HOW MIXED-MODE SURVEYS ARE TRANSFORMING SOCIAL RESEARCH:

THE INFLUENCE OF SURVEY MODE ON MEASUREMENT

IN WEB AND TELEPHONE SURVEYS

By

LEAH MELANI CHRISTIAN

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPY

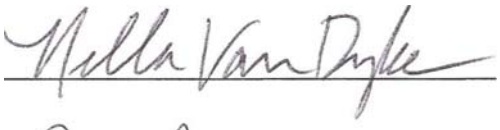
WASHINGTON STATE UNIVERSITY
Department of Sociology

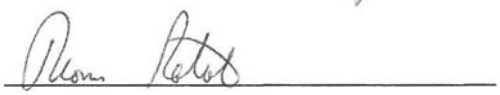
AUGUST 2007

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of LEAH MELANI CHRISTIAN find it satisfactory and recommend that it be accepted.

Chair





ACKNOWLEDGMENTS

There are not enough words to adequately thank Don Dillman, my mentor throughout my graduate career at Washington State University. Without his support, I can honestly say this dissertation would never have been written, and I might not have ventured into the world of survey methodology. Since the summer after my first year when I received the opportunity to work with Don, he has had a profound impact on my professional and personal development. Without his encouragement and assistance, I would likely have left WSU and traveled a different path. I have learned so much from you, Don, and I can only hope that one day I am able to mentor and support graduate students in the ways that you do.

I also want to thank Tom Rotolo for teaching me how to see the world as a methodologist and for his willingness to respond to my numerous questions. Tom was also invaluable in connecting me with Don and survey research. Nella Van Dyke, thank you for coming to WSU and sharing your expertise in political sociology and social movements with all of us. You have encouraged my interests in political issues and indigenous movements, and I look forward to your continued advice on these issues. Nella, you have encouraged my development as a sociologist and have been an amazing mentor during my graduate career. I also want to thank Ray Jussaume. I enjoy every conversation with you Ray and always walk away thinking about an issue in new ways because of your insights.

I cannot imagine the past six years of graduate school without my friend and fellow graduate student, Jolene Smyth. From our first year as officemates in the basement to traveling together and presenting at conferences, you have encouraged my development and challenged me to think about the social world in new ways. I have thoroughly enjoyed working with you and now know the true meaning of collaboration after having worked with you and Don over the past

several years. Many additional friends and fellow graduate students have provided needed fun and insightful conversations on a wide range of social topics. In particular, I want to thank Nick Parsons, Bryan Rookey, Lyssa Thaden and Taj Mahon-Haft. I appreciate the times we have spent together and those yet to come.

My parents, Lauren and Darrel Christian, have guided and encouraged me in all my educational pursuits. Without their continued love and support, I would not have been able to complete a doctoral degree. Thank you, mom, for your emotional and intellectual support! You have inspired me to be a "whole" person. You taught me how to write and organize; two skills I would have never survived graduate school without. Dad, your continued support, has enabled me to pursue undergraduate and graduate education. Your model of hard work and dedication helped me to achieve many goals. To my sister, Kela Quinn, you have constantly helped me grow personally and professionally. I thank you for your friendship and love throughout all of life's ups and downs.

Most importantly, I want to thank my partner and husband, Eugene MacIntyre, for his constant love and support during my graduate education. Thank you for joining me on this journey and for always believing in me, even when my own confidence wavered. Your patience and encouragement throughout this process has been amazing. I also want to thank my in-laws, Dan and Elise MacIntyre, and the rest of the MacIntyre and Cofield families for their support over the past six years.

HOW MIXED-MODE SURVEYS ARE TRANSFORMING SOCIAL RESEARCH:

THE INFLUENCE OF SURVEY MODE ON MEASUREMENT

IN WEB AND TELEPHONE SURVEYS

Abstract

by Leah Melani Christian, Ph.D.
Washington State University
August 2007

Chair: Don A. Dillman

This dissertation is comprised of an introductory chapter that explores the proliferation of mixed-mode surveys, where data is collected using multiple modes, and their influence on social research. Following the introductory chapter, three chapters of journal article length form the main body of the dissertation and address measurement issues facing mixed-mode surveys. Finally, a concluding chapter summarizes the main findings and considers future directions for mixed-mode surveys. The data presented are from methodological experiments embedded in two mixed-mode web and telephone surveys of random samples of Washington State University undergraduate students.

Chapters Two and Three examine how the mode of data collection and the format of response scales influence responses in web and telephone surveys. Respondents surveyed by telephone consistently provide significantly more positive responses than those surveyed by web across a variety of different questions and response scales. In addition, more respondents select the most positive category when all of the categories are labeled and when the scale is presented

in two-steps (respondents first choose the direction and then the intensity of their response) rather than in one-step. These effects are compounded where telephone respondents to the two-step format provide the most extreme responses.

Chapter Four describes the forms of communication available in each mode and addresses how differences in visual and aural communication influence the response process. This chapter presents a multi-stage model that explains how respondents perceive and process visual information and synthesizes research on the visual design of surveys from the past decade to provide a practical framework for surveyors to apply when designing questionnaires. Examples are used to illustrate how visual design can help guide respondents in completing self-administered surveys.

Since social research is often based on data from various types of surveys, threats to survey data quality need to be examined and understood by all social researchers. Features of different survey modes influence responses sometimes threatening the quality of our measurements and therefore the conclusions we draw based on them. Further research is needed to understand the biases that may arise as new technologies and mixed-mode surveys continue to profoundly shape survey research practices.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER ONE

THE IMPACT OF MIXED-MODE SURVEYS ON SOCIAL RESEARCH

The Proliferation of Mixed-Mode Surveys

Surveys are the most frequently used method of data collection in social science research, particularly in the disciplines of sociology and political science (Bradburn and Sudman, 1998). Surveys provide an efficient means of collecting data about a wide range of social topics and their use has increased dramatically over the past fifty years. In addition, once survey data is collected it is frequently archived and made available to the general public or to select groups of people, offering other researchers the opportunity to perform secondary analysis of these data. The use of surveys outside of sociology is extensive; politicians and government officials often rely on survey results when making policy decisions. Businesses utilize surveys for marketing, customer satisfaction research, and to ask questions of their employees. Public opinion surveys influence what topics are debated and how these issues are framed in the media. It is imperative for sociologists and others who utilize survey data to understand the effects of new developments in survey research on the quality of data they analyze.

Recent technological developments in data collection have dramatically influenced survey research practices (Couper, 2005). Survey researchers now have more methods to collect data than ever before and the field has experienced significant advancement in new modes, particularly web surveys. Surveyors now conduct a greater number of mixed-mode surveys where multiple modes are used for data collection and different modes are combined to create various types of data collection strategies (Biemer and Lyberg, 2003; de Leeuw, 2005; Dillman, 2007). Mixed-mode survey designs are extremely flexible; various combinations of modes can

be employed to adapt to the particular needs of each research study. The most common type of mixed-mode design involves surveying members of the same sample using two or more different modes, but some mixed-mode designs involve using multiple modes to survey members of different samples or of subgroups of the population (see de Leeuw, 2005 for a more detailed description of different types of mixed-mode survey designs). Multiple modes can be employed at the same time (e.g. when paper and web versions of the survey are offered concurrently) or they can be employed sequentially over a period of time (e.g. first a paper questionnaire is mailed to respondents and then those who have not responded are called and asked to complete a telephone version of the survey). In addition, longitudinal or panel surveys may switch to alternate modes of data collection after the first wave (Dillman and Christian, 2005).

Mixed-mode surveys are often an optimal choice for researchers who are striving to produce quality data that minimizes survey error while balancing tighter budgets and increasing demand for results to be delivered quickly. Mixed-mode surveys are of particular interest to surveyors looking to compensate for biases in survey estimates that may arise from coverage, sampling, and nonresponse error in single-mode surveys. However, very little research has systematically examined the influence of mixed-mode survey designs on measurement and data quality. The increase in mixed-mode surveys is requiring survey researchers to analyze the effects of using multiple data collection modes on all aspects of survey design, data collection, and analysis.

Survey researchers conducting general population telephone and web surveys must face the error that may arise from coverage and sampling biases. The coverage of traditional landline telephone surveys is threatened as more and more households discontinue their landline telephone service and individuals use only their cell phones (Pew, 2006). Blumberg and Luke

(2006) estimate that the number of households who do not have landlines is now around 16% and the number of cell-only households is about 13%. Of particular concern are the age differences between landline and cell users; half of all cell-only adults are under the age of thirty (Blumberg and Luke, 2006; Pew, 2006). Sampling has also become more difficult as the development of mixed prefixes and number portability has made it increasingly difficult to separate cell from landline numbers. Lastly, cell phone frames sample individuals whereas landline frames sample households and there is no existing file of numbers to construct a list-assisted sampling frame for cell phone users.

Web surveys have proliferated in the past decade and Internet coverage continues to increase; in April 2006, more than 70% of adults in the U.S. were Internet users (Horrigan, 2007). However, there are several concerns with using the web for general population surveys. Web surveys continue to be limited by covering only those people with access to the Internet. In addition, a sampling frame of email addresses that covers all members of the U.S. population has yet to be developed. Similar to cell phones, email addresses are more often used by one individual rather than by all members of a household. Thus, it is important for surveyors to understand the potential differences between those people that are covered and those that are not covered in each mode. Studies have found differences between the general population and cell phone users (Pew, 2006) and Internet users (Horrigan, 2007) on key social demographics (such as age, education, and income) and geographic variables (such as region and whether people live in urban or rural areas).

In addition to improving biases from sampling and coverage in single-mode surveys, mixed-mode surveys can also help correct potential biases from nonresponse error. The development of new communication devices and modifications to existing ones are also

profoundly impacting people's preferences for completing surveys. Based on people's differential access to, use of, and familiarity with different survey modes, some may prefer to respond via one mode and others by another mode. Response rates to landline telephone surveys have also been declining as more respondents use call-screening devices to select which calls they will answer. Overall, surveyors are utilizing mixed-mode survey designs to help increase response rates and decrease coverage, sampling and nonresponse error.

Survey designs involve choosing the optimal mode or combination of modes while minimizing overall total survey error (from coverage, sampling, nonresponse, and measurement bias). The decision of whether to use multiple modes for data collection involves several issues. Surveyors should consider the best mode or modes for the target population and research question under study: some populations may not have access to a particular mode or may prefer to be surveyed by one mode, contact information may only be available for one mode of communication, and some questions or topics may lend themselves to a particular mode. Survey modes can be chosen to increase coverage of the population of interest (e.g. dual sampling frame designs) and to minimize nonresponse bias resulting from differences between respondents and nonrespondents. Mode decisions are almost always influenced by budget constraints of the particular study; often less expensive modes are employed before more expensive modes to reduce overall data collection costs. Lastly, choices about survey mode are also guided by whether the survey must be fielded within a particular time period (e.g. exit polls) and how quickly the data needs to be collected (web and telephone surveys tend to have a shorter data collection period than mail and in-person interview surveys).

The Impact of Mixed-Mode Surveys on Measurement

Mixed-mode surveys, like other developments in survey research, offer new possibilities for surveyors, but they also introduce new challenges and issues that require additional research. It is important that survey researchers conducting mixed-mode surveys balance improvements in coverage, sampling, and nonresponse with the potential for introducing measurement error that may arise from combining data collected by different modes. Mode effects arise because social, cultural, and technological factors associated with particular modes influence how respondents complete the survey response process. There are several features that vary depending on the mode of administration and can help us understand why responses to questions might differ across modes. These features can be grouped into three general categories: media-related, interviewer impact, and information transmission (de Leeuw 1992; 2005); these groups are not independent as the various types of factors influence and relate to one another. It is important to examine how specific features influence responses such that survey modes can be compared in light of these features.

Media-related factors include respondents' familiarity with and use of the mode itself and whether the respondent or an interviewer controls the delivery of the survey. In face-to-face and telephone surveys, interviewers control the delivery of the stimulus including the order in which questions are read to respondents as well as the pace and flow of the conversation (de Leeuw, 1992; Dillman, 2007). However, in mail and web surveys the respondent controls the survey conversation, determining what information to process, the order in which questions are completed, and how quickly they move through the survey (Dillman, 2007).

The impact of interviewer presence in face-to-face and telephone surveys and absence in mail and web surveys can have potential advantages and disadvantages. In telephone and face-

to-face surveys interviewers may facilitate communication between the researcher and the respondent and increase respondent motivation by clarifying questions and respondent answers. However, their presence may also invoke norms of social interaction where respondents tend to provide more culturally acceptable or socially desirable answers (de Leeuw, 2005; Schwarz, Strack, Hippler and Bishop, 1991). In contrast, web and mail survey respondents have more control and privacy making them less likely to be influenced by interactional norms and social desirability; however, respondents must rely solely on the questionnaire itself to infer the researcher's intentions and expectations (de Leeuw, 1992). Recent research on the Internet has focused on utilizing the dynamic nature of web surveys to program various types of interactive features (Conrad, Couper, Tourangeau and Galesic, 2005; Couper, Traugott, and Lamias, 2001; Crawford, Couper, and Lamias, 2001) and to simulate the presence of an interviewer (Krysan and Couper, 2004). The interactivity of web surveys can be used to help increase respondent motivation, provide clarifications or feedback to respondents, and decrease potential confusion and survey terminations.

Survey modes also differ in how information is transmitted between the researcher and the respondent and thus the cognitive stimulus respondents receive (de Leeuw, 2005). Information can be transmitted between the surveyor and the respondent using aural communication, visual communication, or a combination of aural and visual communication. Aural transmission of information requires higher demands on respondents' memory capacity than visual transmission because they must remember the information rather than being able to repeatedly refer to it in the questionnaire (Schwarz et al., 1991). Words are important in aural and visual communication. Additional meaning can also be conveyed visually by other elements, such as: question numbers, arrows to direct respondents, circles or boxes for where respondents

mark their answer, and other shapes and symbols. In additional, the presentation of visual information can be manipulated, such as: the location, contrast, color, size, and font of information.

One of the central issues facing survey researchers is mixed-mode surveys and their influence on data quality. Surveyors need to continue to examine the effects of using different combinations of data collection modes to improve coverage and decrease sampling error and to decrease nonresponse bias. One of the most critical issues facing surveyors conducting mixed-mode surveys is how to best design questionnaires that minimize measurement differences due to particular features of individual survey modes. Decisions to employ mixed-mode survey designs should be made with attention to total survey error, balancing the four sources of error with the needs, cost and time constraints of each particular research project.

Analyzing Survey Mode Effects in Web and Telephone Surveys

The research presented here stems from a series of issues that I have examined over the past five years under the faculty guidance of Dr. Don A. Dillman, and with other graduate students at Washington State University (WSU). This research is supported by USDA-National Agricultural Statistics Service and the National Science Foundation Division of Science Resource Statistics under Cooperative Agreement #43-3AEU-1-80055 with Don A. Dillman as the principal investigator. The data for this research comes from a series of surveys conducted of random samples of undergraduate students at WSU. Multiple versions of each survey were developed so that we could embed methodological experiments testing different question constructions in each of these surveys; students were randomly assigned to complete one version of each survey. Our methodological research focused on different ways of constructing questions

(particularly scalar questions and multiple response questions) as well as the influence of visual design on how people respond to surveys. We have been able to incorporate new ideas for experimentation with each survey, and we have also been able to build upon previous findings with experiments in subsequent surveys since the survey was repeated every academic year. We conducted three single-mode surveys, including one paper and two web surveys, between 2001 and 2003 to test various question formats and design features within self-administered surveys (see the top of Figure 1.1 for general details about these three surveys). The web surveys tested many of the same ideas from the paper survey, but we were also able to test new design experiments unique to the web.

Figure 1.1: Design and Implementation Details for Student Experience Surveys

Survey Date	Survey Mode	Experimental Versions	Number of Questions	Sample size	Completed Responses	Response Rate (AAPOR 1)
Spring 2001	Paper	4	42	1,800	1,042	57.9%
Spring 2002	Web	4	21	3,004	1,591	53.0%
Fall 2003	Web	4	25	3,045	1,705	50.1%
Fall 2004	Web	3	25	1,800	1,054	58.6%
	Phone	3	25	1,608	945	58.8%
Spring 2006	Web	4	27	2,400	1,369	57.0%
	Phone	3	27	1,600	847	52.9%

After the three single-mode surveys, we conducted two mixed-mode web and telephone surveys in the Fall of 2004 and the Spring of 2006 to further understand how different question formats operate across modes (see the bottom of Figure 1.1 with the gray background). To test for mode effects in the mixed-mode surveys, we needed a population with equal access to

completing either a web or telephone version of the survey (i.e. a population with both telephone and web access so we could randomly assign respondents to a web or phone version). Thus, students, who all have web access through the university, provided an ideal population for this research. The opportunity to work on a series of surveys, conducted using different modes, was important in helping me learn all stages of the survey research process – from the development of concepts and design of the survey questionnaires (including methodological experiments) to implementing the contact strategy and monitoring data collection for web and telephone, as well as analysis of the data, summarizing results for conference presentations and writing articles for publication. Additional details about the methodology of the surveys and information about specific methods of data analysis are presented in Chapters Two, Three, and Four and in Appendix A.

In this dissertation, I address an issue that has received little attention by survey researchers conducting mixed-mode surveys. I examine how different features of web and telephone surveys influence measurement. This dissertation allows me to explore questionnaire construction differences in a mixed-mode setting to build upon previous single mode findings by evaluating how specific question formats perform in mixed-mode surveys involving web *and* telephone data collection. The format for this dissertation is an article format where Chapter One provides an introduction to mixed-mode surveys and an overview of the dissertation, Chapters Two, Three and Four are written as article-length journal papers. In Chapters Two and Three, I examine the effects of question format and survey mode on responses to scalar questions within and across web and telephone surveys. In Chapter Four, I discuss in detail how the type of communication influences the survey response process, and present a conceptual framework for understanding how survey respondents process visual information presented to them in a

questionnaire. I conclude with Chapter Five, which provides a brief summary presenting the main conclusions of the dissertation and considers future directions for mixed-mode survey research.

Chapters Two and Three continue research on how different formats for asking response scales influence the answers respondents provide. Previous analyses comparing various ways of constructing response scales in paper (Christian and Dillman, 2004) and web (in my master's thesis Christian, 2003; Dillman and Christian, 2005) single-mode surveys have been presented elsewhere. In Chapters Two and Three, I analyze data from the mixed-mode surveys to extend my previous research by analyzing different scale formats across web and telephone survey modes. These two chapters focus on understanding how the medium and type of communication used to transmit information influence how survey respondents answer scalar questions in web and telephone modes and why respondents answer more positively when surveyed by telephone than web. The results of these experiments using different scalar constructions will help inform the theoretical understanding of how survey mode and scale format can produce measurement error and how different question constructions influence respondents' reports of their opinions and attitudes.

Chapter Two is a paper that I presented at the second annual Telephone Survey Methodology Conference in January, 2006; this paper "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys," with co-authors Don A. Dillman and Jolene D. Smyth, is forthcoming in a monograph from the conference *Advances in Telephone Survey Methodology* and appears in the format accepted for publication. In Chapter Two, I compare responses to five-category fully and polar point labeled scales and eleven-category polar point labeled scales across telephone and web modes. In addition, I also compare

various ways of presenting scales within both web and telephone modes including whether to: provide all or only the endpoint labels, use construct-specific scales instead of agree-disagree scales, and present the positive or negative end first.

In Chapter Three, I continue analyzing different ways of asking scalar questions by comparing responses to a one-step format, where all categories are presented to respondents at the same time (either visually on the web or aurally on the telephone), with responses to a two-step format, where respondents are first asked the direction of their attitude (e.g. positive or negative; satisfied or dissatisfied, etc.) and then asked for the strength or intensity of their attitude (e.g. Extremely, Very, Somewhat, Slightly, etc.). This two-step format is ideally suited for web and computer-assisted telephone interviewing (CATI) where surveyors can utilize computer technology to branch respondents or interviewers to subsequent follow-up questions based on responses they provide to previous questions.

