response-options, and help respondents understand the question. Further motivating instructions have proven to elicit long and rich responses (Israel, 2013; Smyth, et al., 2009). But highlighting the importance of a narrative open-ended question might work a few times only, because not every question can be "very important".

This thesis focused on the use of visual design to reduce the response burden and to motivate respondents. Various studies have focused on the design of associated answer-boxes with respect to open-ended questions and found that the answer-box size can especially influence responses (Christian & Dillman, 2004; Israel, 2010; Smyth, et al., 2009). Despite the influence of varying answer-box sizes on responses, this thesis also tested the use of a counter indicating the number of characters left, dynamic answer-boxes that automatically grow as the respondent types, and respondent-adjusted answer-boxes to improve responses. In addition, the adaptive designs tested made use of the

interactivity of the Web and used prior responses to adapt a questionnaire to different groups of respondents. Both visual and adaptive design were expected to improve data quality in narrative open-ended questions. Any critique of the various designs tested must be done in the context of their intended use and the claims these designs make. Even though written language is more powerful than visual design language, paying attention to the visual design of Web surveys can improve data quality. Visual design must be considered as an additional way of improving questions, along with the actual question wording/formulation itself.

11.1 Main findings and implications

The visual design experiments tried to improve responses by reducing the burden to respond and by motivating respondents. Increasing the answer-box size resulted in longer and more detailed responses rather than in more reported topics. While larger answer-boxes were expected to pose an additional burden, we found no influence of the answer-box size on item-nonresponse. Using a counter indicating the number of characters left curtailed the response length if the default counter value was set low, and increased the response length when the default value was set high. However, a low-value counter limited the number of words used by respondents, but not the number of topics reported. Respondents always seemed to report what they intended to report. Therefore, when designing narrative open-ended questions for a Web survey, the answer-box size should be large rather than small. Of course, answer space size should be appropriate and reflect the question and the response length expected. The use of a counter indicating the number of characters left can increase the length of responses. It is a useful addition to any narrative open-ended question, as long as the default value matches or exceeds the size of the answer-box that it is displayed with.

Automatically growing answer-box designs do not improve response length or the number of topics reported to narrative open-ended question. The dynamic growing design does not correspond to the

question-answer process, since the answer-box size is increased after the respondent has formatted the response; the additional rows are displayed too late in the process of forming a response. In a Web survey with numerous open-ended questions, respondents might get used to the dynamic design and it could help to improve overall data quality. But based on the experiments carried out in this thesis, the dynamically growing answer-box design should be avoided in Web surveys.

In the respondent-adjusted design, respondents were able to set the answer-box size themselves. Since they were aware of the box-size adjustment, the design corresponds better with the question–answer process, as compared to the dynamic growing answer-spaces. As a result, respondents reported more topics and produced longer responses with the self-adjusted answer-box design. At first glance, the respondent-adjusted design appears to be a good design option, but the required instructions for adjustment add to the overall response burden. Therefore, it is harder to recommend the respondent-adjusted answer-box design.

Except for the dynamic answer-box design, all visual designs tested for this thesis increased the overall response length and the elaboration of responses. However, not all four designs get more respondents to answer narrative open-ended questions. The different visual designs do not harm data quality by increasing item-nonresponse, but neither do they get more respondents to answer an open-ended question.

The adaptive designs were attempts to do exactly that: reduce the burden and get more respondents to answer narrative open-ended questions. But adapting individually-sized answer-boxes again only increased the length of responses, while the willingness to respond was not affected by the adaptive box design. Like the varying visual designs tested, the adaptive design was able to evoke longer responses, only.

The closed-ended probe design was aimed at getting at least some information from those respondents who had not answered the question in an open-response format. Almost all nonrespondents to the open-ended question answered the closed-ended probe. The probe therefore

worked as expected. But it is important to point out that closed-ended probing is not suitable for every narrative open-ended question. First of all, it requires a set of answer categories. If a valid set of response categories is available, asking a question in a closed-ended format might be a better choice. However, if the response options are likely to bias responses, a closed-ended probe is not a sensible design option, either. Yet the probe has to be considered as an addition to an open-ended question. It also helps to prove the point that many nonrespondents hold an attitude towards the open-ended question posed. In comparing the responses of the initial narrative open-ended question with the closed-ended probe, it becomes apparent that respondents and nonrespondents differ in their answer distribution. This difference can be a result of the answer-categories, or of respondents' personality differences.

The analysis of personal characteristics in this thesis emphasizes one of the largest problems faced by users of narrative open-ended questions. They are burdensome to answer and the data obtained might be biased, based on the influence of personal characteristics on item-nonresponse. Female respondents and respondents with basic algebra skills answered narrative open-ended questions less frequently than males and those with advanced algebra skills. While personal characteristics had an influence on the willingness to respond, their influence on the response length was negligible. Since personal characteristics and abilities remain more or less constant through time, survey designers can only try to provide questions that are easy to read, understand, and answer. Reducing the response burden or motivating respondents using visual design will further help to improve the response quality of narrative open-ended questions in Web surveys.

11.2 Limitations and future suggestions

The fact that personal characteristics affect responses has to been understood in the context of the samples used. All experiments reported were conducted using a highly educated student population sample that

might even diminish the influence of personal characteristics. The highly educated student population may be more able and willing to answer narrative open-ended questions. The influence of visual design, using homogeneous student population samples, might suggest an even stronger influence in a more widespread sample. Whether the results of the experiments will hold for Web surveys in the general population remains to be tested.

Another important aspect when evaluating the results and implications from this thesis is the fact that every open-ended question asked is unique. Whether the visual designs tested will improve every narrative open-ended question in the very same way is arguable. While the influence of the answer-box size, the automatically growing answer-box, and the counter were tested in more than one experiment, the other experiments should be replicated in further studies.

For a long time, Web surveys mimicked paper questionnaires with respect to layout and design, even though the opportunities and interactivity the Web allows for are almost endless. But the use of such methods only makes sense “when they add value to a survey, not simply for the sake of doing so” (Couper, 2008 p. 114). The interactive nature of the Internet allows a multiplicity of ways to integrate interactive features into a survey that can engage respondents and help them stay motivated throughout the survey (Oudejans & Christian, 2010). Motivating or probing respondents are techniques that previously had been available only in interviewer-administered surveys. Adapting the questionnaire based on prior responses was used in this thesis to improve results of narrative open-ended questions. However, the adaptive design can be useful for other question types as well, such as in burdensome matrix questions that are especially prone to satisficing. Respondents who do not differentiate between the items displayed could be assigned to a design that presents the matrix as single items on the screen design, forcing respondents to spend more time with each item.

Another example would be “all-that-apply” questions (see Figure 6) where respondents are instructed to check all items that apply, from a

list of response options. “Check all that apply” questions are prone to satisficing and many respondents check only the first reasonable item (Krosnick, 1999). In an adaptive design, respondents who leave all checkboxes blank or only use the upper items displayed could be assigned to a forced-choice-question format (e.g. yes or no) for each item in a list.

These are only a few examples of taking advantage of opportunities that Web surveys provide. One very promising approach for improving narrative open-ended questions in particular in Web surveys is the speech-to-text integration in HTML5. The Chrome browser already allows speech input by default. Allowing respondents to dictate their responses might help to increase response rates and reduce the bias induced by personal characteristics and abilities.

11.3 Concluding remarks

The answer-box size and a counter indicating the number of characters remaining affect responses, as does the respondent-adjusted design. Using larger answer-boxes, adding a counter, or allowing respondents to adjust the answer-box size all improved responses. In addition, the use of adaptive box-sizes increased the number of characters to open-ended narrative questions. Results provide preliminary support for the effectiveness of a Web survey design that adapts the type and visual design of survey questions to the motivation and capabilities of the respondent. While previous studies in the design of open-ended narrative questions aimed to enhance the effectiveness of design features that were meant to influence response behavior (in particular of less-motivated respondents), the adaptive design changes the questionnaire in order to get the most out of the respondent, consistent with his motivation and capabilities. The closed-ended probe, particularly, increased response rates. Overall, the experiments demonstrate that it is well worth paying attention to the visual and adaptive design of open-ended questions in Web surveys, and that well-

designed open-ended questions are a powerful tool for collecting specific data from large samples of respondents.