The purpose of Chapter Four is to advance my previous research analyzing how visual design influences respondents as they complete paper (Christian and Dillman, 2004) and web surveys (Christian, Dillman and Smyth, 2007; Smyth, Dillman, Christian, and McBride, 2006; Smyth, Dillman, Christian, and Stern, 2006). Chapter Four will provide a more in-depth analysis of how differences in information transmission between survey modes influence respondents as they complete the steps in the survey response process. To analyze how the type of communication influences the survey conversation, this chapter will briefly cover the differences in how aural and visual information are perceived and processed. The main focus of this chapter is to present a three-stage model illustrating how respondents visually perceive and process visual design elements and their properties in the survey questionnaire at each of these three stages of visual information processing. Specific data reported in this survey build upon the

visual design experiments reported in Christian, Dillman, and Smyth (2007) and in my master's thesis (Christian, 2003).

Chapter Five concludes with a summary of the main findings from Chapters Two, Three, and Four and provides suggestions for future research on the format for rating scales and the visual design of surveys. This chapter discusses the implications of these findings on our understanding of how features of different survey modes influence respondents and for designing survey questions that translate effectively across modes. In addition, the importance of improving the visual design of survey questionnaires is highlighted with attention to understanding how different visual design decisions impact the survey response process. This chapter also considers how changes in survey research practices will influence social research more broadly.

References

Biemer, Paul P. and Lars Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley and Sons.

Blumberg, Stephen J. and Julian V. Luke. 2007. "Wireless substitution: Early release of estimates based on data from the National Health Interview Survey, July – December 2006." *National Center for Health Statistics*. Available from http://www.cdc.gov/nchs.nhis.htm. May 14, 2007.

Bradburn, Norbert M. and Seymour Sudman. 1998. *Polls and Surveys: Understanding What They Tell Us*. San Francisco: Jossey-Bass.

Christian, Leah M. 2003. "The Influence of Visual Layout on Scalar Questions in Web Surveys." Master's Thesis: Washington State University.

Christian, Leah M. and Don A. Dillman. 2004. "The Influence of Symbolic and Graphical Language Manipulations on Answers to Paper Self-Administered Questionnaires." *Public Opinion Quarterly* 68: 57-80.

Christian, Leah Melani**,** Don A. Dillman, and Jolene D. Smyth. 2007. "Helping Respondents Get it Right the First Time: The influence of words, symbols, and graphics in web surveys." *Public Opinion Quarterly* 71: 113-125.

Conrad, Fred, Mick Couper, Roger Tourangeau and Mirta Galesic. 2005. "Interactive Feedback Can Improve the Quality of Responses in Web Surveys." Paper presented at the annual meetings of the *American Association for Public Opinion Research*. Miami Beach, FL.

Couper, Mick P. 2005. "Technology Trends in Survey Data Collection." *Social Science Computer Review* 23: 486-501.

Couper, Mick P., M. Traugott, and M. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly* 65: 230-54.

Crawford, Scott D., Mick P. Couper, and M. J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19(2): 146-62.

de Leeuw, Edith D. 1992. *Data Quality in Mail, Telephone, and Face to Face Surveys*. Amsterdam: TT Publications.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix: Data Collection Modes in Surveys." *Journal of Official Statistics* 21: 233-255.

Dillman, Don A. 2007. *Mail and Internet Surveys: The Tailored Design Method*. 2nd ed. Updated with new Internet, Visual, and Mixed-Mode Guide. New York, NY: John Wiley and Sons, Inc.

Dillman, Don A. and Leah M. Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17: 30-52.

Horrigan, John B. Pew Internet and American Life Project. 2007. *A Typology of Information and Communication Technology Users.* Available from http://www.pewinternet.org/PPF/r/213/report_display.asp. May 17, 2007.

Krysan, Maria and Mick Couper. 2004. "Race of Interviewer Effects: What Happens on the Web?" Paper presented at the annual meetings of the *American Association for Public Opinion Research*. Phoenix, AZ.

Pew Research Center for the People and the Press. 2006. *The Cell Phone Challenge to Survey Research*. Available from http://people-press.org/reports/display.php3?PageID=1054. May 17, 2007.

Schwarz, Norbert, Fritz Strack, Hans-J. Hippler, and George Bishop. 1991. "The Impact of Administration Mode on Response Effects in Survey Measurement." *Applied Cognitive Psychology* 5:193-212.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian and Mallory McBride. 2006. "Open-Ended Questions in Web and Telephone Surveys." Paper presented at the *World Association for Public Opinion Research* Annual Meeting. Montreal, Canada.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Michael J. Stern. 2006. "Effects of Using Visual Design Principles to Group Response Options in Web Surveys." *International Journal for Internet Science and Research* 1(1): 6-16.

CHAPTER TWO

THE EFFECTS OF MODE AND FORMAT ON ANSWERS TO SCALAR QUESTIONS IN TELEPHONE AND WEB SURVEYS[1]

Introduction

The use of mixed-mode surveys has become increasingly popular as surveyors adapt to rapid technological advances in survey methodology and the changing lifestyles of survey respondents (de Leeuw, 2005; Biemer and Lyberg, 2003). The trend toward conducting mixed-mode surveys was only beginning at the time of the first Telephone Survey Methodology Conference in 1986 (Dillman and Tarnai, 1988), but has proliferated since then with the creation of web surveys. Utilizing multiple mode(s) to collect data from respondents allows survey designers to increase response rates and sometimes data quality by exploiting the strengths of particular survey modes while remaining within the time and cost constraints of a study.

Since the data are often combined for analysis, the increased use of mixed-mode survey designs raises concern about whether respondent characteristics are being measured equivalently across modes. Previous research indicates that the mode of data collection can influence how respondents answer survey questions (de Leeuw, 1992). Generally, the various survey modes differ with respect to technical and cultural factors related to the media or mode itself, the impact of interviewer presence (or absence), and how information is transmitted or conveyed during the survey (de Leeuw, 2005). Understanding the effects of these three types of mode effects can help us evaluate the equivalency of data collected across modes. In addition to the effects of

[1] Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. Forthcoming. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In J.P. Lepkowski, C. Tucker, J.M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, M.W. Link, and R.L. Sangster (Eds.), *Advances in Telephone Survey Methodology*. New York: Wiley.

these specific survey mode factors, survey designers often introduce question format effects in mixed-mode surveys by constructing questions differently depending on the mode being used to survey respondents.

Scalar questions are one of the most commonly used types of survey questions and are frequently formatted or constructed differently across modes to maximize the effectiveness of particular questions for each mode. For example, since no visual aid is typically available for telephone interview respondents, response scales are oftentimes simplified by providing only the polar endpoint labels to ease the administrative task for interviewers and the cognitive and memory burden placed on respondents. However, response scales on web surveys are often presented with all of the scale points verbally labeled while for face-to-face respondents the scale might be presented visually on a show card also with all of the categories labeled. Research on scalar questions suggests that we may expect differences in responses because of the differential labeling of response categories and the overall visual presentation of the scale (Christian and Dillman, 2004; Dillman and Christian, 2005; Krosnick and Fabrigar, 1997; Tourangeau, Couper, and Conrad, 2004). Moreover, the effects of survey mode and question construction or format may both independently and/or jointly influence responses.

Our purpose in this paper is to assess the equivalency of data collected using different scalar formats within and across both telephone and web modes. More broadly, our objective is to contribute to theoretical understandings of how differences among survey modes *and* the scalar question formats influence responses to survey questions. The experimental comparisons were designed to identify ways of asking scalar questions that present equivalent stimulus to respondents across telephone and web modes so that optimal questionnaire designs can be constructed for mixed-mode surveys.

Overview of Experimental Comparisons and Procedures

The popularity and frequent use of scalar questions means that many ways of constructing them have been developed over the years. In this paper, we analyze 70 experimental comparisons from one survey using six versions (three telephone and three web) of thirteen scalar questions. We include comparisons of *similar scales across telephone and web modes* and comparisons of *different scales within these modes* to test the independent effects of both mode and scale format, and we also test for interaction effects of survey mode and scale format,

To test for mode effects we compare telephone and web responses to five-point fully and polar point labeled scales as well as eleven-point polar point labeled scales (see Figure 2.1). To test for scale format effects and provide insights into how various aspects of constructing scalar questions can influence responses, we compare different ways of formatting or constructing scales such as whether to label all of the categories or only some, whether to utilize verbal and numerical labels, and the overall presentation of the scale (see Figure 2.2). The comparisons together provide insights into constructing scales optimally for mixed-mode surveys. Since we include a large number of comparisons in this paper and to facilitate connections between the theory and results, we summarize general procedures here but present our discussion of previous research, theoretical rationales, and detailed description of the manipulations with the results for each set of comparisons. We group the comparisons into two sections: mode effects and scalar format effects and have included a summary figure (Figure 2.3) of our experimental comparisons and key findings.

Figure 2.1: Summary and examples of experimental comparisons testing the mode effects across telephone and web

	Telephone example	Web example
a. Fully labeled scales *9 comparisons* 3 satisfied/dissatisfied 3 agree/disagree 3 construct-specific	How satisfied are you with Washington State University as a place to go to school? Would you say you are … Very Satisfied Somewhat Satisfied Neutral Somewhat Dissatisfied Very Dissatisfied	How satisfied are you with Washington State University as a place to go to school? Very Satisfied Somewhat Satisfied Neutral Somewhat Dissatisfied Very Dissatisfied
b. Polar point labeled scales *13 comparisons* 6 satisfied/dissatisfied 3 agree/disagree 2 extremely/not likely 2 best/worst possible *4 comparisons* 11 category polar point w/ midpoint also labeled "average" on two comparisons best/worst possible	On a 5-point scale, where 5 means very satisfied and 1 means very dissatisfied, how satisfied are you with Washington State University as a place to go to school? You may use any of the numbers (read slowly) 5, 4, 3, 2, or 1.	How satisfied are you with Washington State University as a place to go to school? 5 Very Satisfied 4 3 2 1 Very Dissatisfied

Figure 2.2: Summary and examples of experimental comparisons testing the effects of various scalar formats

a. Fully labeled vs. Polar point labeled scales

See Figure 2.1 a and b

6 telephone comparisons
6 web comparisons
3 satisfied/dissatisfied
3 agree/disagree

	Telephone example	**Web example**
b. Agree/disagree vs. Construct-specific scale *3 telephone comparisons* *3 web comparisons*	To what extent do you agree or disagree that your instructors are accessible outside of class. Would you say you … Strongly Agree Agree Neutral Disagree Strongly Disagree How accessible are your instructors outside of class? Would you say … Very Accessible Somewhat Accessible Neutral Somewhat Inaccessible Very Inaccessible	To what extent do you agree or disagree that your instructors are accessible outside of class. Strongly Agree Agree Neutral Disagree Strongly Disagree How accessible are your instructors outside of class? Very Accessible Somewhat Accessible Neutral Somewhat Inaccessible Very Inaccessible
c. Assigning the most positive category 5 versus 1 *3 telephone comparisons* *3 web comparisons* satisfied/dissatisfied	On a 5-point scale, where 5 means very satisfied and 1 means very dissatisfied, how satisfied are you with the quality of advising you have received as a WSU student? You may use any of the numbers (read slowly) 5, 4, 3, 2, or 1. On a 5-point scale, where 1 means very satisfied and 5 very dissatisfied, how satisfied are you with the quality of advising you have received as a WSU student? You may use any of the numbers (read slowly) 1, 2, 3, 4 or 5.	How satisfied are you with the quality of advising you have received as a WSU student? 5 Very Satisfied 4 3 2 1 Very Dissatisfied 1 Very Satisfied 2 3 4 5 Very Dissatisfied

	Telephone example	Web example
d. Presenting the most positive vs. most negative category first *2 telephone comparisons* *2 web comparisons* extremely/not at all likely	On a 5-point scale, where 5 means extremely likely and 1 not at all likely, how likely are you to continue to attend WSU until you finish your degree? You may use any of the numbers (read slowly) 5, 4, 3, 2, or 1. On a 5-point scale, where 1 means not at all likely and 5 means extremely likely how likely are you to continue to attend WSU until you finish your degree? You may use any of the numbers (read slowly) 1, 2, 3, 4, or 5.	How likely are you to continue to attend WSU until you finish your degree? 5 Extremely likely — 1 Not at all likely 4 — 2 3 — 3 2 — 4 1 Not at all likely — 5 Extremely likely
e. <u>TELEPHONE ONLY</u> **Instruction vs. No Instruction** *6 comparisons* satisfied/dissatisfied	On a 5-point scale, where 5 means very satisfied and 1 means very dissatisfied, how satisfied are you with Washington State University as a place to go to school? You may use any of the numbers (read slowly) 5, 4, 3, 2, or 1.	On a 5-point scale, where 5 means very satisfied and 1 means very dissatisfied, how satisfied are you with Washington State University as a place to go to school?
Direction of instruction *2 comparisons* extremely/not at all likely	On a 5-point scale, where 5 means extremely likely and 1 not at all likely, how likely are you to continue to attend WSU until you finish your degree? You may use any of the numbers (read slowly) 5, 4, 3, 2, or 1.	On a 5-point scale, where 5 means extremely likely and 1 not at all likely, how likely are you to continue to attend WSU until you finish your degree? You may use any of the numbers (read slowly) 1, 2, 3, 4 or 5.
f. <u>WEB ONLY</u> **Polar point vs. number box (web)** *6 comparisons* satisfied/dissatisfied	How satisfied are you with Washington State University as a place to go to school? 5 Very Satisfied 4 3 2 1 Very Dissatisfied	On a 5-point scale, where 5 means very satisfied and 1 very dissatisfied, how satisfied are you with Washington State University as a place to go to school? You may use any of the numbers 5, 4, 3, 2, or 1. ☐
5 versus 1 positive (number box) *2 comparisons* satisfied/dissatisfied	On a 5-point scale, where 5 means very satisfied and 1 very dissatisfied, how do you feel about the quality of <u>instruction</u> in the classes you have taken at WSU? You may use any of the numbers 5, 4, 3, 2, or 1. ☐	On a 5-point scale, where 1 means very satisfied and 5 very dissatisfied, how do you feel about the quality of <u>instruction</u> in the classes you have taken at WSU? You may use any of the numbers 1, 2, 3, 4 or 5. ☐

Figure 2.3: Summary of experimental comparisons and findings

Mode effects (telephone vs. web)	T-tests (difference in means) More positive ratings on the telephone than the web	Chi-square tests (2*2; df=1) More likely to select the most positive category on the telephone than the web
Fully labeled 5-point scales (see Table 2.1)	8 of 9 comparisons significant	5 of 9 comparisons significant
Polar point 5-point scales (see Table 2.2)	11 of 13 comparisons significant	7 of 13 comparisons significant
Polar point 11-point scales (see Table 2.3)	3 of 4 comparisons significant	No significant differences (4 comparisons)

Scale effects	
Fully labeled vs. polar point (see Table 2.4)	More positive ratings to fully labeled than polar point labeled scales 6 of 6 comparisons significant on the telephone; 2 of 6 comparisons significant on the web
Agree/disagree vs. construct-specific (see Table 2.5)	No consistent pattern 2 of 3 comparisons significant on both the telephone and web
Most positive category 5 vs. 1 (see Table 2.6)	More positive ratings when 5 is assigned to the most positive category versus 1 1 of 3 comparisons significant on the telephone; 0 of 3 comparisons significant on the web
Most positive vs. negative first	No significant differences - 0 of 2 comparisons significant on both the telephone and web
Instruction vs. no instruction	(telephone only) Slightly more negative ratings w/ instruction - 1 of 6 comparisons significant
Direction of instruction (5 vs. 1)	(telephone only) No significant differences (2 comparisons)
Polar point vs. number box	(web only) No significant differences (6 comparisons)
Number box (5 vs. 1 positive) (see Table 2.7)	(web only) More positive ratings when 5 (vs. 1) is positive - 2 of 2 comparisons significant

Biemer (1988) has noted that most of the literature on mode effects actually compares two (or more) "systems of data collection" where the overall survey instrument and implementation are adapted to maximize the efficiency of each mode such that the estimation of "pure mode effects" is not possible. To overcome this shortcoming, the survey implementation procedures were standardized across the web and telephone modes. We embedded the experimental comparisons in a mixed-mode survey of random samples of undergraduate students about their experiences at Washington State University's (WSU) Pullman campus conducted in the fall of 2004. We randomly assigned each student to one of six experimental versions (three telephone and three web versions) of a 25 question survey and administered the survey to the telephone and web samples simultaneously. To test for "pure mode effects" we needed a population with equal access to completing either a web or telephone version of the survey (i.e. we needed a population with telephone and web access so we could randomly assign respondents to one of the six versions). Thus, students who all have web access through the university provided an ideal population for this experimental research. Response rates for both surveys were comparable with 59% of the telephone respondents completing the survey (945 completes out of 1,608 sampled) and sixty percent of the web respondents completing the survey (1,082 completes of 1,800 sampled).

All of the students were initially contacted via postal mail letter that included a two-dollar incentive. Telephone respondents were then contacted by the WSU Social and Economic Sciences Research Center's telephone lab to complete the telephone survey. Up to ten callback attempts were made. Web respondents for whom we had an email address (about two-thirds of the sample) were sent an initial email, which included a link to the web survey, in addition to the

initial postal mailing. Subsequent contacts to web nonrespondents were sent using postal mail and e-mail.

An interactive or dynamic design was adopted for the web survey where each question appeared on a separate screen. Questions were presented in black text against a colored background with white answer spaces to provide contrast between the text, answer spaces, and background. All of the screens were constructed using HTML tables where proportional widths were programmed in order to maintain a consistent visual stimulus regardless of individual screen or window sizes. Cascading Style Sheets were used to automatically adjust font size and accommodate varying user browsers and screen resolutions. We controlled access to the web survey by assigning each student an individual identification code they had to input to gain entrance to the survey. Their unique access code was provided in each contact to the respondent.

We performed a series of multinomial logistic regression models for nine comparisons testing both the independent effects of survey mode (telephone or web) and scale format and the potential effects of the interaction of mode and format. We found several significant independent effects of survey mode and question format all consistent with the bivariate analyses; however, none of the interactions of survey mode and scale format were significant. Consequently, we present only the results from the difference of means t-tests and chi-square tests of association (two-by-two, with one degree of freedom) and their associated p-values for both tests of mode and scalar format effects.

Mode Effects

How does survey mode influence data quality?

Since data are often combined in mixed-mode surveys, there is growing concern about whether answers collected from respondents surveyed using different modes are comparable. There are several factors that differentiate modes and can help us understand why responses to questions might differ across modes. These factors can be grouped into three general categories: media-related, interviewer impact, and information transmission (de Leeuw, 1992). Although the various mode-related factors are grouped separately into these three categories for purposes of discussion, these categories are not independent and the various types of factors influence and relate to one another.

Media-related factors include respondents' familiarity with and use of the mode itself, which can be of particular concern for web surveys since computers and the Internet are relatively new technologies only now gaining widespread use and not everyone knows how to use or has access to either or a computer or the Internet. One extremely influential media-related factor is the locus of control. In face-to-face and telephone surveys, interviewers control the delivery of the stimulus including the order in which questions are read to respondents as well as the pace and flow of the conversation (de Leeuw, 1992; Dillman, 2000). However, since mail and web surveys are self-administered, the respondent controls the survey conversation, determining what parts of the questionnaire to process when, how to answer, and how quickly to move through the survey (Dillman, 2000). Telephone interviews are often conducted at a quicker pace because the interviewer and respondent try to avoid silences or lapses in conversation over the telephone whereas in face-to-face interviews, nonverbal communication and interaction between the respondent and interviewer make silences less of a problem (de

Leeuw, 1992; Dillman et al., 1996) and in self-administered surveys, norms of silences are not an issue. Thus, telephone respondents tend to be more susceptible to various types of format effects because of this increased time pressure, a factor that may result in more "top of the head" answers (Hippler and Schwarz, 1998).

The impact of interviewer presence in face-to-face and telephone surveys and absence in mail and web surveys can have potential advantages and disadvantages. In telephone and face-to-face surveys interviewers may facilitate communication between the researcher and the respondent and increase respondent motivation by clarifying questions and respondent answers (see Conrad, Schober, and Dijkstra, 2007 on the effect of paralinguistic cues in survey interviews). However, their presence may also invoke norms of social interaction where respondents tend to provide more culturally acceptable or socially desirable answers (de Leeuw, 2005; also see St-Pierre and Béland, 2007 on social desirability in computer assisted telephone versus personal interviewing). In contrast, web and mail survey respondents have more control and privacy making them less likely to be influenced by interactional norms and social desirability. At the same time though, these respondents must rely solely on the questionnaire itself to infer the researcher's intentions and expectations (de Leeuw, 1992; Schwarz et al., 1991a). Recent research on the Internet, however, has focused on utilizing the dynamic nature of web surveys to simulate the presence of an interviewer (Krysan and Couper, 2004) and provide various types of interactive feedback to help increase respondent motivation, provide clarifications or feedback to respondents, and decrease potential confusion and survey terminations (Conrad et al., 2005; Couper, Traugott, and Lamias, 2001; Crawford, Couper, and Lamias, 2001).

The survey modes also differ dramatically in how information is transmitted between the researcher and the respondent and thus the cognitive stimulus respondents receive (de Leeuw, 2005). The primary difference is whether information is transmitted aurally, visually, or both (Schwarz et al., 1991a; Dillman, 2000; de Leeuw, 1992). Aural transmission of information requires higher demands on respondents' memory capacity than visual transmission because they must remember the information rather than being able to repeatedly refer to it in the questionnaire. In addition to the presentation of stimuli aurally or visually, the various modes utilize different types of communication channels. Both aural and visual information transmission rely on verbal communication - the words used to convey meaning. In addition to verbal language, paralinguistic features such as voice inflection, tone or emphasis, and timing also convey meaning to respondents in face-to-face and telephone surveys (de Leeuw, 1992; Conrad, Schober, and Dijkstra, 2007). During face-to-face surveys nonverbal communication transmitted through the use of gestures, facial expressions, and the body is also an important channel of communication. Similarly, for visual surveys symbolic and graphical languages can act as a type of paralanguage where information can be emphasized, using font, bold, or italics. Additional visual features such as arrows, shapes, size, and graphical location provide other means of transmitting paralinguistic-type information to respondents (Redline and Dillman, 2002).

Face-to-face surveys, often considered the most "flexible" of modes, can utilize aural and visual (e.g. through the use of show cards) transmission and can convey information through verbal and nonverbal languages as well as paralinguistic communication. In comparison, telephone interviews lack visual transmission and non-verbal language cues and instead rely only on aural transmission of information through verbal and paralinguistic communication. In

contrast to face-to-face and telephone surveys, mail surveys totally lack aural communication and instead rely solely on visually transmitted information communicated through verbal as well as symbolic and graphical languages to convey meaning to respondents. Finally, web surveys generally use the same visual transmission of information as mail surveys, relying mostly on verbal, symbolic, and graphical communication, but also have the potential to utilize aural communication or other multimedia technologies such as pictures to simulate facial expressions and other types of nonverbal communication.

Mode effects and scalar questions

Previous research comparing responses to scalar questions has found a mode effect where respondents surveyed by telephone are more likely to provide extreme answers than respondents to modes where the scale is presented visually (either using a show card in a face-to-face interview or as part of the mail or web questionnaire). Specifically, two studies showed that respondents were more likely to select the extreme positive category when surveyed by telephone than by face-to-face interview, where a show card was used to visually display the scale (Groves, 1979; Jordan, Marcus, and Reeder, 1980). In these two studies, the most positive category was presented first. All four categories were verbally labeled (strongly agree, agree, disagree, strongly disagree) in the Jordan, et al. (1980) study and the polar endpoints and midpoint were labeled in the Groves (1979) study using a seven-point satisfaction scale with the most positive category labeled "completely satisfied." De Leeuw (1992) also found that telephone respondents were more likely to select the extreme positive response category ("very satisfied") than face-to-face interview (who received a show card) and mail respondents when all of the categories were verbally labeled and the extreme positive category was the last alternative

presented. To explain these findings, de Leeuw (1992) suggests that it is more difficult for telephone respondents to keep multiple categories in memory since they do not have a visual aid displaying the response options.

Dillman and Mason (1984) found that respondents to telephone *and* face-to-face surveys (where a show card was not used) were also more likely than mail survey respondents to choose the extreme positive option "not a problem" on seven-point polar labeled scales independent of whether the category was mentioned first or last. Further, Tarnai and Dillman (1992) confirm the previous results of Dillman and Mason (1984) and also test the independent effects of "visually" presenting the response scale to respondents by experimentally comparing telephone respondents who received a copy of the questionnaire to use during the interview to telephone respondents who did not receive a copy of the questionnaire and thus did not have the visual presentation of the response scale when being interviewed. They found that providing respondents the questionnaire mitigated some of the mode effect but differences persisted between telephone respondents who received the questionnaire and mail survey respondents (Tarnai and Dillman, 1992).

Additional research has also shown that respondents provide more extreme answers to polar point labeled scales when surveyed by telephone and IVR (Interactive Voice Response) modes than by mail and web where the scale is provided visually (Dillman et al., 2001). Dillman et al. (2001) argue that the respondents give more attention to the internal or middle categories when the scale is displayed visually. Thus, the literature suggests that the combination of increased time pressure and the task of requiring respondents to hold categories in their memory (since no visual aid is usually present) increases respondents' selection of the extreme positive endpoint of the response scale when they are surveyed by telephone (compared to other survey

modes), regardless of whether that category is mentioned first or last or whether full or partially labeled scales are used.

Comparing responses to fully labeled and polar point labeled scales across telephone and web

To experimentally test whether telephone respondents provide more extreme answers than web respondents, particularly whether they are more likely to select the extreme positive category, we compare the results of 22 experimental comparisons of five-point fully labeled and polar point labeled scales using various types of verbal labels (see Figure 2.1). First, we test for mode effects across telephone and web among fully labeled scales with five categories using nine questions with three types of verbal labels: satisfied/dissatisfied, agree/disagree, and construct-specific labels. In both the telephone and web modes, all of the verbal labels are read or presented to the respondents as part of the question stimulus (see Figure 2.1, panel a). In all nine comparisons, we find that respondents to fully labeled scales provide higher ratings when surveyed by telephone than by web (see Table 2.1). Eight of the nine mean comparisons indicate that the telephone ratings are significantly higher and a greater percentage of telephone respondents select the extreme positive category for eight of the nine comparisons with five reaching significance.

We also test for mode effects across telephone and web using polar point labeled scales with five categories for thirteen questions and with four types of verbal labels: satisfied/dissatisfied, agree/disagree, extremely likely/not at all likely, and best possible/worst possible. In both the telephone and web modes, only the positive and negative endpoints are labeled (see Figure 2.1, panel b). Again, we find that telephone respondents provide higher mean ratings than web respondents and that they are more likely to select the extreme category.

Table 2.1: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests of the fifth vs. all other categories (and associated p-values) for fully labeled five-point scales across telephone and web modes

		Telephone							Web							Diff. Means t-test		5+	
		percent of R selecting category							percent of R selecting category										
Q	**Scale type**	**5+[1]**	**4**	**3**	**2**	1	**n**	**Mean**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**		**p**	$\chi^2(1)$	**p**
2	Satisfied/ Dissatisfied	54.5	38.1	5.5	1.3	0.6	310	4.45	48.1	39.3	9.4	2.9	0.3	351	4.32	2.10	.018	2.67	.102
16		67.5	23.8	6.4	1.6	0.7	311	4.56	55.2	28.0	9.8	5.2	1.7	346	4.30	3.87	.000	10.46	.000
24		41.9	34.9	15.8	5.8	1.6	310	4.10	29.8	39.6	15.9	10.4	4.3	346	3.80	3.62	.000	10.57	.001
5	Agree/ Disagree	26.7	50.2	19.9	2.9	0.3	311	4.00	18.8	54.9	20.6	5.1	0.6	350	3.86	2.23	.013	5.78	.016
21		28.0	44.1	24.1	3.2	0.6	311	3.95	25.9	40.1	24.2	6.9	2.9	347	3.79	2.24	.013	.35	.556
25		27.4	43.9	18.0	9.7	1.0	310	3.87	25.2	38.5	21.2	11.6	3.5	345	3.70	2.09	.019	.41	.523
5	Construct-specific[2]	36.1	44.3	13.1	6.2	0.3	321	4.10	24.2	51.3	17.1	7.1	0.3	351	3.92	2.65	.004	11.36	.001
21		43.3	41.1	13.1	1.6	0.9	321	4.24	34.3	47.2	13.5	2.9	2.1	341	4.09	2.36	.009	5.63	.018
25		16.9	43.5	23.1	13.1	3.4	320	3.57	18.7	39.2	21.6	14.9	5.6	342	3.51	.79	.216	.38	.537
	Overall	38.0	40.9	15.8	5.3	1.2		4.09	31.1	42.5	17.4	7.7	2.4		3.92				

[1] 5 is the most positive category and 1 is the most negative category

[2] Label types for construct specific scales are Q5 Accessible/Inaccessible Q21 Central/Not very central Q25 Desirable/Undesirable

In twelve of the thirteen comparisons, we find that telephone respondents provide higher mean ratings than web respondents; eleven are significant (see Table 2.2). Telephone respondents select the extreme positive category more frequently than web respondents for ten of the thirteen comparisons (seven of the thirteen chi-square tests are significant).

We also include four comparisons of eleven-category polar point labeled (worst possible/best possible) scales to test for mode effects across telephone and web (see Figure 2.1, panel b). Similarly to the above findings, we find that respondents provide more positive ratings when surveyed by telephone than by web in all four comparisons with three of the four comparisons reaching statistical significance (Table 2.3). However, responses tend to be distributed among the positive categories (6-10), with none of the chi-square tests of the most extreme positive category significant.

Overall, telephone respondents provide more positive ratings than web respondents for five category response scales in nineteen of the twenty-two comparisons and are more likely to select the most extreme positive category to both fully labeled and polar point labeled scales. In addition, telephone respondents to eleven-category polar endpoint labeled scales also provide more positive ratings than web respondents in three of the four comparisons. These findings confirm previous research that telephone respondents are more likely than mail, web, and face-to-face respondents to select the positive endpoint category. Additionally, the findings appear to be quite robust as telephone respondents provide more positive ratings than web respondents regardless of whether all or only the endpoint categories are labeled and across various types of substantive scales (e.g. satisfaction scales, agree/disagree, construct-specific, etc.).

Table 2.2: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests of the fifth vs. all other categories (and associated p-values) for polar point labeled five-point scales across telephone and web modes

		Telephone							Web							Diff. Means		5+	
		percent of R selecting category							percent of R selecting category										
Q	**Scale type**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**	**t-test**	**p**	**$\chi^2(1)$**	**p**
2		27.6	56.6	13.2	2.6	0.0	311	4.09	32.5	48.0	16.1	2.6	0.8	379	4.09	.11	.542	1.86	.172
4		12.2	49.5	33.8	4.2	0.3	311	3.70	8.2	50.7	34.2	5.8	1.1	377	3.59	1.72	.043	3.02	.082
6	Satisfied/	22.3	43.5	26.1	6.1	1.9	310	3.78	14.9	43.5	29.4	8.8	3.4	377	3.58	2.83	.002	6.27	.012
7	Dissatisfied	21.6	27.1	28.4	15.8	7.1	310	3.40	15.9	28.1	28.7	18.0	9.3	377	3.23	1.86	.032	3.67	.056
16		60.8	25.7	10.9	2.3	0.3	323	4.44	48.0	34.4	11.9	4.1	1.6	369	4.23	3.18	.001	11.13	.001
24		29.2	43.7	21.9	4.2	1.0	311	3.96	21.4	40.8	27.1	7.1	3.6	365	3.69	3.68	.000	5.58	.018
5		25.5	42.6	27.4	4.2	0.3	310	3.89	18.4	47.2	28.5	4.3	1.6	375	3.77	1.86	.031	5.03	.025
21	Agree/ Disagree	21.0	45.2	27.4	4.5	1.9	310	3.79	24.5	33.1	28.8	8.4	5.2	368	3.63	1.98	.024	1.16	.281
25		25.1	38.9	23.2	9.3	3.5	311	3.73	25.9	34.9	25.3	7.1	6.8	367	3.66	.79	.214	.06	.811
18	Extremely/	81.9	10.7	2.6	1.6	3.2	310	4.66	70.0	15.3	9.2	3.2	2.3	347	4.48	2.65	.004	12.60	.000
19	Not likely	54.2	29.7	12.2	2.6	1.3	310	4.33	43.9	32.3	18.6	3.5	1.7	344	4.13	2.75	.003	6.92	.009
9	Best/worst	27.3	35.4	25.2	9.1	3.0	297	3.75	22.4	28.0	32.8	13.0	3.8	339	3.52	2.64	.004	2.01	.157
17	possible	20.5	56.5	21.4	1.3	0.3	322	3.96	8.4	55.7	33.0	2.9	0.0	345	3.70	4.93	.000	19.93	.000
	Overall	33.0	39.2	21.3	5.4	1.9		3.96	39.4	55.1	36.3	10.1	4.6		3.79				

Table 2.3: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests of the fifth vs. all other categories (and associated p-values) for eleven point polar labeled scales across telephone and web modes

		percent of respondents selecting category …													Diff. Means		5+	
Q	**Format**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**n**	**Mean**	**t-test**	**p**	**$\chi^2(1)$**	**p**
9	Telephone	1.4	1.0	2.1	3.4	4.8	15.2	11.0	15.9	22.4	13.1	9.7	290	6.82	2.33	.010	.03	.869
	Web	2.2	0.6	5.5	5.7	3.3	21.0	6.8	13.1	16.7	12.6	12.6	366	6.52				
9*	Telephone	2.2	1.8	3.6	2.2	4.7	8.2	12.2	16.8	20.1	16.5	11.8	279	6.94	3.69	.000	1.37	.241
	Web	2.1	3.8	3.2	6.1	5.6	12.6	10.8	14.6	16.9	12.9	11.4	342	6.46				
17	Telephone	0.0	0.0	0.3	1.6	2.3	6.4	10.6	26.5	29.0	18.1	5.2	310	7.42	1.59	.056	1.83	.176
	Web	0.3	0.0	1.1	0.8	3.5	7.6	10.0	26.8	30.4	14.1	5.4	369	7.28				
17*	Telephone	0.0	0.3	0.0	1.0	0.3	3.9	5.1	18.1	43.1	21.2	7.1	311	7.87	1.16	.123	.02	.881
	Web	0.0	0.0	0.3	1.2	3.2	6.9	8.1	23.9	33.1	18.7	4.6	347	7.47				

* The midpoint was also verbally labeled "average" on these comparisons

Previous research comparing responses from face-to-face and telephone interviews suggests that the presence of an interviewer cannot account for these differences since telephone respondents also provided more extreme answers than face-to-face respondents (who were provided a show card). In addition, most of the questions in our survey would not be considered sensitive questions or ones where we might expect more socially desirable responses when an interviewer is present. However, we cannot rule out a self-selection bias where students with more positive attitudes toward the university were more likely to answer by telephone and students with more negative attitudes were more likely to respond by web. Further, we present the most positive category first on the telephone *and* web so recency, where respondents are more likely to select later items when heard aurally, cannot explain these findings. Thus, the faster pace of the telephone interviews (mean 12.1 minutes on the phone and 21.4 minutes on the web)[2] and the lack of visual aids when answering scalar questions seem together to encourage telephone respondents to answer more extremely than respondents to other modes.

Scalar Format Effects

Response scales are often constructed differently depending on the mode used to survey respondents. Thus, surveyors usually must make several decisions when constructing response scales, such as: whether to label all or only some of the categories, what types of labels to choose, and how to present the scale to respondents. For example, fully labeled scales, where all of the categories are given a verbal label, are used on mail and web surveys and frequently in face-to-face interviews using show cards. Since visual communication can be utilized in these modes, the surveyor can visually present multiple category labels to respondents without increasing memory burden. However, to ease the interviewing task and reduce demands on

[2] To calculate overall response time, outliers were removed at two standard deviations away from the mean.

respondents' memories in telephone surveys where visual aids are not available, the same response scale is often changed to provide verbal labels for only the endpoints and respondents must choose a number corresponding to the category that best represents their answer.

In addition to the across mode comparisons reported above, we also include several comparisons of different types of response scales *within* both telephone and web modes to identify how question format affects respondents answers. These within mode comparisons include fully labeled versus polar point labeled scales, agree/disagree versus construct-specific scales, assigning 5 versus 1 to the most positive category, presenting the positive or negative end of the scale first, the use of an instruction to simulate the visual response scale for telephone respondents versus no instruction, and a number box version where the visual scale is removed for web respondents versus polar point scales. These within mode comparisons of different scalar formats are designed to test various ways of asking scalar questions in an effort to help determine effective formats for constructing response scales that provide equivalent stimulus across modes.

Fully labeled vs. polar point labeled scales

Researchers can include words and/or numbers to label categories and they can choose to label all the categories or only some. Partially labeling the scale by only using verbal labels for some of the categories (endpoints and/or midpoint) may differentially attract respondents to these categories because they tend to be drawn to labeled points and the respondent must interpret what the categories mean (Krosnick and Fabrigar, 1997). However, verbally labeling all of the response categories requires respondents to read, process, and interpret all of the labels. Several studies have concluded that fully labeled scales rate higher on various measures of data quality

(reliability and validity as well as respondent satisfaction) than polar point scales as long as care is taken to choose verbal labels that are not overly specific or too vague (Krosnick and Fabrigar, 1997; Schwarz and Hippler, 1991). Since survey designers often construct scales with all of the labels for surveys where the scale can be presented visually (e.g. mail, web, and face-to-face surveys when a show card is used) and as polar point labeled scales for telephone surveys, we include six comparisons of fully labeled and polar point labeled scales within telephone and web modes (Figure 2.2, panel 2a).

Our telephone survey results indicate that respondents provide significantly more positive ratings, as reflected in higher mean ratings, to fully labeled than polar point labeled scales for all six comparisons (see Table 2.4). Additionally, in all of the comparisons, a greater percentage of respondents to the fully labeled scales select the most positive extreme category than respondents to the polar point labeled scales. The web survey results also indicate that respondents provide more positive ratings to the fully labeled than the polar point labeled scales for all six comparisons but only two of the six difference of means tests are statistically significant (see Table 2.4). This finding appears to be linked to the scale type as the magnitude of the differences between fully labeled and polar point labeled scales is much greater for the satisfaction scales than the agree/disagree labeled scales within the web mode.

Overall, respondents seem to provide more positive ratings to fully labeled than polar point labeled scales, a finding that is particularly robust within the telephone mode where all six comparisons yield significant differences compared to only two of the six comparisons within the web. These findings, in conjunction with previous research suggesting that fully labeled scales are more reliable and rate higher in validity measures, indicate that when polar point scales are provided, respondents may answer more negatively than their actual attitude.