12 Deutsche Zusammenfassung (German summary)

Um Informationen, Einstellungen oder Meinungen über bestimmte Gruppen von Personen zu erhalten, ist die standardisierte Befragung als Erhebungsmethode in den Sozialwissenschaften unverzichtbar. Dabei gilt es, für die befragten Personen in der Regel eine zuvor festgelegte Reihe an Fragen zu interpretieren und zu beantworten, gleich, ob diese Fragen persönlich, am Telefon, mittels Papierfragebogen oder in einer Onlinebefragung gestellt werden (Groves et al., 2009). Neben dem Inhalt oder der Funktion einer Frage innerhalb eines Fragebogens ist eine der zentralsten Differenzierungen die des Fragetyps zu treffen: Wird der Typ der geschlossenen oder offenen Frage verwendet. Geschlossene Fragen erfordern von Befragten aus einer festgelegten Reihe an Antwortmöglichkeiten ihre Antwort auszuwählen, während offene Fragen von den Befragten in ihren eigenen Worten frei beantwortet werden. Geschlossene Fragen lassen sich dabei einfacher beantworten und auch einfacher auswerten. Im Gegensatz dazu müssen die Antworten auf offene Fragen von den Befragten selber formuliert und im späteren Analyseprozess aufwändig kodiert werden. Geschlossene Fragen schließen daher Kodierfehler von vornherein aus und sparen Zeit und Geld in der späteren Analyse (Reja, Manfreda, Hlebec & Vehovar, 2003). Aber auch für die Befragten haben geschlossene Fragen Vorteile: So erhöhen die vorgegebenen Antwortkategorien das Frageverständnis und erleichtern das Finden einer adäquaten Antwort. Darüber hinaus müssen Befragte lediglich die passende Antwort markieren und nicht aufwändig einen eigenen Antworttext verfassen. In standardisierten Befragungen finden daher deutlich häufiger geschlossen Fragen Verwendung (Krosnick, 1999).

Allerdings sind geschlossene Fragen nicht für jeden Einsatzzweck und in jeder Situation ideal. Zum Beispiel, wenn keine umfassenden Antwortmöglichkeiten zu einer geschlossenen Frage formuliert werden können oder die genutzten Antwortmöglichkeiten die Antwort des Befragten beeinflussen. Außerdem entsprechen offene Fragen deutlich stärker der alltäglichen und natürlichen Kommunikation und machen sie

gerade deshalb für explorative Fragen interessant, da deren Fokus auf möglichst detaillierten und umfassend beschreibenden Antworten liegt. Dabei können offene Fragen genutzt werden um einfache Häufigkeiten abzufragen (z.B. “An wie vielen Tagen haben Sie innerhalb der letzten vier Wochen Zeitung gelesen?) und für kurze Antworten oder Auflistungen (z.B. „Bitte nennen Sie die Zeitungen, die sie in den letzten vier Wochen gelesen haben?). Aber auch für narrative Fragen (z.B. „Wie beurteilen Sie die Arbeit der gegenwärtigen Bundesregierung?), die von den Befragten in ihren eigenen Worten möglichst umfassend und detailliertet beantwortet werden sollen. Befragte müssen daher zum einen die nötigen Fähigkeiten haben und zum anderen die notwendige Motivation, eine Antwort zu verfassen. In der Praxis haben zahlreiche Studien gezeigt, dass Persönlichkeitseigenschaften das Antwortverhalten beeinflussen (z.B. Knäuper et al., 2004) und das im Vergleich zu anderen Fragetypen speziell narrative offene Fragen häufig nicht beantwortet werden (z.B. Galesic, 2006; Reja et al., 2003; Scholz & Zuell, 2012).

Diese Dissertation befasst sich mit (eben solchen) narrativen offenen Fragen und versucht, diese durch den gezielten Einsatz von visuellem und adaptivem Design für Online-Befragungen zu optimieren. Online-Befragungen haben sich in den letzten Jahren in relativ kurzer Zeit als Erhebungsmodus in der Umfrageforschung etabliert (ADM, 2012). Sie sind im Vergleich zu Papier basierten Befragungen schneller und günstiger zu realisieren und erlauben darüber hinaus nahezu endlose viele Möglichkeiten und Variationen in der Fragebogengestaltung (z.B. Couper, 2008a). Dabei ist es von großer Bedeutung, diese Möglichkeiten der interaktiven Gestaltung stets zielführend und nicht wahllos einzusetzen. Für die Designexperimente zu narrativen offenen Fragen in dieser Arbeit bedeutet dies, dass die unterschiedlichen Designs zum einen mehr Befragte zum Antworten bewegen sollen und zum anderen, dass die gegebenen Antworten detaillierter und umfassender sein sollen.

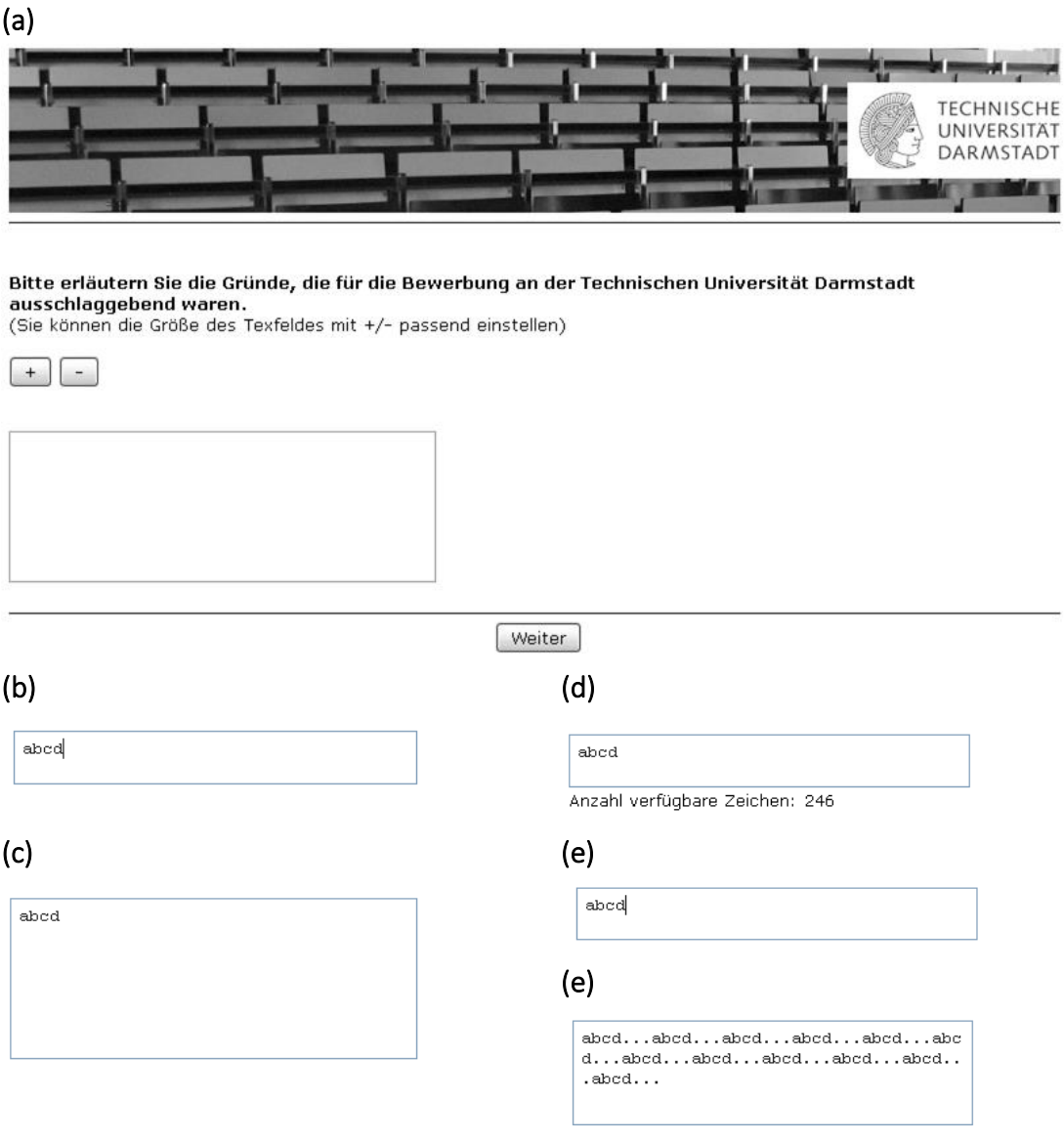
Online-Befragungen sind selbst administriert. Sie werden also von den Befragten in Eigenregie beantwortet, weshalb in diesem Fall die Schriftsprache die elementare Sprache ist, um mit Befragten zu