Table 2.4: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests (and associated p-values) for fully labeled vs. polar point scales for telephone and web respondents

			Fully labeled							Polar point							Diff. Means t-test	p	5+ $\chi^2(1)$	p
			percent of R selecting category							percent of R selecting category										
	Q	Scale type	5+	4	3	2	1-	n	Mean	5+	4	3	2	1-	n	Mean				
PHONE	2	Satisfied/ Dissatisfied	54.5	38.1	5.5	1.3	0.6	310	4.45	27.6	56.6	13.2	2.6	0.0	311	4.09	6.15	.000	51.72	.000
	16		67.5	23.8	6.4	1.6	0.7	311	4.56	60.8	25.7	10.9	2.3	0.3	323	4.44	1.87	.031	5.64	.210
	24		41.9	34.9	15.8	5.8	1.6	310	4.10	29.2	43.7	21.9	4.2	1.0	311	3.96	1.82	.035	14.49	.005
	5	Agree/ Disagree	26.7	50.2	19.9	2.9	0.3	311	4.00	25.5	42.6	27.4	4.2	0.3	310	3.89	1.73	.042	6.42	.131
	21		28.0	44.1	24.1	3.2	0.6	311	3.95	21.0	45.2	27.4	4.5	1.9	310	3.79	2.41	.008	6.51	.168
	25		27.4	43.9	18.0	9.7	1.0	310	3.87	25.1	38.9	23.2	9.3	3.5	311	3.73	1.79	.037	7.76	.101
WEB	2	Satisfied/ Dissatisfied	48.1	39.3	9.4	2.9	0.3	351	4.32	32.5	48.0	16.1	2.6	0.8	379	4.09	3.98	.000	21.59	.000
	16		55.2	28.0	9.8	5.2	1.7	346	4.30	48.0	34.4	11.9	4.1	1.6	369	4.23	.95	.171	5.37	.251
	24		29.8	39.6	15.9	10.4	4.3	346	3.80	21.4	40.8	27.1	7.1	3.6	365	3.69	1.36	.087	17.79	.001
	5	Agree/ Disagree	18.8	54.9	20.6	5.1	0.6	350	3.86	18.4	47.2	28.5	4.3	1.6	375	3.77	1.59	.057	8.79	.066
	21		25.9	40.1	24.2	6.9	2.9	347	3.79	24.5	33.1	28.8	8.4	5.2	368	3.63	2.03	.022	6.73	.151
	25		25.2	38.5	21.2	11.6	3.5	345	3.70	25.9	34.9	25.3	7.1	6.8	367	3.66	.54	.294	9.72	.045

In addition to choosing whether to label all or only some of the response categories, researchers must choose what type of substantive labels to use. Surveyors often use agree/disagree, yes/no, and true/false response categories when designing scales because they are easy to administer and can be applied across a variety of question topics. However, research on acquiescence response bias suggests that people have a tendency to agree regardless of the content of the question and particularly in the social presence of someone else (e.g. an interviewer). Saris and Krosnick (2000) argue that scales with construct-specific response options, where verbal labels are designed using the underlying concept or construct being measured, decrease acquiescence response bias and produce less measurement error than agree/disagree scales because the process of respondents mapping their judgment to the appropriate response option should be more accurate when respondents only have to think along one dimension (the underlying construct) instead of along two dimensions (the underlying construct *and* the agree/disagree response options provided). Experimentally testing these two scale formats, Saris and Krosnick (2000) find that construct-specific scales decrease cognitive burden and acquiescence response bias and yield data of higher quality (with higher reliability and validity ratings). To test whether respondents to agree/disagree scales are more likely to acquiesce, we include three comparisons of agree/disagree labeled scales (e.g. strongly agree, agree, neutral, etc.) and construct specific scales (e.g. very accessible, somewhat accessible, etc.) for both telephone and web respondents (see Figure 2.2, panel 2b).

For two of the three comparisons within both the web mode and the telephone mode, the mean ratings are higher for the construct-specific scales than the agree/disagree scales because respondents are more likely to select the most positive category in the construct-specific scales (see Table 2.5). Respondents seem to avoid the "strongly agree" category, and instead most

Table 2.5: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests (and associated p-values) for agree/disagree vs. construct-specific scales for telephone and web respondents

		Agree/Disagree							Construct-specific							Diff. Means t-test		5+	
		percent of R selecting category							percent of R selecting category										
	Q	5+	4	3	2	1-	n	Mean	5+	4	3	2	1-	n	Mean		p	$\chi^2(1)$	p
PHONE	5	26.7	50.2	19.9	2.9	0.3	311	4.00	36.1	44.3	13.1	6.2	0.3	321	4.10	1.46	.072	13.99	.004
	21	28.0	44.1	24.1	3.2	0.6	311	3.95	43.3	41.1	13.1	1.6	0.9	321	4.24	4.39	.000	23.08	.000
	25	27.4	43.9	18.0	9.7	1.0	310	3.87	16.9	43.5	23.1	13.1	3.4	320	3.57	3.78	.000	15.86	.003
WEB	5	18.8	54.9	20.6	5.1	0.6	350	3.86	24.2	51.3	17.1	7.1	0.3	351	3.92	.92	.178	5.34	.236
	21	25.9	40.1	24.2	6.9	2.9	347	3.79	34.3	47.2	13.5	2.9	2.1	341	4.09	4.11	.000	24.49	.000
	25	25.2	38.5	21.2	11.6	3.5	345	3.70	18.7	39.2	21.6	14.9	5.6	342	3.51	2.37	.009	6.41	.170

respondents choose the "agree" category whereas responses to the construct-specific scales are more evenly spread over the two positive categories (e.g. Very and Somewhat Accessible). However, the pattern for the third comparison is different with significantly more positive ratings to the agree/disagree scales than the construct-specific scales within both telephone and web modes and with more respondents choosing the most positive category on the agree/disagree scales. This question asks students about the desirability of Pullman as a place to live while going to school, and students have tended to provide more negative ratings overall when this question was asked in previous surveys. Thus, both the topic of the question (and whether respondents are more positive or negative on that topic) and the type of response options seem to be influencing respondent answers. The across mode results reported earlier in this paper also show that telephone respondents provide more positive ratings than web respondents regardless of whether agree/disagree or construct-specific labels are used.

Some survey designers also choose to use numbers to label response categories in addition to or instead of verbal labels. Previous research testing the effects of numeric labels suggests that respondents interpret the meaning of word labels differently when the numeric labels run from –5 to 5 than 0 to 10 (Schwarz et al., 1991b). Other research on the web suggests that labeling polar point scales 1 to 5 does not produce significantly different answers than when the numeric labels are omitted on polar point scales (Christian, 2003). Overall, scales with numbers often take longer because respondents are required to process additional information, so unless the numeric labels are essential in helping respondents interpret the scale, they seem to unnecessarily increase respondent burden (Krosnick and Fabrigar, 1997).

Respondents tend to culturally associate higher numbers with more positive categories and lower numbers with more negative categories. Thus, it would seem that when numeric

labels are chosen to be consistent with this expectation it should ease the response task. Research using mail surveys suggests that when this a priori expectation is not met and respondents are asked to select a number from 1 to 5 where 1 is very satisfied and 5 is very dissatisfied, they often confuse which end of the scale is positive and which is negative and have to correct their answers (Christian and Dillman, 2004; Dillman and Christian, 2005). To test whether assigning higher versus lower numbers to more positive ratings influences respondent answers, we include three comparisons of five-category polar point labeled scales where the most positive category is numerically labeled 5 versus 1 on both the telephone and web modes (see Figure 2.2, panel 2c). We find that within both the telephone and web modes respondents provide more positive ratings when 5 is assigned to the most positive category; however, only one of the telephone and none of the web comparisons are statistically significant (see Table 2.6). Thus, assigning one versus five to the positive end of the scale does not seem to substantially impact how respondents answer polar point scales on the telephone or web.

Presenting the most negative or most positive category first

There is a tendency for researchers to present the most positive category first on telephone surveys and the most negative category first on mail or web surveys to avoid primacy/recency effects on respondents' answers. However, respondents gain information about each category from its labels and its position in relation to other categories. In other words, they interpret additional meaning from the overall presentation of the response scale; and therefore, their responses may be different depending on whether the positive or negative end of the scale is presented first. Tourangeau, Couper, and Conrad (2004) have suggested five heuristics respondents use to interpret meaning from the visual presentation of the response scale. Two of

Table 2.6: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests (and associated p-values) for polar point reversals for telephone and web respondents

			Polar point 1+ 5- percent of R selecting category							Polar point 1+ 5- percent of R selecting category							Diff. Means t-test	5+		
	Q	Scale type	5+	4	3	2	1-	n	Mean	1+	2	3	4	5-	n	Mean		p	$\chi^2(1)$	p
PHONE	4	Satisfied/ Dissatisfied	12.2	49.5	33.8	4.2	0.3	311	3.70	9.6	47.3	24.8	17.0	1.3	311	3.47	3.28	.000	31.45	.000
	6		22.3	43.5	26.1	6.1	1.9	310	3.78	24.0	40.0	23.4	11.0	1.6	308	3.74	.56	.287	5.59	.228
	7		21.6	27.1	28.4	15.8	7.1	310	3.40	22.7	24.3	27.5	19.7	5.8	309	3.38	.22	.412	2.33	.676
WEB	7	Satisfied	15.9	28.1	28.7	18.0	9.3	377	3.23	15.4	23.7	32.9	18.6	9.4	350	3.17	.70	.241	2.46	.652
	18	Extremely/	70.0	15.3	9.2	3.2	2.3	347	4.48	68.7	14.5	8.4	5.5	2.9	345	4.41	.92	.179	2.66	.616
	19	Not Likely	43.9	32.3	18.6	3.5	1.7	344	4.13	46.2	29.5	15.5	7.3	1.5	342	4.12	.18	.427	6.32	.172

these heuristics, "left and top mean first" and "up means good," suggest that respondents to visual surveys expect scales to begin with the most positive category (i.e., very satisfied) and expect the successive categories to follow logically from that point (i.e., somewhat satisfied, neutral, etc.). Since these heuristics have been only tested using web surveys, it seems important to test whether they also apply to telephone surveys where the response scale is not presented visually.

To test whether presenting/mentioning the positive versus negative category first influences responses, we include two comparisons of scales presenting the most positive versus the most negative category first within both telephone and web modes (see Figure 2.2, panel 2d). Within the telephone mode the means are slightly higher when the positive (versus the negative) category is presented first, but within the web mode, the means are slightly higher when the negative (versus the positive) category is presented first. However, none of the comparisons yield significant differences within the telephone or web modes (analysis not shown[3]). Thus, we find that presenting/mentioning the positive versus negative end of the scale first does not seem to influence how telephone or web respondents answer scalar questions; however, findings from other research on the web has found that designing scales with the positive end of the scale first facilitates the response task, suggesting that additional research is still needed.

Instruction vs. no instruction (telephone only)

Several researchers have found that visually manipulating the presentation of the response scale influences how respondents answer scalar questions in both mail (Christian and Dillman, 2004; Schwarz, Grayson and Knäuper, 1998; Smith, 1995) and web surveys (Christian, 2003; Dillman and Christian, 2005; Tourangeau et al., 2004). The findings from these studies suggest

[3] All analyses not shown are available from the authors upon request.

that response differences might also occur across modes when the scale is presented visually in one mode (e.g. web) and not at all in another mode (e.g. telephone).

Dillman et al. (2000) suggest that an additional instruction such as, "You may use any of the categories 5, 4, 3, 2, or 1 to answer" might help telephone respondents visualize the response scale, thus simulating the visual display seen by web respondents, and strengthening the stimulus for the internal categories when only the endpoint labels are provided in the query. Additionally, having the interviewers read this instruction should help slow the pace of the interview and allow respondents more time to complete the question/answer process, perhaps reducing the number of "top of the head" answers. We discuss six comparisons where telephone respondents are asked five-category satisfaction scales with the polar endpoints labeled with and without the additional instruction, "You may use any of the numbers 5, 4, 3, 2, or 1 for your answer" (see Figure 2.2, panel 2e). Interviewers were instructed to read the instruction slowly, pausing between each number. We also include the results from two comparisons in which we reverse the direction of the instruction (beginning with the most positive versus the most negative category) to determine whether this influences respondent answers (also in Figure 2.2, panel 2e).

Overall, the means are slightly lower for the version with the additional instruction but only one of the six tests is significant (analysis not shown). Thus, including the additional instruction does not seem to influence how respondents answer polar point labeled scales on the telephone. We also find no significant differences in responses when the instruction is presented beginning with the most positive versus the most negative category first (analysis not shown). This particular instruction does not seem to significantly influence telephone responses and thus does not seem to provide a powerful enough stimulus to overcome the lack of visual presentation of the response scale on telephone surveys.

Polar point vs. number box (web only)

Previous research has shown that on mail and web surveys, respondents provide more negative ratings when asked to report a number corresponding to their answer than when the scale is displayed visually with the polar endpoints labeled (Christian and Dillman, 2004; Dillman and Christian, 2005). In these studies, the most positive category was assigned the number 1 and the most negative category the number 5, a format that is inconsistent with respondents' a priori expectations as discussed above. We include six comparisons of polar point labeled and number box scales where the numbers are assigned consistent with respondents' a priori expectations (i.e. where 5 is the most positive category and 1 the most negative category) (see Figure 2.2, panel 2f). We find that respondents provide slightly more negative ratings to the number box scales than the polar point scales but none of the differences are significant for any of the six comparisons (results not shown). Thus, when the numeric labels attached to the scale match respondents expectations (i.e. higher numbers indicate more positive ratings), respondents appear to answer polar point labeled scales and number box entries similarly.

Since previous research indicated that respondents to number box scales confused which end of the scale was positive when the most positive category was numerically labeled 1 in the question stem, we include two experimental comparisons on the web of five-category number box scales where the query indicates that either 5 or 1 is the most positive category. Our results show that web respondents give significantly higher ratings when 5 is labeled the most positive category (see Table 2.7). Respondents are more likely to write "4" when 1 is labeled the most positive category. These findings support previous research that respondents confuse which end of the scale is positive when the numeric labels are not consistent with the verbal labels (i.e. the

Table 2.7: Response differences, t-test of the difference of means (and associated p-values), and chi-square tests (and associated p-values) for number box reversal for web respondents

		Number box 5+ 1-							Number box 1+ 5-							Diff. Means			
		percent of R selecting category							percent of R selecting category							t-test	p	5+ $\chi^2(1)$	p
Q	**Scale type**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**	**1+**	**2**	**3**	**4**	**5-**	**n**	**Mean**				
4	Satisfied/	7.1	48.6	35.4	8.0	0.9	350	3.53	6.6	37.3	33.2	20.6	2.3	349	3.25	4.30	.000	27.31	.000
6	Dissatisfied	11.8	44.7	32.1	9.7	1.7	349	3.55	9.7	36.1	30.4	19.5	4.3	349	3.28	3.80	.000	19.20	.000

highest number is assigned to the most positive category). Once the scales are constructed so that higher numbers indicate more positive ratings, no significant differences are found in how respondents answer polar point labeled and number box versions.

Discussions and Conclusions

In this paper, we included a large number of comparisons in part to integrate various findings from previous research on ways to construct scalar questions. The results from our 70 experimental comparisons indicate that survey mode and scalar format independently influence respondents' answers (see Figure 2.3 for a summary of the experimental comparisons and the main findings). We find no significant interaction effects of mode and format. Overall, we find that telephone respondents provide more positive ratings and are more likely to use the extreme positive endpoint than web respondents to five-category scales (nineteen of twenty-two difference of means comparisons significant) and to eleven-category scales (three of four mean tests significant). This mode effect finding appears quite robust as telephone respondents provide more positive ratings to both five- and eleven-point scales, regardless of whether all or only the endpoint categories are labeled, and across various types of substantive scales (e.g. satisfaction scales, agree/disagree, construct-specific, etc.). These findings confirm previous research that telephone respondents are more likely than mail, web, and face-to-face respondents to select the positive endpoint category. It appears that neither social desirability nor recency can explain the extremeness tendency of telephone respondents. The lack of a show card or other visual presentation to help telephone respondents remember the response categories and perhaps the faster pace of telephone interviews together seem to encourage telephone respondents to select the extreme positive category more frequently than respondents to other modes.

Our comparisons of different scalar formats within modes, particularly in combination with findings from previous research, provide several suggestions for constructing response scales. First, it seems desirable to use fully labeled scales because they are more reliable and yield higher validity ratings than polar point labeled scales (Krosnick and Fabrigar, 1997). In addition, respondents to agree/disagree scales seem to avoid the most positive category "strongly agree" with most respondents choosing the second positive category "agree" whereas respondents to construct-specific scales are more likely to select the first two positive categories. However, since Saris and Krosnick (2000) found that construct-specific scales yield data of higher quality and decrease measurement error, construct-specific scale labels should be used when possible. If numerical labels are also used when labeling scales, it seems optimal to assign higher numbers to more positive categories particularly if no visual presentation of the scale is provided (e.g. the number box format on the web) since this is consistent with respondents' a priori expectations. Finally, our additional instruction on the telephone survey ("You may use any of the numbers 5, 4, 3, 2, or 1 for your answer") is designed to simulate the response scale and help slow down the pace of the telephone interview in an effort to provide greater equivalency across modes; however the non-significant findings suggest that this stimulus is not powerful enough to visually represent the scale to telephone respondents. Further analyses of variance and covariance could also help us understand how these various formats influence respondent answers within and across modes. Also, since our experimental comparisons were tested using students, a population with typically higher levels of education and younger in age (eighteen to twenty-five years), it is important for future research to determine whether these same results are found in samples of the general population both in the United States and internationally.

Overall, the results from our experimental comparisons show that mode effects exist independent of format effects – telephone respondents provide more positive ratings and select the extreme positive category more often than web respondents across various types of scales. Our finding of independent mode effects suggests caution for mixed-mode surveys where data are often combined across modes as it appears that these mode effects cannot be overcome by scale format changes; it seems that none of the scalar formats tested here translate equivalently across telephone and web modes. This suggests that combining responses to scalar questions across telephone and web modes is a tenuous practice. Further, this mode effect seems rather robust historically with telephone respondents providing more extreme positive ratings than respondents to all other modes where the scale is presented visually (either on a show card or directly in the stimulus of the mail or web questionnaire). This suggests that the lack of visual presentation of the scale and perhaps the slower pace are the differentiating factors between telephone and other modes. Additionally, since neither social desirability nor recency can explain these findings, it seems urgent to develop a cohesive theoretical explanation for why telephone respondents provide more extreme positive ratings.

In addition to developing a theoretical explanation for why telephone respondents are more likely to select the positive extreme endpoint, one important direction for future research is to test whether other scalar constructions might mitigate the mode effects found in this paper and by other researchers. Previous research has found that unfolding or branching, where scalar questions are asked using two questions, one asking respondents to indicate the direction of their attitude (e.g. satisfied or dissatisfied) and another question asking respondents to indicate the strength or intensity of their attitude (e.g. very or somewhat … satisfied or dissatisfied), decomposes the response task for respondents making it easier for them to provide an answer

(Groves, 1979; Krosnick and Berent, 1993). In addition, branching has been shown to improve reliability and take less time for mail and telephone respondents to answer than presenting the scale all at once in face-to-face, telephone, and mail surveys. Thus, future experimentation should compare responses to scales using branching on telephone *and* web modes to determine whether this construction decreases the extremeness found for telephone respondents.

Mixed-mode surveys will continue to proliferate as they attract the interest of surveyors attempting to balance the competing demands of survey quality, response rates, and limited budgets in today's survey world. It also seems likely that telephone and web modes will figure prominently in many of the mixed-mode surveys being conducted because these two modes together provide a relative cost efficient combination. As a result, it is increasingly important to understand the effects of mixed-mode designs on methodological quality and to minimize the effects of various changes within different "systems of data collection" on responses to survey questions. In addition, much of the literature on mixed-mode surveys focuses on ways of reducing nonresponse error. However, that focus needs to be balanced with appropriate attention given to understanding the causes and consequences of measurement differences, like those revealed in this paper, to help identify question formats that present the equivalent stimulus to respondents across modes so that responses can be combined for analysis

References

Biemer, Paul P. and Lars Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley and Sons.

Biemer, Paul P. 1988. "Measuring Data Quality." In Robert M. Groves, Paul Biemer, Lars Lyberg, J. Massey, W. Nicholls, and J. Wakesberg (Eds.), *Telephone Survey Methodology*. New York: Wiley. 321-340.

Christian, Leah M. 2003. "The Influence of Visual Layout on Scalar Questions in Web Surveys." Master's Thesis: Washington State University.

Christian, Leah M. and Don A. Dillman. 2004. "The Influence of Symbolic and Graphical Language Manipulations on Answers to Paper Self-Administered Questionnaires." *Public Opinion Quarterly* 68: 57-80.

Conrad, Fred, Michael Schober, and Wil Dijkstra. Forthcoming 2007. "Cues of comprehension difficulty in telephone interviews." *Monograph for International Telephone Survey Methodology II Conference.*

Conrad, Fred, Mick Couper, Roger Tourangeau and Mirta Galesic. 2005. "Interactive Feedback Can Improve the Quality of Responses in Web Surveys." Paper presented at the annual meetings of the *American Association for Public Opinion Research*. Miami Beach, FL.

Couper, Mick P., M. Traugott, and M. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly* 65: 230-54.

Crawford, Scott D., Mick P. Couper, and M. J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19(2): 146-62.

de Leeuw, Edith D. 1992. *Data Quality in Mail, Telephone, and Face to Face Surveys*. Amsterdam: TT Publications.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix: Data Collection Modes in Surveys." *Journal of Official Statistics* 21: 233-255.

Dillman, Don A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. 2nd ed. New York, NY: John Wiley and Sons, Inc.

Dillman, Don A. and Leah M. Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17(1): 30-52.

Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell and Jodi Berck. 2001. "Response Rate and Measurement Differences in Mixed Mode Surveys: Using Mail, Telephone, Interactive Voice Response and the Internet." Available at http://www.sesrc.wsu.edu/dillman/papers.htm.

Dillman, Don A., Roberta L. Sangster, John Tarnai, & Todd H. Rockwood. 1996. "Understanding Differences in People's Answers to Telephone and Mail Surveys." In M. T. Braverman & J. K. Slater (Eds.), *New Directions for Evaluation Serie*s, 70 (Advances in survey research). San Francisco: Jossey-Bass. 46-52.

Dillman, Don A. and John Tarnai. 1988. "Administrative Issues in Mixed-Mode Surveys." In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, II, and J Wakesberg (Eds.) *Telephone Survey Methodology*. New York: Wiley, 509-528.

Dillman, Don A. and Robert G. Mason. 1984. "The Influence of Survey Method on Question Response." Paper presented at the annual meeting of the *American Association for Public Opinion Research*. Delavan, WI.

Groves, Robert M., Paul Biemer, Lars Lyberg, J. Massey, W Nicholls, and J. Wakesberg. 1988. *Telephone Survey Methodology*. New York: Wiley.

Groves, Robert M. 1979. "Actors and Questions in Telephone and Personal Interview Surveys." *The Public Opinion Quarterly* 43(2): 190-205.

Hippler, Hans and Norbert Schwarz. 1998. "Response Effects in Surveys." In Robert M. Groves, Paul Biemer, Lars Lyberg, J. Massey, W. Nicholls, and J. Waksberg (Eds.), *Telephone Survey Methodology*. New York: Wiley. 102-122.

Jordan, Lawrence A., Alfred C. Marcus, & Leo G. Reeder. 1980. "Response Styles in Telephone and Household Interviewing: A Field Experiment." *Public Opinion Quarterly* 44(2): 210-222.

Krosnick, Jon A. and L. R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (Eds.), *Survey Measurement and Process Quality*. New York: Wiley-Interscience.

Krosnick, Jon A. and Matthew K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 37(3): 941-964.

Krysan, Maria and Mick Couper. 2004. "Race of Interviewer Effects: What Happens on the Web?" Paper presented at the annual meetings of the *American Association for Public Opinion Research*. Phoenix, AZ.

Redline, Cleo and Don A. Dillman. 2002. "The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Questionnaires." In R. Groves, D. Dillman, J. Eltinge, and R. Little (Eds.), *Survey Nonresponse*. New York: John Wiley.

Saris, William, Jon A. Krosnick. 2000. "The Damaging Effect of Acquiescence Response Bias on Answers to Agree/Disagree Questions." Paper presented at the annual conference of the *American Association for Public Opinion Research*. Portland, OR.

Schwarz, Norbert, C. E. Grayson, and B. Knäuper. 1998. "Formal Features of Rating Scales and the Interpretation of Question Meaning." *International Journal of Public Opinion Research* 10:177-83.

Schwarz, Norbert and Hans Hippler. 1991. "Response Alternatives: The Impact of Their Choice and Presentation Order." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (Eds.). *Measurement Errors in Surveys.* New York: John Wiley and Sons, Inc.

Schwarz, Norbert, Fritz Strack, Hans-J. Hippler, and George Bishop. 1991a. "The Impact of Administration Mode on Response Effects in Survey Measurement." *Applied Cognitive Psychology* 5:193-212.

Schwarz, Norbert, B. Knäuper, H. J. Hippler, E. Noelle-Neumann, and F. Clark. 1991b. "Rating Scales: Numeric Values may Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55:570-82.

Smith, Tom W. 1995. "Little Things Matter: A Sample of How Differences in Questionnaire Format Can Affect Survey Responses." Paper presented at the *American Association for Public Opinion Research*. Ft. Lauderdale, FL: May.

St-Pierre, Martin and Yves Béland. Forthcoming 2007. "Mode effects in the Canadian Community Health Survey: a comparison of CATI and CAPI." *Monograph for International Telephone Survey Methodology II Conference.*

Tarnai, John and Don A. Dillman. 1992. "Questionnaire Context as a Source of Response Differences in Mail versus Telephone Surveys." In N. Schwarz and S. Sudman (Eds.), *Context Effects in Social and Psychological Research.* New York: Springer-Verlag. 115-129.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68:368-93.

CHAPTER THREE

THE INTERACTIVITY OF COMPUTER TECHNOLOGY: COMPARING ONE-STEP AND TWO-STEP FORMATS FOR RESPONSE SCALES IN A WEB AND TELEPHONE MIXED-MODE SURVEY

Introduction

The invention of computer technology, the Internet, cell phones, and other communication devices has dramatically changed the survey landscape and opened up new possibilities for how surveyors interact with respondents. Web surveys provide the possibility of continuous interaction between the respondent and the survey questionnaire, interaction that is mediated by computer technology. Interviewers interact with respondents in telephone and face-to-face surveys; however, web surveys can incorporate many interactive features into a self-administered survey and often for minimal extra costs. Some of these interactive features have been used in computer assisted telephone and personal interviewing, but new interactive features are also being developed that are unique to web surveys. The flexibility of programming interactive features into web surveys and their increasing use suggests the need to measure the effects of these new and adapted features on responses to web surveys. Since web surveys are often used in mixed-mode survey designs, it is also important to see how responses collected via the web compare to responses from other modes.

Some interactive features that are frequently used in web surveys are progress indicators, next and back navigation buttons, hyperlinks to additional information such as definitions or clarifying instructions, randomization of questions and response options, and numerical calculations (see Conrad et al., 2005 for a discussion of some of these features). One of the most

popular interactive features of web surveys is the ability to branch or skip respondents to subsequent questions based on their responses to a previous question or any combination of previous questions. Computer programming can branch web survey respondents to different follow-up questions, similar to the use of branching in computer-assisted interviewing, and respondents do not have to follow navigation instructions, like in paper surveys, thereby reducing navigation errors.

Rating scales are one of the most frequently used question types in social, political, and psychological research where respondents are asked to rate their attitude, opinion, belief or behavior along a continuum by choosing from an ordered set of response categories. Rating scales are designed in many ways: with varying number of categories, with all or only some of the categories verbally labeled, with or without numerical labels, with the positive end first or last, and with various visual presentations or forms of the scale. The interactivity of web survey design allows for rating scales to be asked in two-steps where respondents are branched to a follow-up question based on their response in the first step. For example, bipolar scales may be asked in two steps where the first step asks the respondent the direction of their attitude or opinion on the issue (e.g. satisfied or dissatisfied, republican or democrat, agree or disagree) and then the second question asks for the strength or intensity of their attitude or opinion (e.g. extremely, strongly, very, somewhat, moderately, slightly, etc.). Branching scalar questions in two-steps is common in telephone surveys to improve data quality by reducing cognitive burden because respondents are asked to process a fewer number of categories at one time.

This paper examines the effects of scale format and survey mode by comparing responses from a one-step scale format and a two-step branched scale format within and across web and telephone modes for ten questions embedded in a mixed-mode survey of college students.

Previous studies comparing the one-step and two-step formats in different survey modes have found that branching improves the reliability and validity of responses (Groves, 1979; Krosnick and Berent, 1993; Malhotra, Krosnick and Thomas, 2007; Miller, 1984). The current study extends previous research by comparing the one and two-step formats within the web mode and by analyzing responses to the one--and two-step scale formats across web and telephone modes. First, the impact of scale format is examined by comparing responses to the one-step format, with all of the categories and only the endpoints verbally labeled, with responses to the two-step format *within* web and telephone modes. Then, the influence of survey mode is analyzed by comparing responses to the one- and two-step formats *across* web and telephone modes. Finally, the results for the multinomial logistic regressions model the effects of scale format, survey mode, and the interaction of format and mode.

Alternative scalar formats for attitude questions: one-step versus two-step

Along with the rise of telephone surveys, surveyors began asking attitude scales in two-steps where respondents are first asked the direction of their attitude and then are branched to the appropriate follow-up question based on whether they expressed a positive or negative attitude in the first step; the follow-up step asks for the strength or intensity of their attitude. Branching rating scales in two-steps decomposes the response task making it easier for respondents to provide an answer when compared to a one-step format where respondents have to consider both the direction and intensity of their attitude at one time (Krosnick and Berent 1993). This two-step branched format for asking rating scales was suggested to help aid respondents to telephone surveys, where a visual display of the response scale is rarely provided, because they are asked to consider a fewer number of categories at one time. This format is rarely used in other survey

modes because it was considered cumbersome in mail surveys where space constraints often limit surveyors from decomposing questions into multiple parts and where interactive branching is not possible. In face-to-face surveys branching seemed unnecessary since interviewers and the use of visual aids (e.g. show cards) help respondents process complex questions with multiple components.

Because of the proliferation of computer technology and the increasing use of web surveys, surveyors have the capability to ask questions in new ways. Web surveys have given expanded attention to the use of branching formats for scalar and other types of questions (and the use of complex skip patterns) because the computer can easily skip respondents to the appropriate follow-up question(s) based on the responses they provide to earlier questions (and any information the surveyor may already have about the respondent). The greater use of mixed mode survey designs that collect data from respondents by telephone and web has also heightened the potential for using a two-step scale format and other interactive branching formats since these design features can be easily integrated into both modes.

In the earliest study of the two-step branched format, Groves (1979) compared responses to a two-step format with responses to a partially labeled one-step scale, where respondents were asked to report a numerical response to a scale, with the endpoints labeled completely satisfied and completely dissatisfied and the midpoint labeled neutral. In the two-step format, respondents were first asked if you are mostly satisfied, mostly dissatisfied, or mixed and then asked how satisfied or dissatisfied you are (delighted/terrible, pleased/unhappy, or mostly satisfied/dissatisfied) unless they chose mixed in the first step. Groves (1979) found more extreme responses to the one-step format and greater use of the midpoint on the two-step format; the two-step format resulted in less extremity bias and higher intercorrelations across items.

Groves also compared these two formats across telephone and face-to-face interviews and found larger mode differences for the two-step version than for the one-step version. Telephone respondents were more likely to choose the most positive category on both the one- and two-step scales than respondents to a personal interview where a show card was provided (Groves, 1979).

Miller (1984) embedded several experiments in a national telephone survey comparing responses from a two-step format with responses from a one-step partially labeled format asking for a numerical response with the endpoints labeled completely satisfied and completely dissatisfied and the midpoint labeled neutral. The two-step format first asked whether they were satisfied, dissatisfied, or somewhere in the middle; those who answered satisfied or dissatisfied were then asked if they were completely, mostly, or somewhat satisfied/dissatisfied and those who answered right in the middle were asked if they were satisfied, dissatisfied or right in the middle. More respondents to the two-step version chose the extreme positive category "completely satisfied" whereas more respondents receiving the one-step version reported being "right in the middle" and "somewhat satisfied" (Miller, 1984). Miller (1984) concludes that attitude scales do not need to be branched or unfolded in two steps if they can be converted to a numerical scale, and he indicates some preference for the one-step version because interviewers preferred this version and it produced less extreme responses, less missing data, and higher intercorrelations across items.

Both of the Groves (1979) and Miller (1984) studies found differences between a partially labeled one-step numerical response scale and a fully labeled two-step scale. But the nature of the differences varied in the two studies where the one-step format produced more extreme responses in the Groves study and the two-step format produced more extreme responses in Miller's study. Miller attributes the differences between his findings and those

reported by Groves (1979) to the different verbal labels between the one- and two-step formats and because respondents choosing the middle category were not asked a follow-up question in Groves's study.

Since the prior two studies (Groves, 1979; Miller, 1984) compared a partially labeled one-step format to a fully labeled two-step format, the differences between formats could also be influenced by the differences in verbal labeling. Research on different types of attitude scales has found that respondents provide more extreme positive responses to fully labeled scales than to polar endpoint labeled scales (Christian, 2003; Christian, Dillman, and Smyth, Forthcoming; Dillman and Christian, 2005). Krosnick and Fabrigar (1997) also argue that fully labeled scales result in more reliable attitude reports than partially labeled scales. In the Groves and Miller studies, we cannot disentangle the effects of branching and verbal labeling on responses.

An important study by Krosnick and Berent (1993) examined both the effects of branching and verbal labeling on political attitudes across telephone, face-to-face, and paper surveys using different respondent populations. Krosnick and Berent (1993) found that branching scales in two steps *and* verbally labeling all of the response categories improve the reliability of respondents' reports of their political attitudes. They argue "our results strongly suggest that verbal labeling and branching should be implemented whenever possible in surveys. Especially because fully labeled branching questions take *less* time to ask than partially labeled nonbranching questions, the former clearly seem preferable to the later" (Krosnick and Berent, 1993, p. 961 *emphasis* in original).

In the most recent study of the two-step branched format and the only study using web surveys, Malhotra, Krosnick and Thomas (2007) examined various ways of branching respondents to attitude questions measuring liking and approval. They found that branching

respondents who select the endpoints significantly increases predictive validity whereas branching respondents who select the midpoint does not improve the validity of responses. They also found that branching the endpoints into three categories produced more valid measures than branching the endpoints into two or four categories (Malhotra, Krosnick and Thomas 2007). Their findings reinforce previous research stating that the optimal length for bipolar constructs is seven categories (Krosnick and Fabrigar 1997).

Only one of these studies investigates the two-step format within web surveys and it does not compare responses to the one- and two-step formats but rather examines different ways of branching respondents within a two-step format. This paper will address this issue by analyzing the one and two-step formats *within* web and telephone modes. The interactivity of web surveys allows for branching two-step scales within a self-administered survey to improve responses without increasing respondent burden. In addition, it seemed important to include telephone comparisons of these formats because of the popularity of the two-step format in telephone surveys and since several of the earlier studies analyzed responses to telephone surveys. To examine the effects of scale format, the results are presented for six comparisons of the fully labeled one-step format, where respondents were presented all seven categories at once, with the two-step format, where respondents were first asked whether they were positive, negative, or neutral (e.g. satisfied, dissatisfied, or neutral) and then asked whether they were very, moderately, or slightly positive or negative (e.g. satisfied or dissatisfied); neutral respondents were not asked a follow up question (see Figure 3.1). Then, results are discussed for four comparisons of the polar endpoint labeled one-step format, where only the endpoint categories were verbally labeled, to the same two-step format (also see Figure 3.1).

Figure 3.1: Summary of experimental comparisons

COMPARING ONE-STEP AND TWO-STEP SCALE FORMATS WITHIN WEB AND TELEPHONE MODES

Fully labeled one-step versus fully labeled two-step	6 comparisons (Table 3.1 web and Table 3.2 phone; Figures 3.2 and 3.3)
Polar point labeled one-step versus fully labeled two-step	4 comparisons (Table 3.3 web and Table 3.4 phone; Figures 3.4 and 3.5)

COMPARING ONE-STEP AND TWO-STEP SCALE FORMATS ACROSS WEB AND TELEPHONE MODES

Fully labeled one-step	6 comparisons (Table 3.5 – top)
Polar point labeled one-step	4 comparisons (Table 3.5 – middle)
Fully labeled two-step	10 comparisons (Table 3.5 – bottom)

MULTINOMIAL LOGISTIC REGRESSION MODELS (Survey Mode, Scale Format, and the interaction of Mode and Format)

Fully labeled one-step versus fully labeled two-step	6 comparisons (Table 3.6 – top; Figure 3.6)
Polar point labeled one-step versus fully labeled two-step	4 comparisons (Table 3.6 – bottom; Figure 3.7)

FULLY LABELED ONE-STEP FORMAT	POLAR POINT LABELED ONE-STEP FORMAT	FULLY LABELED TWO-STEP FORMAT
How satisfied or dissatisfied are you with the housing options in the Pullman and Moscow area? Would you say … ○ Very Satisfied ○ Moderately Satisfied ○ Slightly Satisfied ○ Neutral ○ Slightly Dissatisfied ○ Moderately Dissatisfied ○ Very Dissatisfied	**On a 7-point scale where 7 means very satisfied and 1 means very dissatisfied, how satisfied or dissatisfied are you with the housing options in the Pullman and Moscow area?** ○ Very Satisfied ○ ○ ○ ○ ○ ○ Very Dissatisfied	**Are you satisfied or dissatisfied with the housing options in the Pullman and Moscow Area? Would you say…** ○ Satisfied ○ Dissatisfied ○ Neutral **How (satisfied/dissatisfied) are you with the housing options in the Pullman and Moscow Area? Would you say…** ○ Very ○ Moderately ○ Slightly

Mode differences for alternative scalar formats

Several studies have also found that respondents provide more positive ratings when surveyed by telephone than when surveyed by face-to-face, paper, and web surveys where a visual display of the response scale is provided (Christian, Dillman, Smyth, Forthcoming; de Leeuw, 1992; Dillman and Mason, 1984; Dillman et al., 2001; Groves, 1979; Jordan, Marcus, and Reeder, 1980; Tarnai and Dillman, 1992). These studies have compared four-, five-, seven- and eleven-category partially and fully labeled scales using a variety of scale labels (satisfied/dissatisfied, agree/disagree, etc.). The scales were presented with the positive categories first and last. In one comparison including web responses, Dillman et al. (2001) found that respondents provided more positive ratings when surveyed by telephone and IVR than by web and paper. Similarly, Christian et al. (Forthcoming) found that telephone respondents provided more positive ratings than web respondents in twenty-one of twenty-six comparisons. The literature suggests three possible explanations for more positive ratings on the telephone. First, the increased time pressure and the difficulty of requiring respondents to hold categories in their memory since no visual aid is usually present in telephone interviews. Second, the theory of satisficing suggests that primacy and recency effects occur based on whether the positive or negative end of the scale is presented first. Lastly, respondents may give more attention to internal categories when they are displayed visually.

The only study that has compared the two-step format across modes is the Groves (1979) study discussed earlier where he found that telephone respondents were more likely to select the most positive and midpoint category and less likely to select the sixth category than face-to-face respondents (who received a show card). Additional comparisons of the two-step format across modes is important since branching scales in two-steps could potentially mitigate the mode

effects found for rating scales using fully or partially labeled one-step formats by decreasing the time pressure and by asking respondents to consider a fewer number of categories at once. Also, web surveys are often used in combination with other modes for general population surveys because of Internet coverage issues so the comparisons of the two-step format across web and telephone modes can be helpful in evaluating whether to use the two-step scale format in mixed-mode survey designs.