kommunizieren. Aber Antworten werden nicht alleine von der Frage bzw. dem Fragetext beeinflusst. Die visuelle Darstellung, Anordnung oder Reihenfolge von Antwortkategorien, der Einsatz von Bildern oder Farben hat zum Beispiel ebenfalls einen Einfluss auf Antworten (z.B. Schwarz, Grayson, und Knäuper, 1998; Toepoel & Couper, 2011; Smyth, Dillman, Christian & McBride, 2009; Tourangeau, Couper & Conrad, 2007a). In offenen Fragen hat neben dem Fragetext selber insbesondere die Darstellung des Antwortfeldes einen Einfluss auf das Antwortverhalten. So beeinflusst zum Beispiel die Größe eines Antwortfeldes die Länge gegebener Antworten (z.B. Israel, 2010; Smyth et al., 2009; Stern et al., 2007). Während kleine Antwortfelder dem Befragten suggerieren, nur eine kurze Antwort geben zu müssen, erzeugt ein großes Antwortfeld beim Befragten die Wahrnehmung, dass eine längere elaboriertere Antwort von ihm erwartet wird. Aufgrund des wahrgenommenen Aufwandes sollten offene Fragen, die an ein kleines Textfeld gekoppelt sind, häufiger beantwortet werden. Im Gegenzug ist zu erwarten, dass die Antworten bei einem größeren Textfeld umfangreicher und detaillierter ausfallen. Vor diesem Hintergrund ist die verwendete Größe eines Textfeldes stets ein Kompromiss zwischen dem zu leistenden Aufwand eines Befragten und der zu erwartenden Antwortlänge und -qualität. Genau an diesem Punkt setzt diese Arbeit an und versucht, anhand des gezielten Einsatzes von visuellem und adaptiven Design mehr Befragte dazu zu bringen, zu antworten (Aufwand zu Antworten reduzieren) und gleichzeitig die Befragten dazu zu motivieren, längere und detailliertere Antworten zu geben. Hierzu wurden im Rahmen dieser Dissertation fünf Experimente unter Verwendung verschiedener Studierendenstichproben online durchgeführt.

Insgesamt vier verschiedene visuelle Designs wurden dabei in drei Experimenten getestet (Siehe Abbildung 1). Zunächst wurde die Größe des Antwortfeldes in mehreren Abstufungen variiert (b und c), mit der Erwartung, dass kleinere Antwortfelder kürzere, dafür aber mehr Antworten erzielen (im Vergleich zu größeren Antwortfeldern). Darüber hinaus wurde die Wirksamkeit eines Counters (d), der die verbleibende Anzahl der verfügbaren Zeichen für den Befragten anzeigt, getestet. Der

Counter ermöglicht dabei die Nutzung eines relativ kleinen Antwortfeldes, dass einen geringen Antwortaufwand für den Befragten suggeriert, während der Counter Befragte kontinuierlich motiviert, alle zur Verfügung stehenden Zeichen für die Antwort zu nutzen. Zusätzlich wurden dynamisch wachsende Antwortfelder (e) getestet, die während der Befragte schreibt, kontinuierlich an Größe gewinnen.

Abbildung 1: Getestete visuelle Designs: (a) in der Größe verstellbares Antwortfeld, (b) kleines Textfeld, (c) großes Textfeld, (d) Textfeld mit Counter nach der Eingabe von 4 Zeichen, (e) dynamisch wachsendes Textfeld nach der Eingabe von 4 und 84 Zeichen



Wie das Counter Design so ermöglicht auch das dynamische wachsende Antwortfeld den Einsatz eines relativ kleinen Textfeldes zu Beginn, dessen kontinuierlicher Größenzuwachs Befragte dazu motivieren sollte, eine detailliertere Antwort zu verfassen. Während das dynamisch wachsende Antwortfeld auf der Idee basiert, Befragten eine relativ geringe Anfangshürde aufzuzeigen, die durch das sukzessive Anwachsen der Größe des Antwortfeldes erhöht wird, setzt das vierte getestete Design auf den genau gegenteiligen Weg. Anstatt das Textfeld automatisch mitwachsen zu lassen, wurde das Antwortfeld um einen Plus- und einen Minusknopf ergänzt (a), mit dem Befragte selber die Größe des Antwortfeldes individuell festlegen konnten.

Innerhalb der durchgeführten Experimente zeigte sich ein deutlicher Einfluss von visuellem Design auf das Antwortverhalten. So beeinflusste die Größe des Antwortfeldes die Länge gegebener Antworten und die Anzahl genannter Themen. Dabei stieg die Antwortlänge deutlich mit der Größe des Textfeldes an, während die Anzahl genannter Themen relativ gleich blieb. Der Einsatz eines Counters wirkte in einer ähnlichen Weise, wobei auch hier Befragte insgesamt längere Antworten verfassten und auch zusätzliche Themen in ihre Antwort integrierten - im Vergleich zu einem gleich großen Textfeld ohne Counter. Das dynamisch mitwachsende Antwortfeld konnte die Antwortqualität nicht verbessern, während das von den Befragten per plus und minus Knopf in der Größe verstellbare Antwortfeld längere Antworten und mehr Themen erbrachte. Mit Ausnahme des dynamischen Textfeldes konnten alle Designs die Antwortlänge vergrößern und die Anzahl der beschriebenen Themen anteilig erhöhen. Die Antwortbereitschaft der Befragten konnte allerdings keines der getesteten Designs erhöhen, so dass, je nach Experiment, 10 bis 30 Prozent der Befragten keine Antwort gaben.

Neben den Möglichkeiten in der visuellen Gestaltung von Fragen erlaubt die Interaktivität von Online Befragungen zahlreiche weitere Möglichkeiten, Befragte bei der Beantwortung eines Fragebogens zu unterstützen und zu motivieren. Bisher wurden in Online-Befragungen die Antworten von Befragten häufig nur zur Filterführung benutzt, um

beispielsweise die Frage nach einem Studium nur den Befragten zu stellen, die auch zuvor in einer Frage angaben, Student oder Studentin zu sein. Diese individuelle Anpassung eines Fragebogens auf Basis zuvor gegebener Antworten beschreibt die Funktionsweise des adaptiven Designs. Insgesamt wurden zwei unterschiedliche adaptive Designs getestet, die insbesondere die Responseraten offener Fragen erhöhen sollten. Das erste adaptive Design passte dabei die Textfeldgröße im Laufe der Befragung an, so dass Befragte, die eine offene Frage unbeantwortet ließen, in der Folge ein kleineres Antwortfeld zu einer offenen Frage angezeigt bekamen. Befragte, die einen kurzen Text in das erste Antwortfeld der Frage schrieben, bekamen zur nächsten offenen Frage ein gleich großes (mittelgroßes) Textfeld angeboten. Hingegen erhielten Befragte, die mehr als die Hälfte des ersten Textfeldes nutzten, ein großes Antwortfeld mit der nächsten offenen Frage. Ein kleines Textfeld sollte dabei mehr Befragte zu einer Antwort verleiten, die ein normal großes Textfeld sonst frei gelassen hätten. Befragten, die eher umfassende Antworten formulierten, wurde so mehr Platz eingeräumt, denn insgesamt sollten mehr Befragte zu einer Antwort bewegt werden um so insgesamt mehr Informationen zu erhalten. Das zweite adaptive Design zielte insbesondere auf die Problematik, dass Befragte offene Fragen häufig nicht beantworten (z.B. Galesic, 2006; Reja et al., 2003; Scholz & Zuell, 2012; Smyth et al., 2012). Auch wenn Onlinebefragungen Pflichtfragen ermöglichen, um diesem Problem zu begegnen, kommt diese Verpflichtung häufig einer Bevormundung gleich. Die Vorgehensweise der adaptiven Designs ist genau gegenteilig ausgelegt und versucht, den Fragebogen an den Befragten anzupassen und nicht umgekehrt. Vor diesem Hintergrund reicht gegebenenfalls die zuvor beschriebene Verkleinerung eines Antwortfeldes alleine nicht aus, um Befragte zur Beantwortung einer offenen Frage zu bewegen. Um zumindest zu einer Teilinformation zu gelangen, stellt das zweite adaptive Design deshalb die gleiche Frage in einem geschlossenen Antwortformat, wenn diese zuvor, offen gestellt, nicht beantwortet wurde. Damit wird die Anforderung an den Befragten deutlich gesenkt

und es ist zu erwarten, dass gerade weniger motivierte Befragte die geschlossen formulierte Frage beantworten.

Im Ergebnis zeigte sich für das erste adaptive Design, dass das Anpassen der Textfeldgröße basierend auf zuvor gegebenen Antworten zu insgesamt detaillierteren Antworten führt, aber nicht mehr Befragte dazu animiert, zu antworten. Dass die Responserate nicht gesteigert werden konnte, ist insbesondere deshalb problematisch, weil sich in den Analysen zeigte, dass sich Befragte, die keine Antwort gaben, signifikant von jenen unterschieden, die antworteten. Darüber hinaus wurden für die Experimente relative homogene Studierendenstichproben verwendet, weshalb sich in allgemeinen (heterogenen) Bevölkerungsstichproben noch stärkere Verzerrungen zeigen können. Item-Nonresponse und daraus resultierende Verzerrungen sind daher eines der großen Probleme in Bezug auf offene Fragen. Diesem Umstand trug speziell das zweite adaptive Design Rechnung, dass bei nicht erfolgter Beantwortung die gleiche Frage erneut in einem geschlossenen Antwortformat stellte. Auf die geschlossene Nachfrage antworteten hier rund 80 Prozent der Befragten, die zuvor keine Antwort gaben.