In addition to presenting the results comparing the fully and partially labeled one-step format with the two-step format within web and telephone modes, results are presented of twenty comparisons *across* web and telephone modes: six comparisons of the fully labeled one-step format, four comparisons of the partially labeled one-step format, and ten comparisons of the branched two-step format (see Figure 3.1). The mixed-mode design of this survey, where implementation procedures were similar for web and telephone respondents, allows for comparisons across the two modes. The wording of the questions was identical in both the web and telephone surveys; the only difference between modes was that the categories were not visually displayed on the telephone survey. To analyze the independent effects of scale format (one-step versus two-step) and survey mode (web and telephone) and the interaction of format and mode, the results of multinomial logistic regression models for each of the ten questions are also discussed (see Figure 3.1).

Methods

The experimental comparisons of the one- and two-step scale formats were embedded in a mixed-mode survey conducted in the Spring of 2006. Undergraduate students were randomly sampled from a list of current students and asked to respond to a short survey (twenty-seven

questions) about their experiences at Washington State University's (WSU) Pullman campus. Students were randomly assigned to one of seven experimental versions (four web and three telephone versions) and the survey was administered to the telephone and web samples simultaneously. For this experimental research, it was necessary to have a population with equal access to completing either a web or telephone version of the survey so respondents could be randomly assigned to one of the seven versions; all students at WSU have computer and web access through the university and students provide postal mail and telephone contact information to the university upon admission. Response rates for both surveys were comparable with 57% of the web respondents completing the survey (1,369 completes out of 2,400 sampled) and 53% of the telephone respondents completing the survey (847 completes of 1,600 sampled) for a combined overall response rate of 55% (2,216 completes of 4,000 sampled). The number of completed surveys for each version ranged from 329 to 353 for the web versions and from 272 to 289 for the telephone versions.

The survey implementation procedures were standardized across the web and telephone modes. All of the students were initially mailed a letter describing the survey and a two-dollar incentive. Students assigned to the web version were instructed in the letter how to access the web survey and students who provided an email address to the university (about 2/3 of the web sample) were also sent an email that included a direct link to the web survey. Subsequent contacts to web nonrespondents were sent using postal mail and email. Students were also provided a unique access code in each contact; each student was asked to entire their unique code at the opening screen to verify their eligibility and begin the survey. Students assigned to the telephone versions were instructed in the letter that they would be contacted by the WSU Social

and Economic Sciences Research Center's telephone lab to complete the telephone survey. Up to 10 callback attempts were made to interview those in the telephone sample.

Every question in the survey appeared on a separate screen where each screen included three regions: a banner or header, the main question area, and a footer. The header included the logo for the university and the title of the survey. The center of the screen is where the question and response options or answer spaces were presented using black font for the text and white for the answer choices/spaces to provide contrast with the colored background; the next and back navigation buttons were also in this region. Lastly, the footer provided contact information for the research organization. All of the screens were constructed using HTML tables where proportional widths were programmed in order to maintain a consistent visual stimulus across differing individual screen or window sizes. Cascading Style Sheets were used to automatically adjust font size and to accommodate varying user browsers and screen resolutions.

The first section presents the results comparing the fully labeled one-step format to the two-step format within web and telephone survey modes for six questions (Q2, 4, 5, 7, 8, and 9) and then the results comparing the polar endpoint labeled one-step scale to the two-step format within each mode for four questions (Q19, 23, 25, and 26). Figure 3.1 displays an overview of the comparisons discussed in this chapter. The means and response distributions are analyzed for each of the ten comparisons because when researchers discuss results for scalar questions, they often report the percentage or frequency of respondents selecting each category (e.g. very satisfied or strongly agree) and/or the average or mean response across all respondents. For each question where responses to the one- and two-step formats are compared, the mean response, difference of mean t-test (one-sided), and Mann-Whitney/Wilcoxon rank-sum test are reported. To examine the response distributions, the percent selecting each category, overall chi-square

tests comparing all seven categories (7*2, df=6), and chi-square tests of individual categories (2*2, df=1) are presented; chi-square tests reported are Fisher exact tests when less than five respondents selected a category in one of the versions being compared. For all of the analyses, the three lowest categories "Very," "Moderately" and "Slightly" negative (categories 1 through 3) are combined because only a few or none of the respondents selected these categories. To summarize across the comparisons of the one- and two-step formats, the average means and response distributions are graphically displayed for the four different versions: web one-step, web two-step, phone one-step, and phone two-step.

In the second section, each of the three scale formats (fully labeled one-step, polar endpoint labeled one-step, and fully labeled two-step) are compared across web and telephone modes. In addition to examining the bivariate effects of survey mode and scale format, a series of multinomial logistic regression models was performed for each of these ten questions to analyze the independent effects of survey mode (telephone or web) and scale format (one-step or two-step) and the effects of the interaction of mode and format. For the multinomial logistic regression models (df=4), the pseudo R^2 and likelihood-ratio chi-square tests are reported for the overall model and for the three independent variables (survey mode, scale format, and the interaction of mode and format). The average predicted probabilities (the expected likelihood of respondents selecting that category) across questions for each category are also graphed for the four versions: web one-step, web two-step, phone one-step, and phone two-step. In all tables, results for statistically significant values (p-values $< .05$) are shown in black and those that are not statistically significant (p-values $\geq .05$) in gray.

Results

Comparing one-step and two-step scale formats within web and telephone modes

Fully labeled one-step versus fully labeled two-step format

Averaging across the six comparisons, web respondents provided more positive responses to the one-step format than the two-step format (see Table 3.1). However, mean responses for individual questions reflect that for the three satisfaction scales respondents provided more positive responses to the one-step format (two of three comparisons significant) whereas for the other three questions (desirability and accessibility scales) respondents provided more positive responses to the two-step format (one of three comparisons significant). The results from the Mann-Whitney tests also show conflicting results where there is a higher rank sum on the one-step format for two of the satisfaction scales (one of two significant) and a higher rank sum on the two-step format for one of the satisfaction and for the desirability and accessibility scales (two of four significant). Similar to web respondents, telephone respondents also provided on average more positive responses to the fully labeled one-step format than the two-step format (see Table 3.2). However, the difference of mean t-tests and Mann-Whitney tests are only significant for one of the six questions (q4). Figure 3.2 displays the average mean response across the six questions for each of the four versions: web fully labeled one-step (5.51), web fully-labeled two-step (5.44), phone fully-labeled one-step (5.81), and phone fully-labeled two-step (5.67).

All six of the overall chi-square tests, comparing all seven categories, for both web and phone respondents are significant indicating that the response distributions for the fully labeled one- and two-step formats are significantly different (see Tables 3.1 and 3.2). Figure 3.3 displays the response distributions for each of the four versions. A greater percentage of the respondents

Table 3.1: Comparing the fully labeled one-step version with the two-step version within the web mode

Q	Scale labels	Scale format	n	Mean	t-test	Mann-Whitney	% selecting each category					Overall chi-square	Chi-square for individual categories				
							7	6	5	4	321		7	6	5	4	321
2	Desirable/ Undesirable	One-step	329	5.95	-1.11	-2.99	31.6	42.9	19.2	3.0	3.3	82.50	11.40	1.37	52.31	18.31	0.02
		Two-step	341	6.05			44.3	38.4	2.1	11.7	3.5						
4	Satisfied/ Dissatisfied	One-step	328	4.84	2.21	1.93	17.4	29.6	17.1	11.3	24.7	87.02	1.24	7.42	37.74	36.87	0.05
		Two-step	342	4.52			20.8	20.5	2.9	30.4	25.4						
7	Satisfied/ Dissatisfied	One-step	329	5.58	0.90	-1.05	18.2	48.6	17.0	7.6	8.5	81.08	15.28	12.65	29.99	21.19	0.26
		Two-step	342	5.48			31.3	35.1	4.1	19.9	9.6						
8	Satisfied/ Dissatisfied	One-step	327	5.62	3.65	2.22	18.4	51.1	16.2	6.7	7.7	90.49	1.89	12.18	38.51	42.34	3.64
		Two-step	340	5.22			22.7	37.7	2.4	25.3	12.1						
5	Accessible/ Inaccessible	One-step	328	5.39	-0.22	-0.70	28.1	33.8	15.6	7.0	15.6	41.00	1.17	0.00	20.54	12.68	0.60
		Two-step	342	5.42			31.9	33.9	5.0	15.8	13.4						
9	Accessible/ Inaccessible	One-step	328	5.65	-3.16	-3.83	24.1	40.2	17.7	14.0	4.0	34.26	11.75	0.11	24.21	0.01	0.55
		Two-step	342	5.93			36.3	41.5	5.6	13.7	3.0						
	Overall	One-step	6	5.51			23.0	41.0	17.1	8.3	10.7						
		Two-step	6	5.44			31.2	34.5	3.7	19.5	11.1						

Notes: For each comparison, I report the question number, scale labels, scale format, and number completing each version; the mean for each version, the difference of means t-test, and nonparametric Mann-Whitney test; the percent selecting each category (where 7 is "Very" positive, 6 is "Moderately" positive, 5 is "Slightly" positive, 4 is "Neutral" nor "Neither" positive or negative, and 321 includes "Slightly," "Moderately" and "Very" negative), the overall chi-square test, and chi-square tests for individual categories. Significant p-values $p<.05$ are shown in black and p-values $\geq .05$ in gray.
The overall mean and percentages combine across all six comparisons for each version.

Table 3.2: Comparing the fully labeled one-step version with the two-step version within the phone mode

Q	Scale labels	Scale format	n	Mean	t-test	Mann-Whitney	% selecting each category					Overall chi-square	Chi-square for individual categories				
							7	6	5	4	321		7	6	5	4	321
2	Desirable/ Undesirable	One-step	286	6.26	0.63	-1.00	38.8	50.7	8.4	1.4	0.7	82.50	11.40	1.37	52.31	18.31	0.02
		Two-step	272	6.21			47.1	39.0	2.9	10.3	0.8						
4	Satisfied/ Dissatisfied	One-step	285	5.34	3.03	2.74	27.7	33.7	14.7	6.0	17.9	87.02	1.24	7.42	37.74	36.87	0.05
		Two-step	271	4.89			26.2	22.5	1.5	34.0	15.9						
7	Satisfied/ Dissatisfied	One-step	286	5.84	1.22	-0.55	28.7	44.4	14.7	7.3	4.9	81.08	15.28	12.65	29.99	21.19	0.26
		Two-step	272	5.71			40.8	28.3	1.1	25.4	4.4						
8	Satisfied/ Dissatisfied	One-step	286	5.70	0.20	-1.07	17.8	50.0	20.3	8.7	3.2	90.49	1.89	12.18	38.51	42.34	3.64
		Two-step	272	5.68			35.3	30.5	2.6	30.9	0.8						
5	Accessible/ Inaccessible	One-step	284	5.73	1.50	0.89	37.3	31.7	14.8	5.3	11.0	41.00	1.17	0.00	20.54	12.68	0.60
		Two-step	270	5.53			35.9	32.6	2.6	18.2	10.7						
9	Accessible/ Inaccessible	One-step	281	5.97	-0.25	-1.03	38.4	34.9	17.1	5.7	4.0	34.26	11.75	0.11	24.21	0.01	0.55
		Two-step	271	6.00			42.4	36.5	6.6	11.4	3.0						
	Overall	One-step	6	5.81			31.5	40.9	15.0	5.7	7.0						
		Two-step	6	5.67			38.0	31.6	2.9	21.7	5.9						

Notes: For each comparison, I report the question number, scale labels, scale format, and number completing each version; the mean for each version, the difference of means t-test, and nonparametric Mann-Whitney test; the percent selecting each category (where 7 is "Very" positive, 6 is "Moderately" positive, 5 is "Slightly" positive, 4 is "Neutral" nor "Neither" positive or negative, and 321 includes "Slightly," "Moderately" and "Very" negative), the overall chi-square test, and chi-square tests for individual categories. Significant p-values $p<.05$ are shown in black and p-values $\geq .05$ in gray.
The overall mean and percentages combine across all six comparisons for each version.

Figure 3.2: Average mean responses for the fully labeled one-step and two-step formats across web and phone (n=6)

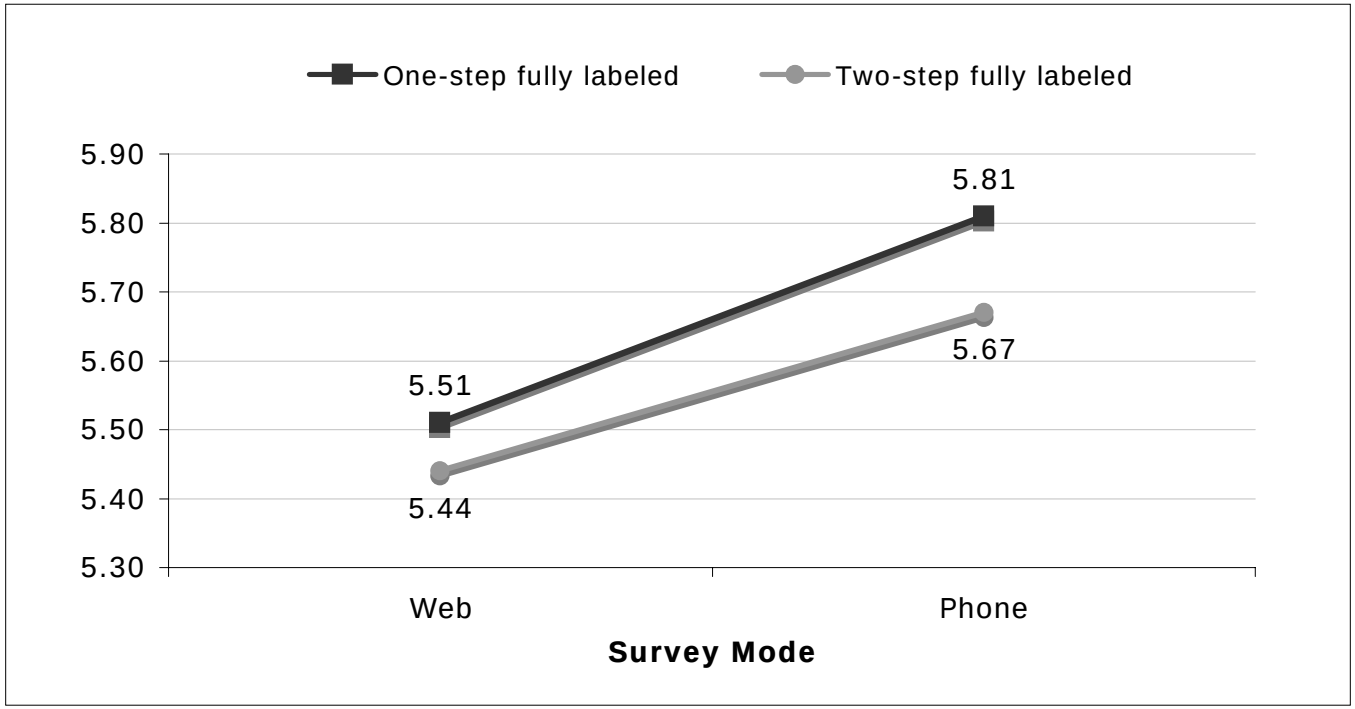
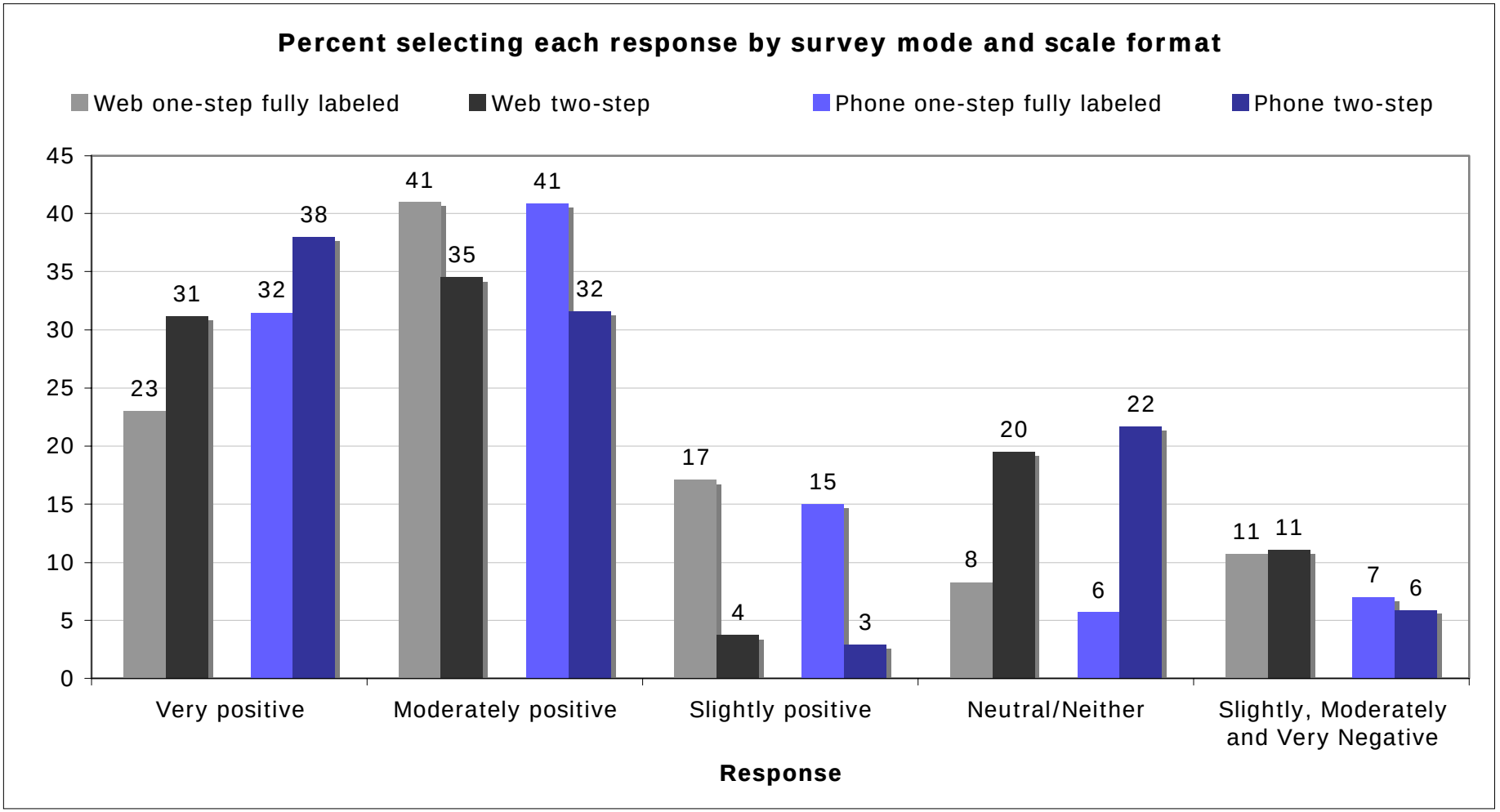


Figure 3.3: Average response distributions for the web and phone fully labeled one-step and two-step formats (n=6 questions)



to the two-step format (compared to the one-step format) chose the most positive category, "Very" positive (*Web* – 8% difference, three of six significant; *Phone* – 6% difference, three of six significant) and the midpoint "Neutral" or "Neither" (*Web* – 12%, five of six tests significant; *Phone* – 16%, six of six tests significant). In contrast, a greater percentage of respondents to the one-step format, compared to the two-step format, chose the sixth category, "Moderately" positive (*Web* – 6% difference overall, three of six chi-square tests significant; *Phone* - 9% difference, four of six significant) and the fifth category, "Slightly" positive (*Web* – 13% difference, six of six significant; *Phone* – 12% difference, six of six significant).

Polar endpoint labeled one-step versus fully labeled two-step format

Averaging across the four comparisons, web respondents provided more positive responses to the fully labeled two-step format than the polar point labeled one-step format; however, only two of the four comparisons are significant (see Table 3.3). The results from the Mann-Whitney tests also show a higher rank sum for respondents to the two-step format than those to the polar one-step format (three of the four significant). Similar to web respondents, phone respondents also provided more positive responses to the two-step format (two of four t-tests significant – see Table 3.4) and higher rank sums for the two-step format (three of four significant) than the polar one-step format. Figure 3.4 displays the average mean response across the four questions for each of the four versions: web polar endpoint labeled one-step (5.19), web fully-labeled two-step (5.33), phone polar one-step (5.44), and phone fully-labeled two-step (5.69).