Die Ergebnisse der im Rahmen dieser Dissertation durchgeführten Experimente belegen die Wirksamkeit von visuellem und adaptivem Design. Es lohnt sich demnach bei der Gestaltung von Fragen und Fragebögen abseits des reinen Fragetextes und der Frageformulierung auch auf die visuelle Darstellung einer Frage zu achten. Und, auch wenn sich deutliche Effekte speziell für unterschiedliche Textfeldgrößen und den Einsatz von Countern in den Experimenten zeigten, bleibt eine gut geschriebene und leicht verständliche Frage immer noch der beste Garant für eine hohe Antwort- und Datenqualität. Die getesteten adaptiven Designs passten den Fragebogen, basierend auf den zuvor gegebenen Antworten an, ähnlich wie sich ein Interviewer in einer Interviewer-administrierten Befragung an einen Befragten individuell anpasst. Die Interaktivität von Online Befragungen bieten hier perspektivisch noch weitere Möglichkeiten auf Befragte einzugehen, sie

stärker zu unterstützen und offene aber auch geschlossene Fragen in Online Befragungen weiter zu optimieren.

13 References

AAPOR. (2011). *Standard definitions. Final dispositions of case codes and outcome rates for surveys. Revised 2011*: AAPOR.

ADM. (2012). *Jahresbericht 2012*: ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute.

Banaji, M. R., Blair, I. V., Schwarz, N., & Sudman, S. (1995). Implicit memory and survey measurement. *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. Jossey-Bass.

Bandilla, W., Bosnjak, M., & Altdorfer, P. (2003). Survey administration effects? A comparison of web-based and traditional written self-administered surveys using the ISSP environment module. *Social Science Computer Review, 21*(2), 235-243.

Beatty, P., & Herrmann, D. (2002). To answer or not to answer: decision processes realted to survey item nonresponse. In R. M. Groves, D. A. Dillmann, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresposne* (pp. 71-86). New York: Wiley.

Bethlehem, J. G., & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken, New Jersey: Wiley.

Bethlehem, J. G., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, New Jersey: Wiley.

Biemer, P. P. (2010a). Overview of design issues: Total survey error. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2 ed., pp. 27-57): Emerald Group Publishing Ltd.

Biemer, P. P. (2010b). Total survey error. Design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817-848.

Biemer, P. P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, New Jersey: Wiley.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annu. Rev. Psychol., 55*, 803-832.

Blom, A. G., Gathmann, C., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., et al. (2013). *The German internet panel: recruitment and beyond*. Paper presented at the Workshop of Longitudinal Research in Internet Panels. Mannheim, Germany.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods, 25* (4), 1-22.

Bosnjak, M., and Tuten, T. L. (2002), "Prepaid and Promised Incentives in Web Surveys – An Experiment. *Social Science Computer Review, 21* (2): 208–217.

Bradburn, N. M. (2004). Understanding the question-answer process. *Survey Methodology, 30* (1), 5-15.

Callegaro, M. (2005). *Origins and developments of the cognitive models of answering questions in survey research.* Paper presented at the first annual meeting of the European Association for Survey Research (EASR), Barcelona, Spain.

Chaiken, S., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J.S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 212–52). New York: Guilford

Chaudhuri, A. (2011). *Randomized response and indirect questioning techniques in surveys*. Boca Raton: Chapman & Hall CRC.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 73-96). New York, NY: Guilford.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68* (1), 57-80.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007). Helping the respondents get it right the first time: the influence of words, symbols, and graphics in Web surveys. *Public Opinion Quarterly, 71*(1), 113-125.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2010). The impact of progress indicators on task completion. *Interacting with computers, 22* (5), 417-427.

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1-28.

Cornilleau, A. (2013). *The French example*. Paper presented at the Workshop of Longitudinal Research in Internet Panels. Mannheim, Germany.

Couper, M. P. (2000). Web surveys. A review of issues and approaches. *Public Opinion Quarterly, 64*, 464-494.

Couper, M. P. (2001). *Web surveys: the questionnaire design challenge*. Proceedings of the ISI 2001. The 53rd Session of the ISI, Seoul, South Korea.

Couper, M. P. (2008a). *Designing effective Web surveys*. New York: Cambridge University Press.

Couper, M. P. (2008b). Technology and the survey interview/questionnaire. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 58-76). New York: Wiley.

Couper, M. P. (2011). *Web survey methodology: interface design, sampling and statistical inference*: Presented at the international statistical seminar of the Basque Statistical Institute (EUSTAT).

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an internet survey. *Social Science Research, 36*, 131-148.

Couper, M. P., Kennedy, C., Conrad, F. G., & Tourangeau, R. (2011). Designing input fields for non-narrative open-ended responses in web surveys. *Journal of Official Statistics, 27*(1), 65-85.

Couper, M. P., Tourangeau, R., & Conrad, F. G. (2007). Visual context effects in Web surveys. *Public Opinion Quarterly, 71*(4), 623-634.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the Effectiveness of Visual Analog Scales A Web Experiment. *Social Science Computer Review, 24*(2), 227-245.

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual design effects in Web surveys. *Public Opinion Quarterly, 68*(2), 255-266.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*(2), 230-253.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review, 19*(2), 146-162.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making, 21*, 241-251.

d'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med, 17*(19), 2265-2281.

de Leeuw, E. D., & De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M.

Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York: Wiley.

de Leeuw, E. D., & Hox, J. J. (2008). Self-administered questionnaires: mail surveys and other applications. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 239-263): Tailor and Francis.

de Leeuw, E. D., & Hox, J. J. (2010). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral reseach and the internet* (pp. 45-76). New York: Routledge.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). The cornerstones of survey research. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 1-16): Tailor and Francis.

de Vos, K. (2010). *Representativeness of the LISS-panel 2008, 2009, 2010*. Tilburg: CentERdata Institute for data collection and research.

Deming, W. E. (1944). On errors in surveys. *American Sociological Review, 9*(4), 359-369.

Denscombe, M. (2008). The length of responses to open-ended questions. A comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review OnlineFirst, 26*(3), 359-368.

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological review, 70*(1), 80.

Dillard, J. P. (1991). The current status of research on sequential-request compliance techniques. *Personality and Social Psychology Bulletin, 17*(3), 283-288.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method -- 2007 Update with new Internet, visual, and mixed-mode guide*. New York: Wiley.

Dillman, D. A., & Bowker, D. K. (2001). The web questionnaire challenge to survey methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 159-178). Lengerich: Pabst Science Publishers.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods, 17*(1), 30-52.

Dillman, D. A., & Messer, B. L. (2010). Mixed-mode surveys. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 551-574): Emerald.

Dillman, D. A., Phelps, G., Tortora, R. D., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode

surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research, 38*, 1-18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. New York: Wiley.

Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for constructing web surveys.* (Technical Report 98–50) Pullman, Wash.: Social and Economic Sciences Research Center, Washington State University.

Emde, M., & Fuchs, M. (2012). Exploring animated faces scales in web surveys: drawbacks and prospects. *Survey Practice,* (February 2012). Retrieved from http://surveypractice.wordpress.com/2012/02/21/exploring-animated-faces-scales/

Emde, M., & Fuchs, M. (2013). *Response Rate and Nonresponse in a Web Surveys When Using Text Message Invitations: Results of an Experiment with Traditional E-mail Invitations.* Paper presented at the 5th conference of the European Survey Research Association (ESRA). Ljubljana, Slovenia.

Eurostat. (2012). Internet use in households and by individuals 2012. Retrieved from http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-12-050/EN/KS-SF-12-050-EN.PDF

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

Frick, A., Bächtiger, M.-T., & Reips, U.-D. (2001). Financial incentives, personal information and dropout in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 209-219). Lengerich: Pabst.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392.

Fuchs, M. (2003). Kognitive Prozesse und Antwortverhalten in einer Internet-Befragung. *Österreichische Zeitschrift für Soziologie, 28*(4), 19-45.

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*(4), 701-725.

Fuchs, M. (2007). *Asking for Numbers and Quantities: Visual Design Effects in Web Surveys*. Paper presented at the 61th annual conference of the American Association for Public Opinion Research (AAPOR), Anaheim, CA.

Fuchs, M. (2008). Total survey error (TSE). In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (Vol. 2, pp. 896-902). Thousand Oaks: SAGE.

Fuchs, M. (2009a). Differences in the visual design language of paper-and-pencil surveys vs. Web surveys. A field experimental study on the length of response fields in open-ended frequency questions. *Social Science Computer Review, 27*(2), 213-227.

Fuchs, M. (2009b). *Dynamic screen design in open-ended questions. A field-experiment on the visual design language of Web surveys*. Paper presented at the 1st International Workshop of Internet Science, Bergamo, Italy.

Fuchs, M. (2009c). *The video-enhanced web survey. Data quality and cognitive processing of questions*. Paper presented at the Eurostat conference, Brussels, Belgium.

Fuchs, M. (2010). *Beyond question wording. The use of visual design and multi-media elements in web surveys*. Paper presented at the The 2nd International Workshop on Internet Survey Methods, Daejeon, South Korea.

Fuchs, M., Bossert, D., & Stukowski, S. (2013). Response rate and nonresponse bias - impact of the number of contact attempts on data quality in the European Social Survey. *Bulletin de Méthodologie Sociologique, 117*, 26-45.

Fuchs, M., & Busse, B. (2009). The coverage bias of mobile Web surveys across European countries. *International Journal of Internet Science, 4*(1), 21-33.

Funke, F. (2010). *Internet-based measurement with visual analogue scales: An experimental investigation.* Ebehard Karls Universität Tübingen, Tübingen.

Funke, F. (2011). *Explaining more variance with visual analogue scales: A web experiment*. Paper presented at the 4th conference of the European Survey Research Association (ESRA). Lausanne, Switzerland.

Funke, F., Reips, U.-D., & Thomas, R. K. (2011). Sliders for the smart: Type and rating scale on the web interacts with educational level. *Social Science Computer Review, 29*(2), 221-231.

Galesic, M. (2006). Dropouts on the Web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22*(2), 313-328.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a Web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly, 52*(3), 365-371.

Geer, J. G. (1991). Do open-ended questions measure "salient" issues? *Public Opinion Quarterly, 55*(3), 360-370.

Glasser, G. J., & Metzger, G. D. (1972). Random-digit dialing as a method of telephone sampling. *Journal of Market Research, IX* (February 1972), 59-64.

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology, 59*(6), 1216.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2 ed.). Hoboken, New Jersey: Wiley.

Groves, R. M., & Lyberg, L. (2010). Total survey error. Past, present, and future. *Public Opinion Quarterly, 74*(5), 849-879.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly, 72*(2), 167-189.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly, 68*(1), 2-31.

Haddock, G., & Zanna, M. P. (1998). On the use of open-ended measures to assess attitudinal components. *British Journal of Social Psychology, 37*, 129-149.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus Web surveying in a high-Internet-coverage population. Differences in response quality. *Public Opinion Quarterly, 72*(5), 836-846.

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys *Social Science Computer Review, 27*(2), 196-212.

Hox, J. J., De Leeuw, E., & Kreft, I. G. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 439-461): Wiley & Sons, Inc.

Israel, G. D. (2006). *Visual cues and response format effects in mail surveys*. Paper presented at the Annual Meeting of the Southern Rural Sociological Association, Tampa, FL.

Israel, G. D. (2010). Effects of answer space size on responses to open-ended questions in mail surveys. *Journal of Official Statistics, 26*(2), 271-285.

Israel, G. D. (2013). *Using motivating prompts to increase responses to open-ended questions in mixed-mode surveys: Where should the prompt be placed and to what effect?* Paper presented at the 67th annual conference of the American Association for Public Opinion Research (AAPOR), Boston, MA.

Jabine, T. B., Straf, M. L., Tanur, J. M., & Tourangeau, R. (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, D.C.: National Academy Press.

Kaczmirek, L. (2009). *Human-survey interaction. Usability and nonresponse in online surveys*. Köln: Herbert von Halem Verlag.

Kalton, G., & Stowell, R. (1979). A study of coder variability. *Applied Statistics*, 276-289.

Kaplowitz, M., Hadlock, T., & Vevinde, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly, 68*(1), 94-101.

Keusch, F. (2012). *Open-ended questions in web surveys: one large vs. ten small boxes*. Paper presented at the 66th annual conference of the American Association for Public Opinion Research (AAPOR), Orlando, FL.

Keusch, F. (2013). *The influence of answer box format, personal topic interest, and respondent characteristics on response behavior in open-ended questions*. Paper presented at the 67th annual conference of the American Association for Public Opinion Research (AAPOR), Boston, MA.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question diffiulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics, 13*(2), 181-189.

Knäuper, B., Schwarz, N., & Park, D. (2004). Frequency reports across age groups. *Journal of Official Statistics, 20*(1), 91-96.

Knäuper, B., Schwarz, N., Park, D., & Fritsch, A. (2007). The perils of interpreting age differences in attitude reports: question order effects decrease with age. *Journal of Official Statistics, 23*(4), 515-528.

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

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J. A. (2006). The handbook of questionnaire design: insights from social and cognitive psychology.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory on response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, M. W., Kopp, R. J., et al. (2002). The impact of "no opinion" response options an data quality. Non-attitude reduction or an invitation to satifice? *Public Opinion Quarterly, 66*(3), 371-403.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: initial evidence. *New Directions for Evaluation, 70*, 29-44.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. *Handbook of Survey Research. 2nd edition. Bingley, UK: Emerald*, 263-314.

Kruglanski, A. W. (1989). *Lay epistemics and human knowledge: Cognitive and motivational bases*: Plenum Press.

Kunz, T., & Fuchs, M. (2012). *Positioning of clarification features in web surveys: evidence from eye tracking data*. Paper presented at the 66th annual conference of the American Association for Public Opinion Research (AAPOR), Orlando, FL.

Kwak, N., & Radler, B. (2002). A comparison between mail and Web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18*(2), 257-273.

Lee, S., & Valliant, R. (2008). Weighting telephone samples using propensity scores. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 170-186). New York: Wiles.

Lenski, G. (1963). *The Religious Factor.* Rev. ed. Garden City, New York: Anchor Books.

Lensvelt-Mulders, G., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods & Research, 33*(3), 319-348.

Lensvelt-Mulders, G., Lugtig, P. J., & Hubregtse, M. (2009). Separating selection bias and non-coverage in Internet panels using propensity matching. *Survey practice*.

Lohr, S. L. (2010). *Sampling: Design and Analysis* (2 ed.). Pacific Grove: Duxbury Press.

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods, 4*(1), 43-59.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2007). Web surveys versus other survey modes: a meta-analysis comparing response rates. *International Journal of Market Research*.

Manfreda, K. L., & Vehovar, V. (2008). Internet surveys. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 264-284): Tailor and Francis.

Markus, H., & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (pp. 137-230). New York: Random House.

Menon, G. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 161-172). New York: Springer.

Messer, B. L., Edwards, M. L., & Dillman, D. A. (2012). Determinants of item nonresponse to web and mail respondents in three address-based mixed-mode surveys of the general public. *Survey Practice*, 1-7.

Mussweiler, T., Englich, B., & Strack, F. (2004). Anchoring effect. *Cognitive illusions—A handbook on fallacies and biases in thinking, judgment, and memory*, 183-200.

Neubarth, W. (2006). *Ranking vs. Rating in an Online Environment*. Paper presented at the 8th annual General Online Research (G.O.R.) conference of the The German Society for Online Research (D.G.O.F.), Bielefeld, Germany.

Niedomysl, T., & Malmberg, B. (2009). Do open-ended survey questions on migration motives create coder variability problems? *Population Space and Place, 15*, 79-87.

Oudejans, M., & Christian, L. M. (2010). Using interactive features to motivate and probe responses to open-ended questions. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and rehavioral research and the internet* (pp. 215-244). New York: Routledge.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion *Communication and Persuasion* (pp. 1-24): Springer.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. (2006). Web survey design. Paging versus scrolling. *Public Opinion Quarterly, 70*(4), 596-607.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., et al. (2004). *Methods for testing and evaluating survey questionnaires*. Hoboken, N.J.: Wiley.

Redline, C. D., & Dillman, D. A. (2002). The influence of alternative visual designs on respondents' performance with branching instructions in self-administered surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 179-193). New York: Wiley.

Redline, C. D., Dillman, D. A., Dajani, A. N., & Scaggs, M. A. (2003). Improving Navigational Performance in U.S. Census 2000 by Altering the Visually Administered Languages of Branching Instructions. *Journal of Official Statistics, 19*(4), 403-419.

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology, 49*(4), 243-256.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in Web questionnaires. *Metodoloski zvezki, 19*, 159-177.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research, 41*, 1415-1428.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. P. (2009). Selection bias in Web surveys and the use of propensity scores. *Sociological Methods Research, 37*(3), 291-318.

Schonlau, M., Zapter, K., Payne Simon, L., Sanstad, K., Marcus, S., Adams, J., et al. (2003). A comparison between responses from a propensity-weighted Web survey and an identical RDD survey. *Social Science Computer Review, 21*(1), 128-138.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitudes surveys* (reprint 1996 by Sage ed.). San Diego, California: Academic Press.