Table 3.3: Comparing the polar point labeled one-step version with the fully labeled two-step version within the web mode

Q	Scale labels	Scale format	n	Mean	t-test	Mann-Whitney	% selecting each category					Overall chi-square	Chi-square for individual categories				
							7	6	5	4	321		7	6	5	4	321
19	Desirable/ Undesirable	One-step	352	4.50	-2.56	-3.61	8.5	15.9	31.0	20.5	24.1	101.25	21.72	12.49	45.34	0.71	3.04
		Two-step	342	4.83			21.1	26.9	10.2	23.1	18.7						
23	Satisfied/ Dissatisfied	One-step	348	4.83	-0.99	-2.24	10.6	25.6	27.9	18.7	17.2	98.45	9.60	6.96	57.47	2.89	0.17
		Two-step	342	4.95			19.0	34.8	6.1	24.0	16.1						
25	Satisfied/ Dissatisfied	One-step	353	5.65	-2.06	-3.96	20.7	45.6	19.8	8.8	5.1	72.54	26.02	3.80	48.74	6.81	0.01
		Two-step	342	5.85			38.3	38.3	2.9	15.2	5.3						
26	Confident/ Unconfident	One-step	351	5.76	0.72	-1.38	31.6	32.5	23.4	8.3	4.3	57.91	7.75	1.01	44.11	3.46	11.47
		Two-step	342	5.69			41.8	29.0	5.6	12.6	11.2						
	Overall	One-step	4	5.19			17.9	29.9	25.5	14.1	12.7						
		Two-step	4	5.33			30.1	32.3	6.2	18.7	12.9						

Notes: For each comparison, I report the question number, scale labels, scale format, and number completing each version; the mean for each version, the difference of means t-test, and nonparametric Mann-Whitney test; the percent selecting each category (where 7 is "Very" positive, 6 is "Moderately" positive, 5 is "Slightly" positive, 4 is "Neutral" nor "Neither" positive or negative, and 321 includes "Slightly," "Moderately" and "Very" negative), the overall chi-square test, and chi-square tests for individual categories. Significant p-values p<.05 are shown in black and p-values ≥ .05 in gray.
The overall mean and percentages combine across all four comparisons for each version.

Table 3.4: Comparing the polar point labeled one-step version with the two-step version within the phone mode

Q	Scale labels	Scale format	n	Mean	t-test	Mann-Whitney	% selecting each category					Overall chi-square	Chi-square for individual categories				
							7	6	5	4	321		7	6	5	4	321
19	Desirable/ Undesirable	One-step	287	4.69	-1.16	-1.79	9.8	17.1	34.2	22.0	17.0	88.40	6.31	16.34	72.97	4.59	0.07
		Two-step	271	4.85			17.0	31.7	5.2	29.9	16.3						
23	Satisfied/ Dissatisfied	One-step	286	5.29	-1.86	-2.67	18.2	29.0	29.4	14.7	8.8	93.50	17.29	0.05	70.43	14.15	1.64
		Two-step	271	5.51			33.6	29.9	3.0	27.7	5.9						
25	Satisfied/ Dissatisfied	One-step	288	5.84	-5.77	-7.60	24.0	45.5	23.6	4.9	2.1	114.90	67.60	16.14	58.50	5.29	0.85
		Two-step	272	6.32			58.1	29.0	1.8	9.9	1.1						
26	Confident/ Unconfident	One-step	286	5.93	-1.40	-2.99	36.4	33.9	19.9	7.0	2.9	77.52	13.55	2.04	51.49	11.58	0.03
		Two-step	272	6.07			51.8	28.3	1.1	16.2	2.5						
	Overall	One-step	4	5.44			22.1	31.4	26.8	12.2	7.8						
		Two-step	4	5.69			40.1	29.7	2.8	20.9	6.4						

Notes: For each comparison, I report the question number, scale labels, scale format, and number completing each version; the mean for each version, the difference of means t-test, and nonparametric Mann-Whitney test; the percent selecting each category (where 7 is "Very" positive, 6 is "Moderately" positive, 5 is "Slightly" positive, 4 is "Neutral" nor "Neither" positive or negative, and 321 includes "Slightly," "Moderately" and "Very" negative), the overall chi-square test, and chi-square tests for individual categories. Significant p-values p<.05 are shown in black and p-values ≥ .05 in gray.
The overall mean and percentages combine across all four comparisons for each version.

Figure 3.4: Average mean responses across questions for the polar point labeled one-step and two-step formats across web and phone (n=4)

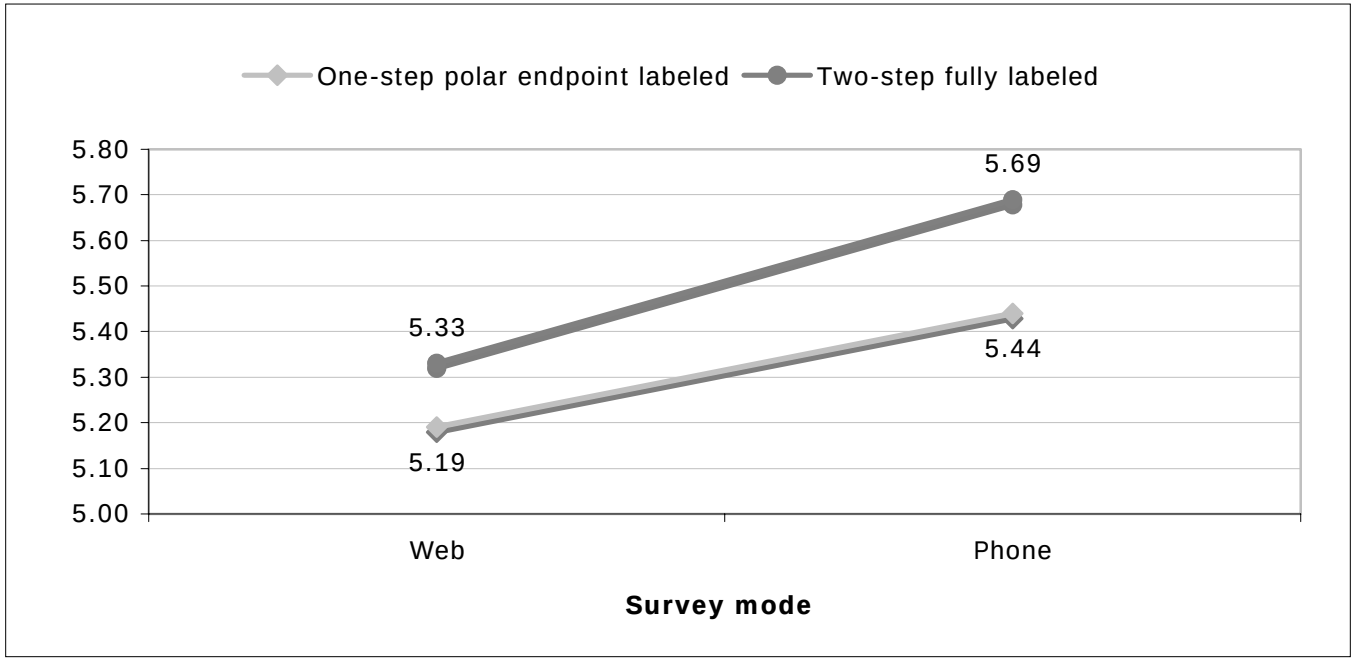
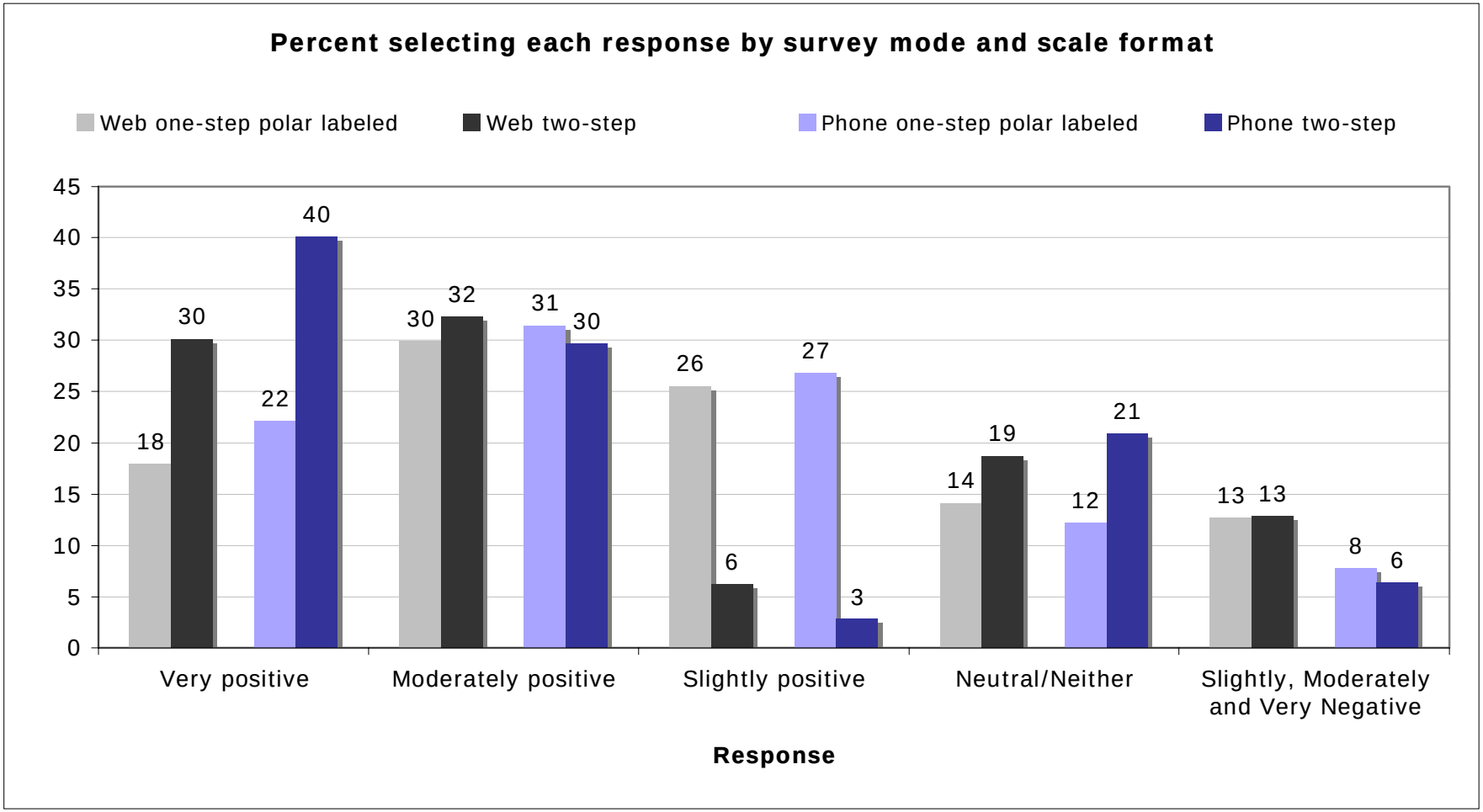


Figure 3.5: Average response distributions for the web and phone polar point labeled one-step and two-step formats (n=4 questions)



All four of the overall chi-square tests for both web and phone respondents are significant, indicating that the response distributions for the polar endpoint labeled one-step and fully labeled two-step formats are significantly different (see Tables 3.3 and 3.4). Figure 3.5 displays the response distributions for each of the four versions. A greater percentage of respondents to the two-step format chose the seventh category, "Very" positive (*Web* – 12% difference, four of four significant; *Phone* – 18% difference, four of four significant) and the fourth category, the "Neutral" or "Neither" midpoint (*Web* – 5% difference, one of four significant; *Phone* – 9% difference, four of four significant). In contrast, a greater percentage of respondents to the polar one-step format (compared to the two-step format) chose the fifth category "Slightly" positive (*Web* – 20% difference, four of four significant; *Phone* – 24% difference, four of four significant) and a slightly greater percentage chose the sixth category "Moderately" positive (*Web* – 2% difference, two of four significant; *Phone* – 1% difference, two of four significant).

Larger differences in the means between the one-step and two-step scale formats were found for the fully labeled one-step format than for the polar point labeled one-step format. When compared to the two-step format, the differences in the percentage of respondents choosing the seventh (most positive) and sixth categories were larger for respondents to the polar endpoint labeled one-step whereas the differences in the percentage of respondents choosing the fifth and fourth (midpoint) categories were larger for the respondents to the fully labeled one-step format. Across both comparisons of the one-step format (fully labeled and polar endpoint labeled) with the two-step format, more respondents to the two-step format selected the most positive category whereas more respondents to the one-step format selected the fifth category.

The finding that more respondents to the two-step format selected the most positive category support Miller's (1984) finding because he also reported that the two-step format produced more positive response whereas Groves (1979) reported higher use of the extreme positive category on the one-step format. These differences may be due to the variations in verbal labeling for the most positive category; in the Groves study the label "delighted" in the two-step format may be considered more extreme than "completely satisfied" in the one-step format and than "very satisfied" in the one- and two-step formats in Miller's study. In contrast, the results of this study support Groves's finding that more respondents to the two-step format selected the midpoint category on the two-step than the one-step format (Miller found that the one-step format produced higher use of the midpoint category). This difference is likely due to the fact that respondents who reported being "neutral" in the first step were not asked a follow-up question in this study or the Groves study whereas they were asked a follow up question in Miller's study.

Comparing one-step and two-step scale formats across web and telephone modes

Telephone respondents provide higher mean ratings than web respondents in all twenty across-mode comparisons of fully labeled one-step, polar point labeled one-step, and fully labeled two-step formats (see Table 3.5). Fifteen of the twenty t-tests comparing the difference between the two means are statistically significant (five of six fully labeled one-step, three of four polar one-step, and seven of ten fully labeled two-step). Phone respondents also showed higher rank sums than web respondents in twelve of twenty comparisons (five of six fully labeled one-step, one of four polar one-step, and six of ten fully labeled two-step). Chi-square tests comparing all seven categories indicated statistically significant differences in response

Table 3.5: Comparing responses by web and phone for the one and two-step scale formats

	Q	Scale labels	Web mean	Phone Mean	Difference of means	t-test	Mann Whitney	Overall chi-square	Chi-square category 7
Fully labeled one-step	2	Desirable/Undesirable	5.95	6.26	-.31	-4.21	-1.03	23.90	3.49
	4	Satisfied/Dissatisfied	4.84	5.34	-.50	-3.56	-3.86	17.38	9.45
	7	Satisfied/Dissatisfied	5.58	5.84	-.26	-2.79	-2.99	12.20	9.38
	8	Satisfied/Dissatisfied	5.62	5.70	-.08	-0.85	0.04	8.66	0.03
	5	Accessible/Inaccessible	5.39	5.73	-.34	-2.62	-2.65	8.81	5.98
	9	Accessible/Inaccessible	5.65	5.97	-.32	-3.52	-3.82	23.76	14.64
		Overall (n=6)	5.51	5.81	-.30				
Polar one-step	19	Desirable/Undesirable	4.50	4.69	-.19	-1.62	-1.60	5.34	0.29
	23	Satisfied/Dissatisfied	4.83	5.29	-.46	-4.19	-3.92	18.40	7.41
	25	Satisfied/Dissatisfied	5.65	5.84	-.19	-2.29	-1.55	11.65	0.99
	26	Confident/Unconfident	5.76	5.93	-.17	-1.88	-1.72	8.20	1.58
		Overall (n=4)	5.19	5.44	-.25				
Two-step branched version	2	Desirable/Undesirable	6.05	6.21	-.16	-1.76	-1.08	6.77	0.47
	19	Desirable/Undesirable	4.83	4.85	-.02	-0.10	0.36	12.73	1.62
	4	Satisfied/Dissatisfied	4.52	4.89	-.36	-2.37	-2.36	12.61	2.51
	7	Satisfied/Dissatisfied	5.48	5.71	-.23	-1.79	-1.98	19.16	6.00
	8	Satisfied/Dissatisfied	5.22	5.68	-.46	-3.82	-3.48	40.33	11.92
	23	Satisfied/Dissatisfied	4.95	5.51	-.56	-4.13	-3.97	32.78	16.93
	25	Satisfied/Dissatisfied	5.85	6.32	-.47	-4.78	-5.08	29.38	23.80
	5	Accessible/Inaccessible	5.42	5.53	-.11	-0.77	-0.97	6.48	1.11
	9	Accessible/Inaccessible	5.93	6.00	-.07	-0.68	-1.30	12.37	2.43
	26	Confident/Unconfident	5.69	6.07	-.38	-3.24	-2.99	29.99	6.13
		Overall (n=10)	5.39	5.68	-.29				

Notes: For each comparison, I report the question number, scale labels, the mean for web and phone respondents, the difference between the means and the t-test, the overall chi-square test, and the chi-square tests for 7 the most positive category. Significant p-values $p<.05$ are shown in black and p-values $\geq .05$ in gray. The overall mean and percentages combine across all comparisons for each scale format.

distributions for twelve of twenty comparisons (four of six fully labeled one-step, one of four polar one-step, and seven of ten fully labeled two-step). A greater percentage of phone respondents than web respondents chose the most positive category (3.2 to 10% overall difference depending on the scale format); ten of twenty individual chi-square tests for the most positive category are significant (four of six fully labeled one-step, one of four polar one-step, and five of ten fully labeled two-step). Overall, the results show that telephone respondents provide higher mean ratings and are more likely to select the most positive category than web respondents across a variety of scale formats, including the two-step format. These findings support previous studies where telephone respondents provide more extreme positive responses than those surveyed by personal interview (and provided show cards) and those surveyed by mail or web.