Schwarz, N. (1990). Assessing frequency reports of mundane behaviors. In C. Hendrick & M. S. Clark (Eds.), *Research Methods in Pesonality and social psychology* (pp. 98-119). Newbury Park: SAGE Publications.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology, 45*(3), 513.

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*(2): 177-183.

Schwarz, N., & Hippler, H.-J. (1987). What response scales may tell your respondents: Information functions of response alternatives. In N. Schwarz, H.-J. Hippler & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 163-178). New York: Springer.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*(3), 388-395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*(2), 127-160.

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55*(1), 3-23.

Shell, D. (2000). *Jugend 2000. 13. SHELL Jugendstudie, Bd. 1*.

Shih, T.-H., & Fan, X. (2007). Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. *International Journal of Internet Science, 2*(1), 59-82.

Simon, H. A. (1957). *Models of man: social and rational.* New York: Wiley.

Singer, E. (2012). *The use and effects of incentives in surveys*. Paper presented at the NSF Conference The Future Of Survey Research: A Pair of Conferences at the National Science Foundation.

Smith, T. (1995). *Little things matter: A sample of how differences in questionnaire format can affect survey responses.* Paper presented at the 50th annual conference of the American Association for Public Opinion Research (AAPOR), Fort Lauderdale, FL.

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2007). *Improving response quality in list-style open-ended questions in Web and telephone surveys*. Paper presented at the 61th annual conference of the American Association for Public Opinion Research (AAPOR), Anaheim, CA.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Mcbride, M. (2009). Open-ended questions in web surveys. Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly, 73*(2), 325-337.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly, 70*(1), 66-77.

Smyth, J. D., Powell, R., Olson, K., & Libman, A. (2012). *Understanding the relationship between literacy and data quality in self-administered surveys*. Paper presented at the 66th annual conference of the American Association for Public Opinion Research (AAPOR), Orlando, FL.

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods, 1*(3), 121-138.

Stiglbauer, B., Gnambs, T., & Gamsjäger, M. (2011). The interactive effects of motivations and trust in anonymity on adolescents' enduring participation in web-based social science research: A longitudinal behavioral analysis. *International Journal of Internet Science, 6*(1), 29-43.

Stoop, I., Billiet, J., & Vehovar, V. (2009). *Nonresponse bias in a cross-national study*. Paper presented to the International Statistical Institute, Durban, South Africa.

Strack, F. (1999). Beyond dual-process models: Toward a flexible regulation system. *Psychological Inquiry, 10*(2), 166-169.

Strack, F., Martin, L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*, 429-442.

Sudman, S., & Bradburn, N. (1982). *Asking questions*. San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N., & Schwarz, N. (2010). *Thinking about answers. The application of cognitive processes to survey methodology* (2 ed.). San Francisco: Jossey-Bass.

Thompson, K. J., & Washington, K. T. (2012). *A response propensity based evaluation of the treatment of unit nonresponse for selected business surveys*. Paper presented at the Federal Committee on Statistical Methodology Research Conference.

Toepoel, V., & Couper, M. P. (2011). Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly, 75*(1), 1-18.

Toepoel, V., Das, M., & Van Soest, A. (2009). Design of web questionnaires: The effects of the number of items per screen. *Field Methods, 21*(2), 200-213.

Toepoel, V., & Dillman, D. A. (2010). Words, numbers, and visual heuristics in web surveys. Is there a hierarchy of importance? *Social Science Computer Review, 29*(2), 1-15.

Toepoel, V., Vis, C., Das, M., & van Soest, A. (2009). Design of Web questionnaires. An information-processing perspective for the effect of response categories. *Sociological Methods Research, 37*(3), 371-392.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*(3), 368-393.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007a). Color, labels, and interpretive Heuristics for response scales. *Public Opinion Quarterly, 71*(1), 91-112.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007b). *The impact of the visible: the design of Web surveys*. Paper presented at the Workshop on Internet Survey Methodology.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science, 185*(4157), 1124-1131.

Van Der Vaart, W., & Glasner, T. (2007). Applying a timeline as a recall aid in a telephone survey: a record check study. *Applied Cognitive Psychology, 21*(2), 227-238.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics, 21*(3), 367-387.

Waksberg, J. (1978). Sampling methods for random digit dialling. *Journal of the American Statistical Association, 73*(361), 40-46.

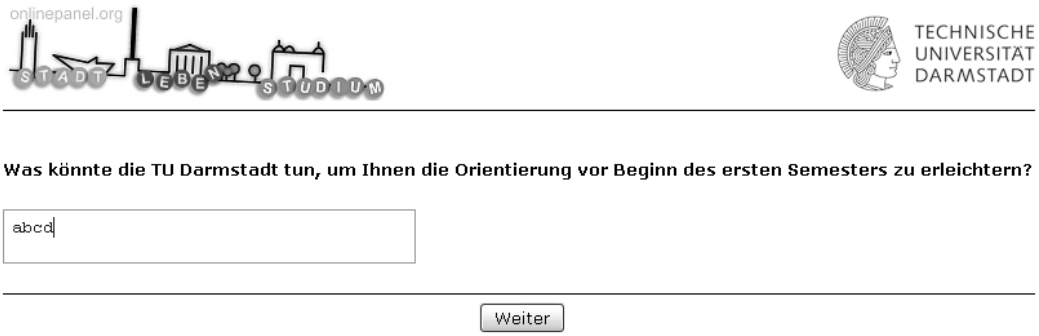
Zickuhr, K., & Smith, A. (2012). *Digital differences.* Washington, D.C.: Pew Research Center.

APPENDIX

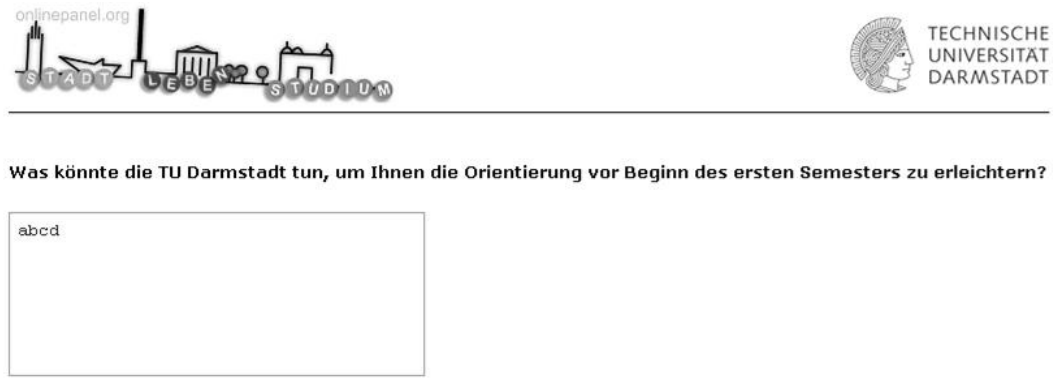
Appendix A: Screenshots of Experiment I

Translation: *"In what ways could the university assist new students?"*

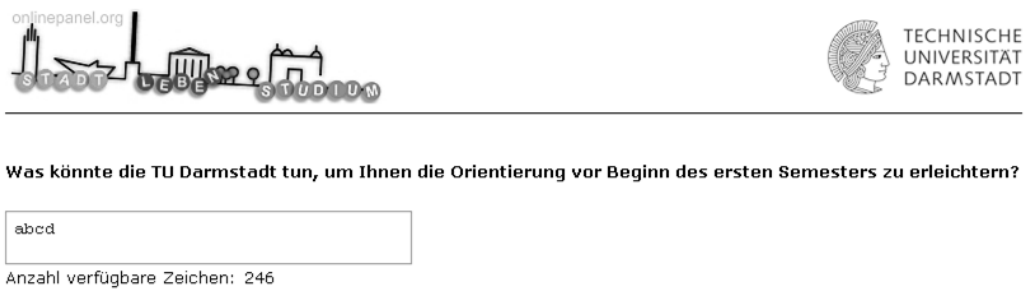
A.1 Small answer-box



A.2 Large answer-box

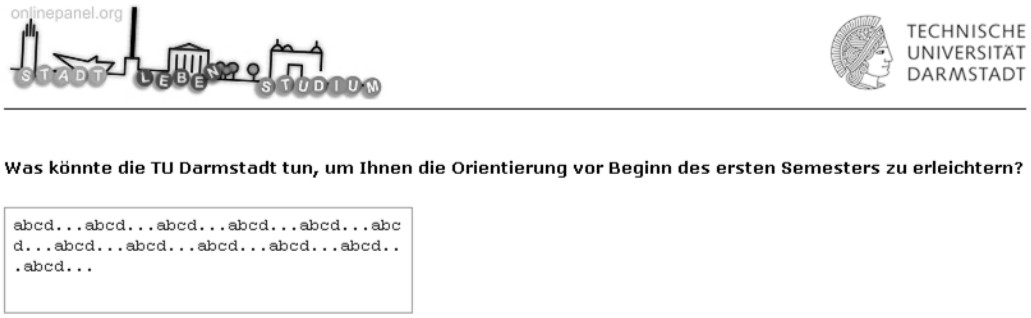


A.3 Small box with counter



A.4 Dynamic box after 4 typed characters

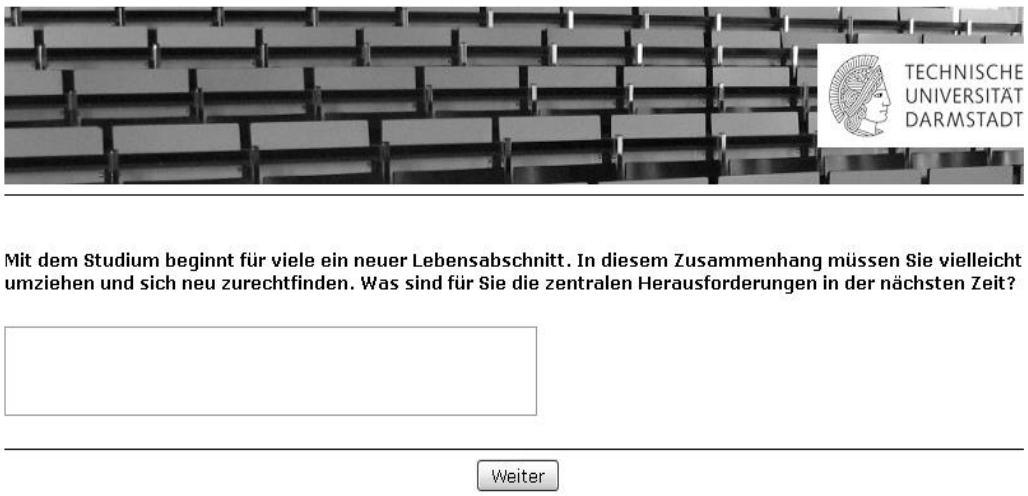
A.5 Dynamic box after 84 typed characters



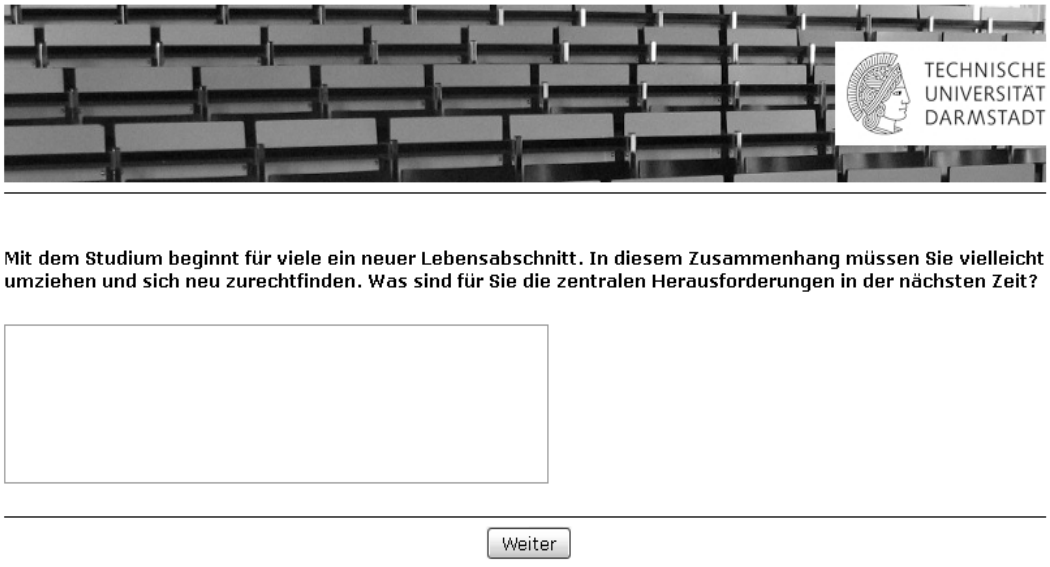
Appendix B: Screenshots of Experiment II

Translation: "*When beginning your studies, you might have to relocate and find a way to manage this new situation. What do you think are the challenges for you in the near future?*"

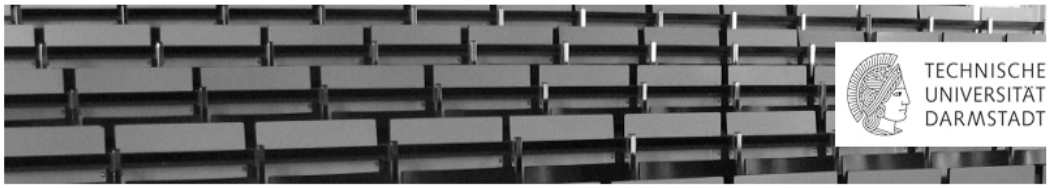
B.1 Small answer-box



B.2 Medium answer-box



B.3 Large answer-box with counter



Mit dem Studium beginnt für viele ein neuer Lebensabschnitt. In diesem Zusammenhang müssen Sie vielleicht umziehen und sich neu zurechtfinden. Was sind für Sie die zentralen Herausforderungen in der nächsten Zeit?

hallo...

Anzahl verfügbare Zeichen: 592

Weiter

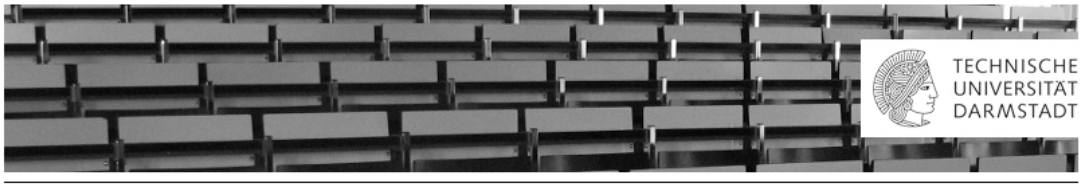
Appendix C: Screenshots of Experiment III

Translation : *"Please name the reasons you apply to study at Darmstadt University of Technology?"*

C.1 Small answer-box (initial size of the small dynamic growing answer-box)

C.2 Medium answer-box (initial size of the medium dynamic growing answer-box)

C.3 Medium respondent-adjusted answer-box (initial size of the large dynamic growing answer-box and the large answer-box)



Bitte erläutern Sie die Gründe, die für die Bewerbung an der Technischen Universität Darmstadt ausschlaggebend waren.
(Sie können die Größe des Texfeldes mit +/- passend einstellen)

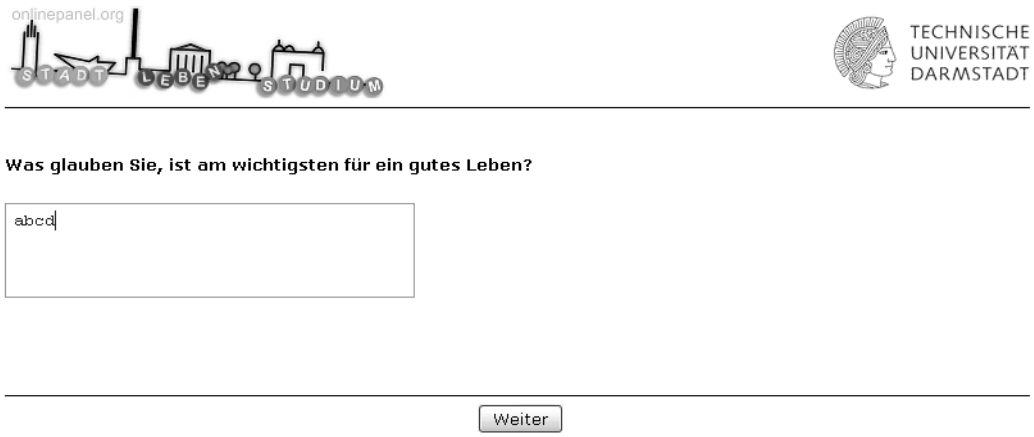
+ -

Weiter

Appendix D: Screenshots of Experiment IV

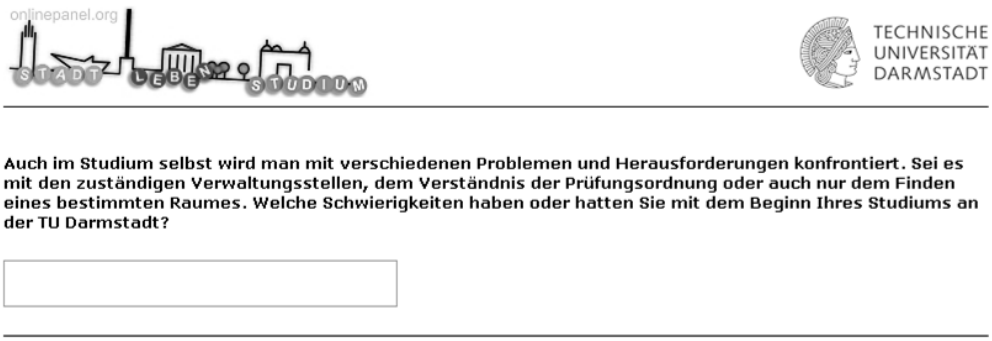
Translation: "*In your opinion, what is most important to lead a good life?"*

D.1 Adaptive design first question: medium answer-box

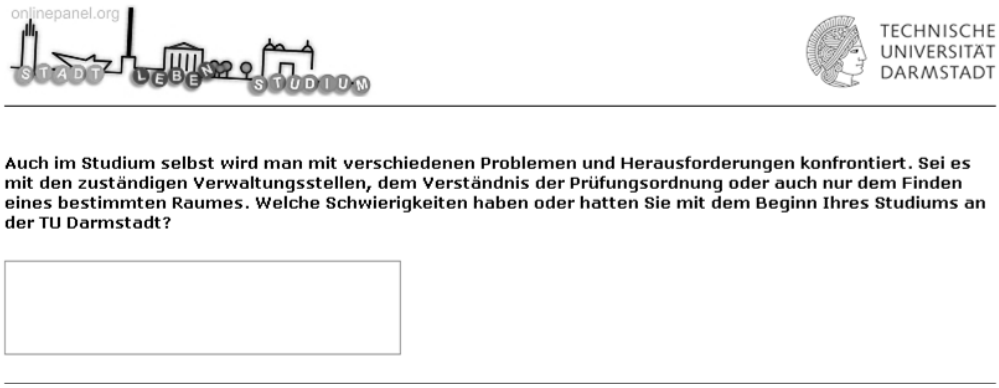


Translation: *"In your opinion, what is essential to succeed in academic studies?"*

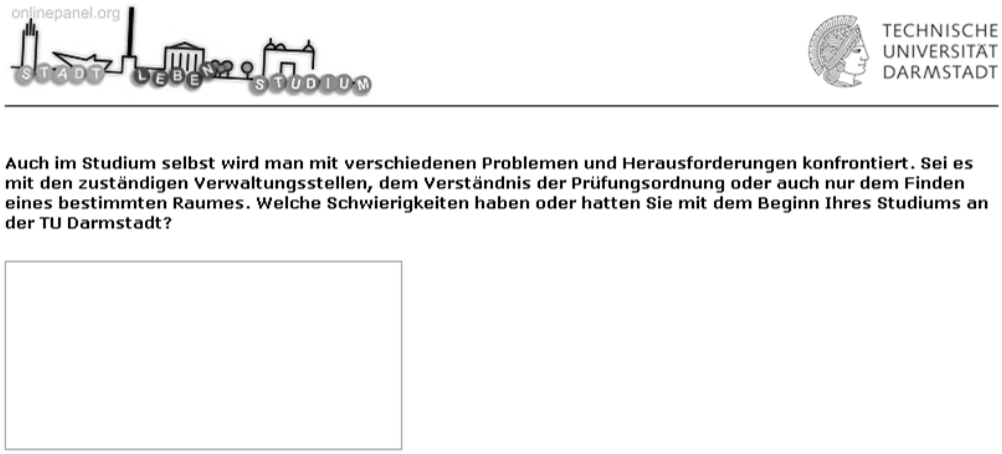
D.2 Adaptive design second question: small answer-box



D.3 Adaptive design second question: medium answer-box



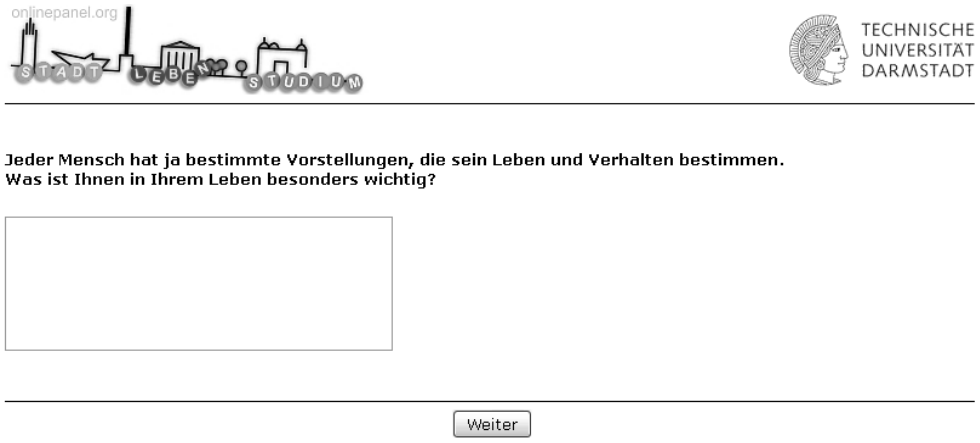
D.4 Adaptive design second question: large answer-box



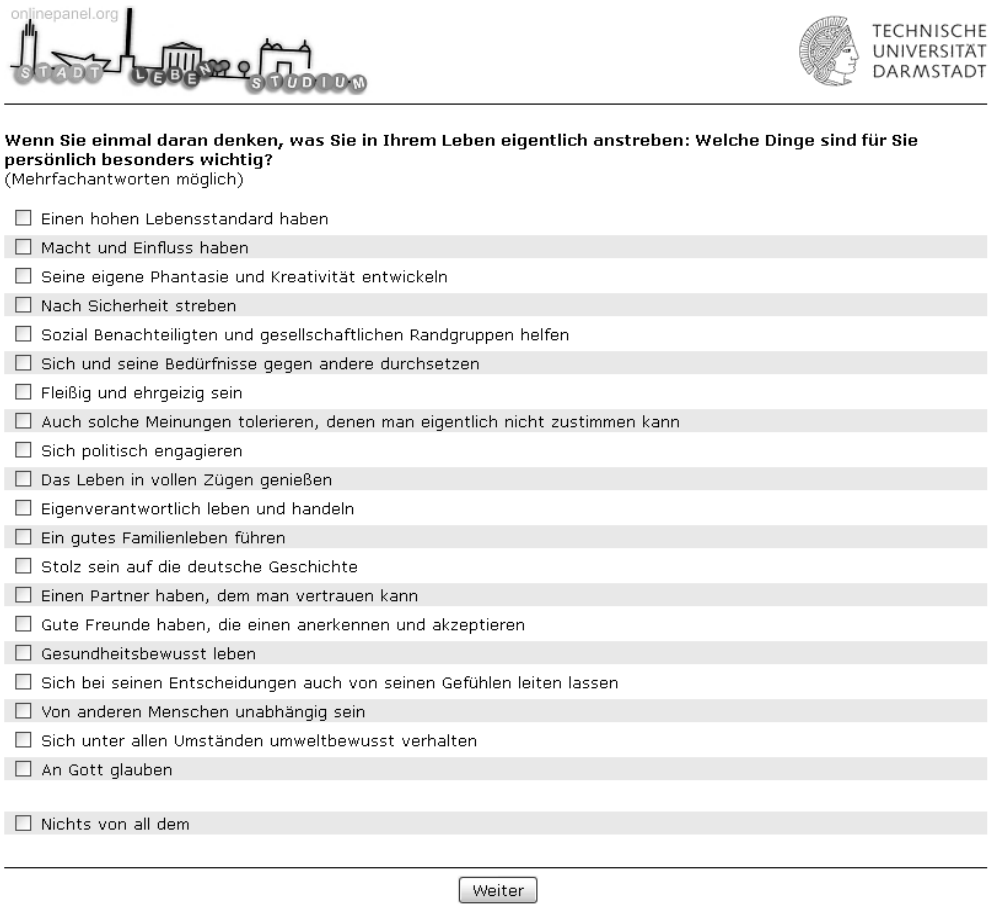
Appendix E: Screenshots of Experiment V

Translation: “Everybody has some major things to accomplish in his life: What are yours?”

E.1 Adaptive design: open-ended question



E.2 Adaptive design: closed-ended probe (see Table 15 for translation of response categories)



Appendix G: Erklärung (Declaration)

Hiermit bestätige ich, dass ich diese zur Promotion eingereichte Arbeit selbständig verfasst habe. Es wurden nur die angegebenen Quellen und Hilfsmittel benutzt und alle übernommenen Zitate wurden als solche gekennzeichnet.

Die vorgelegte Dissertation ist bisher weder ganz noch teilweise als Dissertation oder sonstige Prüfungsarbeit eingereicht worden. Es wurde von mir noch kein Promotionsversuch, auch nicht an einer anderen Universität unternommen.

Hamburg, den 02.05.2014

Appendix H: Lebenslauf (curriculum vitae)

Der Lebenslauf ist in der Online-Verison aus Gründen des Datenschutzes nicht enthalten.