Modeling the independent effects of scale format and survey mode and their interaction

Fully labeled one-step versus two-step

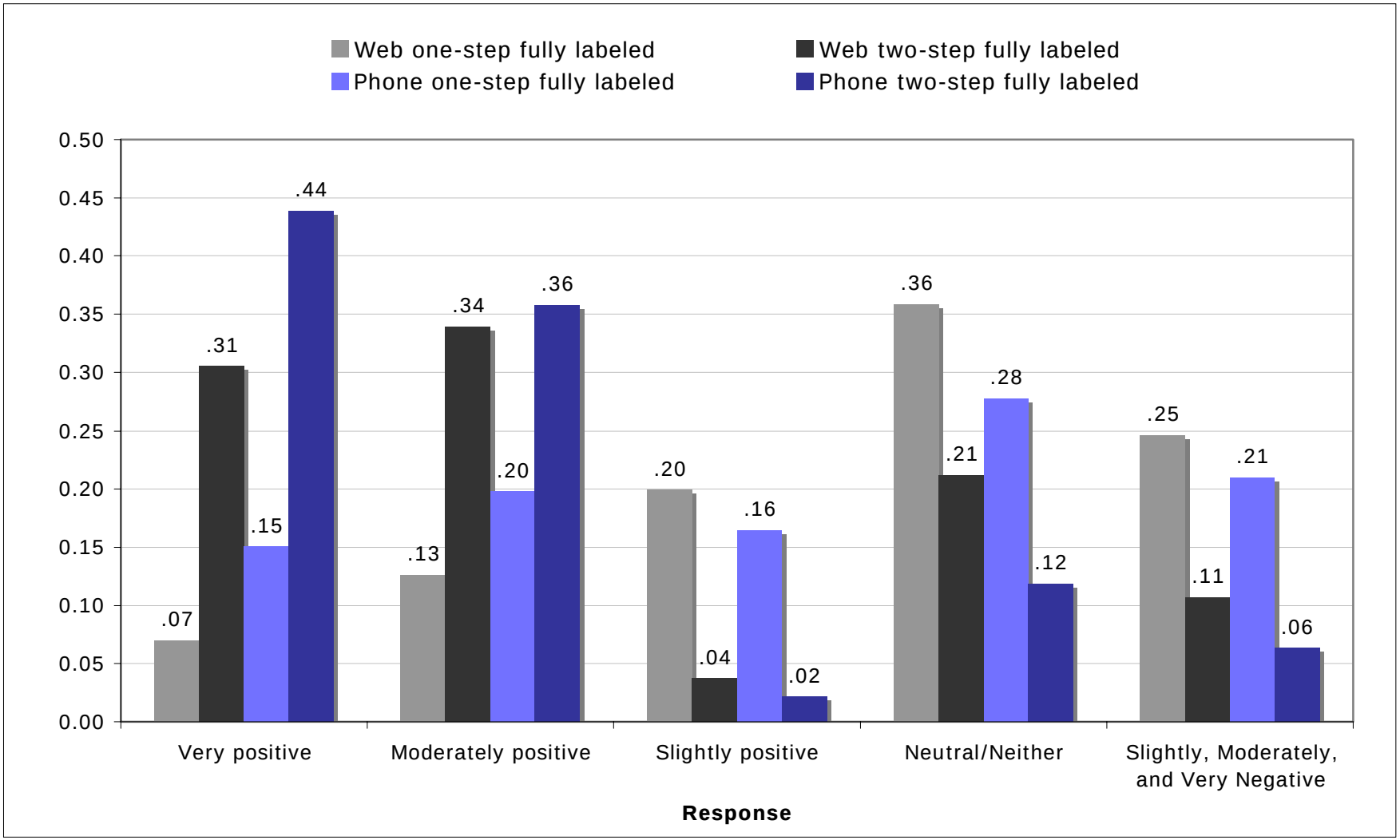
The overall likelihood ratio chi-square for the multinomial logistic regression models is shown for all six comparisons of the fully labeled one-step and two-step formats in Table 3.6; all six tests indicate that the overall model (including the survey mode and scale format independent variables and the interaction variable) is significant. In addition, the likelihood ratio chi-square tests for the mode and format variables and the interaction variable are significant in all six comparisons indicating that the format of the scale and the mode of administration are independently and interacting to influence the answers respondents provide (also shown in Table 3.6). Figure 3.6 displays the probability of choosing each category for the web and phone fully labeled one-step and two-step formats. The probability of choosing the most

Table 3.6: Multinomial logistic regression results

	Q	Scale labels	n	Pseudo R^2	Overall model	Independent variables		
						Survey Mode	Scale Format	Mode*Format
Fully labeled one-step versus fully labeled two-step	2	Desirable/ Undesirable	2214	.093	652.92	81.38	166.94	16.40
	4	Satisfied/ Dissatisfied	2213	.046	307.32	54.81	95.22	14.10
	7	Satisfied/ Dissatisfied	2215	.066	456.94	49.03	140.95	21.63
	8	Satisfied/ Dissatisfied	2210	.062	418.02	27.26	180.21	39.86
	5	Accessible/ Inaccessible	2208	.058	412.04	43.03	111.76	11.36
	9	Accessible/ Inaccessible	2200	.084	581.09	61.04	143.01	12.89
Polar point labeled one-step versus fully labeled two-step	19	Desirable/ Undesirable	2210	.013	91.43	13.47	46.30	10.47
	23	Satisfied/ Dissatisfied	2204	.023	156.72	36.14	56.43	5.45
	25	Satisfied/ Dissatisfied	2211	.026	153.03	19.30	67.49	6.42
	26	Confident/ Unconfident	2206	.018	112.11	11.84	49.69	12.86

Notes: For the Multinomial logistic regression models the df=4 and the pseudo R^2 is reported; likelihood-ratio chi-square tests are reported for the overall model, the survey mode and scale format variables, and for the interaction variable (mode*format) and significant p-values $p<.05$ are shown in black (and p-values $\geq .05$ in gray,

Figure 3.6: Average predicted probability for each response for the web and phone fully labeled one-step and two-step formats (n=6 questions)



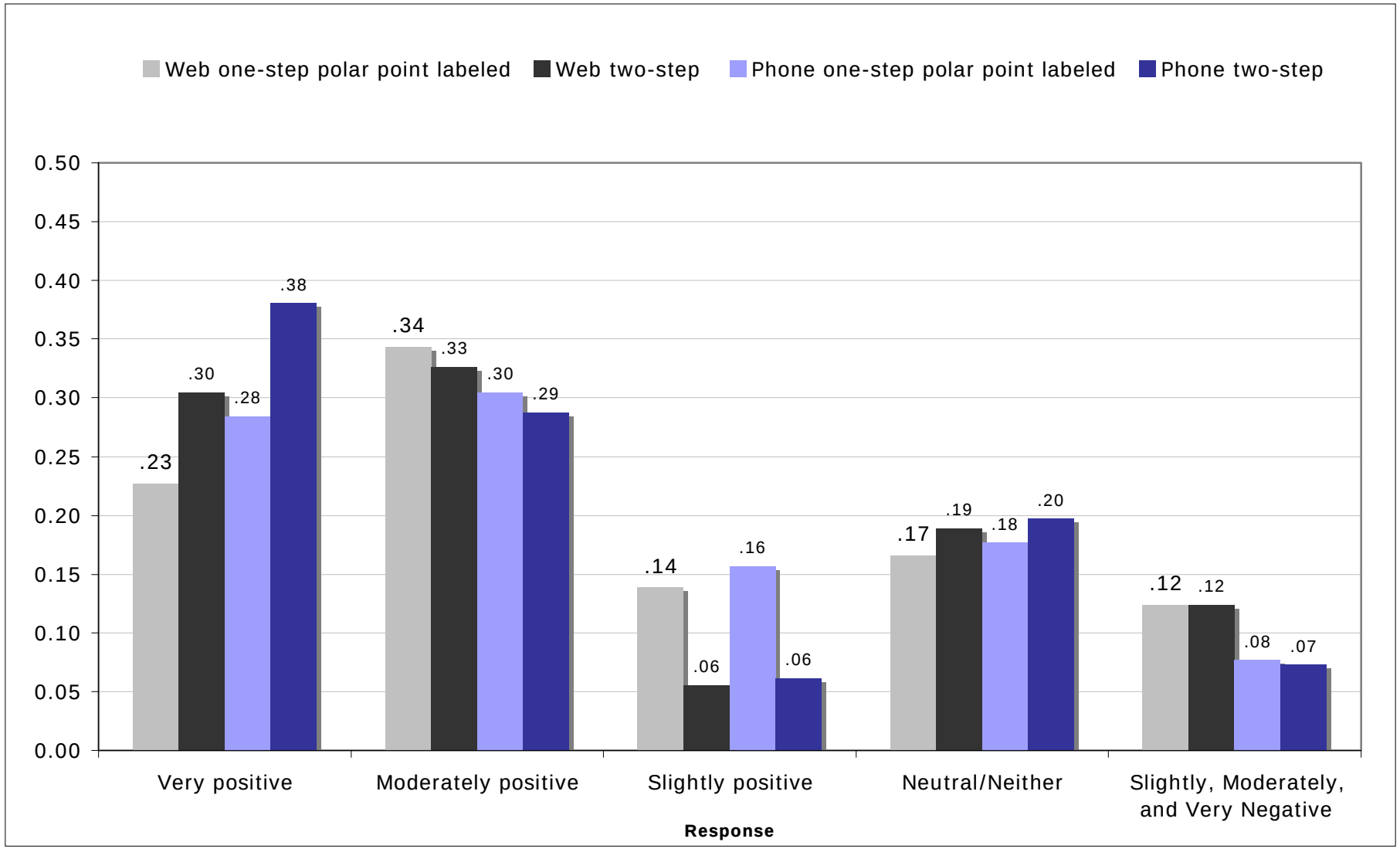
positive category (seven "Very" positive) is highest for respondents to the two-step format (Web .31, Phone .44) and lowest for respondents to the one-step format (Web .07, Phone .15). The probability of choosing the sixth category ("Moderately" positive) is also highest for respondents to the two-step format (Web .34, Phone .36) and lowest for respondents to the one-step format (Web .13, Phone .20). In contrast, the likelihood of choosing the fifth category ("Slightly" positive) is highest for respondents to the one-step format (Web .20, Phone .16) and lowest for respondents to the two-step format (Web .04, Phone .02). Similarly, the probability of choosing the midpoint category (four "Neutral or Neither") is highest for respondents to the one-step format (Web .36, Phone .28) and lowest for respondents to the two-step format (Web .21, Phone .12). Lastly, the probability of choosing one of the three negative categories (321) is also highest for respondents to the one-step format (Web .25, Phone .21) and lowest for respondents to the two-step format (Web .11, Phone .06). Figure 3.6 shows that there are large differences in the probabilities of selecting all response categories by scale format, consistent differences by survey mode with more positive responses on the phone, and a clear interaction of format and mode (with telephone respondents to the two-step format providing more positive responses and the web respondents to the one-step format providing more negative responses).

Polar endpoint labeled one-step versus fully labeled two-step (four comparisons)

The overall likelihood ratio chi-square for the multinomial logistic regression models is significant for all four comparisons of the polar endpoint labeled one-step and fully labeled two-step formats indicating that mode, format, and their interaction are influencing responses (see bottom of Table 3.6). In addition, the likelihood ratio chi-square tests from the logistic regression models are significant for the independent effects of mode and format in all four comparisons

and for the interaction of mode and format in two of the four comparisons (also in Table 3.6). Figure 3.7 displays the probability of choosing each category for the web and phone polar endpoint labeled one-step and fully labeled two-step formats. The probability of choosing the most positive category, "Very" positive, is highest for respondents to the two-step format (Web .30, Phone .38) and lowest for respondents to the one-step format (Web .23, Phone .28). In contrast, the probability of choosing the fifth category, "Slightly" positive, is highest for respondents to the one-step format (Web .14, Phone .16) and lowest for respondents to the two-step format (Web and Phone .06). The probability of choosing the sixth category, "Moderately" positive is slightly higher for web than phone respondents (Web .33–.34, Phone .29–.30) and the probability of choosing the three negative categories (321) is also higher for web respondents than phone respondents (Web .12, Phone .07–.08). Lastly, the probability of choosing the midpoint varies only slightly across the four versions (.17–.20). Looking at both the fully labeled and polar endpoint labeled one-step formats, respondents are more likely to select the fifth category on both one-step formats and to select the most positive category on the two-step format. However, the differences in probabilities of choosing these categories for the one- and two-step formats are much larger for the fully labeled than the polar endpoint labeled one-step format (comparing Figure 3.6 and Figure 3.7); the differences in probabilities are also larger for all the other categories on the fully labeled format. Overall, greater mode differences were found for the fully labeled one-step and two-step comparisons than for the polar endpoint labeled one-step and two-step comparisons.

Figure 3.7: Average predicted probability for each response for the web and phone polar point labeled one-step and two-step formats (n=4 questions)



Discussion

Overall, both scale format and survey mode independently influence responses to scalar questions. Using a one- versus two-step scale format and surveying respondents by web or phone both had independent effects on respondents' answers to ten rating scale questions and the multivariate models show larger effects for scale format than for survey mode. Respondents to the two-step format were more likely to select the most positive category whereas respondents to the one-step format were more likely to select the fifth category. Telephone respondents to the one- and two-step scale formats provided higher mean ratings than web respondents and were more likely to select the most positive category.

Scale format and survey mode also interacted to influence responses to scalar questions, especially when comparing the fully labeled one-step format with the two-step format. Telephone respondents to the two-step format were more likely to select the seventh category whereas web respondents to the one-step format were more likely to select the fifth category. There were also larger scale format and survey mode differences between the one- and two-step formats when all of the categories were verbally labeled in the one-step format compared to when only the endpoints were verbally labeled.

As web surveyors introduce more interactive features into their designs, they should consider how computer programming can help reduce respondent burden by having respondents answer only questions that apply to them. The respondent experience can be improved by breaking up complex questions into multiple steps and asking respondents an initial question so the computer can skip them to the appropriate follow-up questions. The two-step interactive scale format provides another choice for researchers when considering how to design rating scale questions to best measure their construct of interest. The two-step format is conceptually easier

for respondents because it decomposes the response task into two steps with fewer categories. Responses over time have been shown to be more reliable for the two-step branched format than for the one-step format (Krosnick and Berent, 1993) and more valid when the endpoints are branched into three categories (Malhotra, Krosnick and Thomas, 2007). In cognitive interviews testing the web survey, some respondents had difficulty with the two-step format not realizing a different question was being asked in the second-step; their experiences illustrated how it is important to introduce the branched format and/or to clearly distinguish the two steps such that respondents realize a new follow-up question is being asked. Debriefing with the telephone interviewers for this project revealed that they preferred the two-step format to the fully labeled one-step format.

These findings raise particular concern for combining web and telephone responses to one- and two-step rating scales. The evidence is now growing that telephone respondents provide more positive responses to a variety of different types of rating scales than respondents to other modes where the scale is conveyed visually (face-to-face with a show card, paper, and web). Surveyors need to continue to experiment with new rating scale formats to attempt to minimize the differences in responses to scalar questions between telephone respondents and respondents to other modes. However, this may be a challenge as the research shows that differences in responses between telephone and other modes persist across a variety of scale lengths (4, 5, 7, and 11 category scales), a variety of scale labels, with all or only some of the categories labeled, with the positive or negative end presented first, and when the scale is presented in one or two steps. When considering mixed-mode designs for data collection that include the telephone mode, it is important to evaluate the effects of mode on responses to rating scales before combining responses gathered by telephone with those from other modes.

References

Christian, Leah Melani. 2003. "The Influence of Visual Layout on Scalar Questions in Web Surveys." Master's Thesis: Washington State University.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. Forthcoming. "The Effects of Mode and Format On Answers to Scalar Questions in Telephone and Web Surveys." In Lepkowski, J., Tucker, C., Brick, M., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., Sangster R. (Eds.), *Advances in Telephone Survey Methodology.* New York: Wiley-Interscience.

Conrad, Fred, Mick Couper, Roger Tourangeau, and Mirta Galesic. 2005. "Interactive Feedback Can Improve the Quality of Responses in Web Surveys." Paper presented at the annual meetings of the *American Association for Public Opinion Research*. Miami Beach, FL.

de Leeuw, Edith D. 1992. *Data Quality in Mail, Telephone, and Face to Face Surveys*. Amsterdam: TT Publications.

Dillman, Don A. and Leah Melani Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17(1): 30-52.

Dillman, Don A. and Robert G. Mason. 1984. "The Influence of Survey Method on Question Response." Paper presented at the annual meeting of the *American Association for Public Opinion Research*. Delavan, WI.

Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell and Jodi Berck. 2001. "Response Rate and Measurement Differences in Mixed Mode Surveys: Using Mail, Telephone, Interactive Voice Response and the Internet." Draft Paper?

Groves, Robert M. 1979. "Actors and Questions in Telephone and Personal Interview Surveys." *The Public Opinion Quarterly* 43(2): 190-205.

Jordan, Lawrence A., Alfred C. Marcus, & Leo G. Reeder. 1980. "Response Styles in Telephone and Household Interviewing: A Field Experiment." *Public Opinion Quarterly* 44(2): 210-222.

Krosnick, Jon A. and Matthew K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 37(3): 941-964.

Krosnick, Jon A. and L. R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (Eds.), *Survey Measurement and Process Quality*. New York: Wiley-Interscience.

Malhotra, Neil, Jon A. Krosnick and Randall Thomas. 2007. "Optimal Design of Branching Questions to Measure Bipolar Constructs." Paper presented at the 62nd Annual Meeting of the *American Association for Public Opinion Research*. Anaheim, CA.

Miller, Peter V. 1984. "Alternative Question Forms for Attitude Scale Questions in Telephone Interviews." *Public Opinion Quarterly* 48:766-778.

Tarnai, John and Don A. Dillman. 1992. "Questionnaire Context as a Source of Response Differences in Mail versus Telephone Surveys." In N. Schwarz and S. Sudman (Eds.), *Context Effects in Social and Psychological Research.* New York: Springer-Verlag. 115-129.

CHAPTER FOUR

THE INFLUENCE OF VISUAL DESIGN ON THE SURVEY RESPONSE PROCESS

Introduction

Surveyors are increasingly utilizing mixed-mode survey designs, instead of conducting single-mode surveys, to decrease errors that arise from coverage, sampling, and nonresponse bias. As survey researchers balance the four sources of error, improvements in coverage, sampling, and nonresponse must be understood with attention to potential measurement error that may arise when responses to different survey modes are combined for analysis. Particular features of survey modes and variations in questionnaire design can help surveyors understand why one mode may measure responses differently than another mode.

Information can be transmitted between the surveyor and the respondent using visual communication, aural communication, or a combination of aural and visual communication. The form of communication influences the process respondents perform when responding to surveys. Before they can complete the cognitive steps to respond to individual questions, respondents must first visually and/or aurally perceive, attend to, and process the information presented to them as part of the survey conversation. How respondents first perceive and process survey information influences their comprehension of and ultimately their responses to the questions. Since visual and aural information are perceived and processed separately, the form of communication can help to explain why we see variations in responses collected using different modes.

Web surveys have the flexibility of communicating information to respondents visually through the questionnaire screens, and aurally through recorded audio. Although very little

research has examined the potential of utilizing recorded audio in web surveys, the proliferation of web surveys has brought increased attention to how visual design can influence respondent behavior, increase respondents' motivation to respond, and improve their overall experience with the survey questionnaire. Surveyors can strategically manipulate visual design elements and their properties to guide respondents in completing web surveys.

First, this paper discusses the forms of communication available in each survey mode and briefly explains the differences in how survey respondents perceive and process aural and visual information. A three-stage model is presented to describe the process by which survey respondents perceive visual elements and their properties on the page, organize the information presented to them in the questionnaire, and then complete the task of responding to the survey. Finally, examples are shown to illustrate that web surveyors can strategically use visual design to help respondents understand the basic organization of information on the screen, navigate the web questionnaire, and respond to individual questions.

How differences in communication influence the response process

A survey is a special type of conversation where the surveyor guides the communication and the type of information elicited from respondents (Schwarz, 1996). This information exchange is guided by conversational norms where contributions from the surveyor and respondent should be truthful, clear or comprehensible, relevant to the conversation, and informative. The survey mode influences the conversation between survey researchers and respondents. In face-to-face and telephone surveys, an interviewer exchanges information with the respondent on behalf of the surveyor; however, in paper and web surveys, the questionnaire represents the researcher's contributions to the survey conversation. In addition to the presence

of an interviewer to communicate with respondents, communication can also be mediated by computer technology. Researchers may have interviewers use a computer to help improve interaction with respondents (e.g. computer-assisted personal and telephone interviewing) or the respondent can interact directly with a computer in a self-administered setting (e.g. web surveys and computer-assisted self-interviewing usually during a face-to-face interview). Thus, the survey mode influences the survey conversation because an interviewer and/or computer may mediate communication between the surveyor and the respondent in some modes but not in others.

The survey mode also influences the form of communication during the conversation. Some survey modes are limited to using only aural communication or only visual communication to transmit information, whereas other survey modes can utilize both aural and visual communication (see Figure 4.1). Web surveys can communicate information visually to respondents through the questionnaire screens and aurally by audio recordings. Face-to-face surveys also utilize both aural and visual communication where information is conveyed through the interviewer's voice, body language, facial expressions, and eye contact. The interviewer can also present show cards to visually display information or can ask respondents to complete questions on paper or a computer. Mail and telephone surveys are more limited because they primarily rely on only one form of communication. In mail surveys, communication is limited to the information presented visually on the questionnaire pages. In contrast, communication with respondents in telephone surveys is usually restricted to information conveyed aurally via the interviewer's voice. Therefore, features of the survey mode, such as the presence of an interviewer and the form(s) of communication, influence the survey conversation and how respondents complete the steps in the response process.

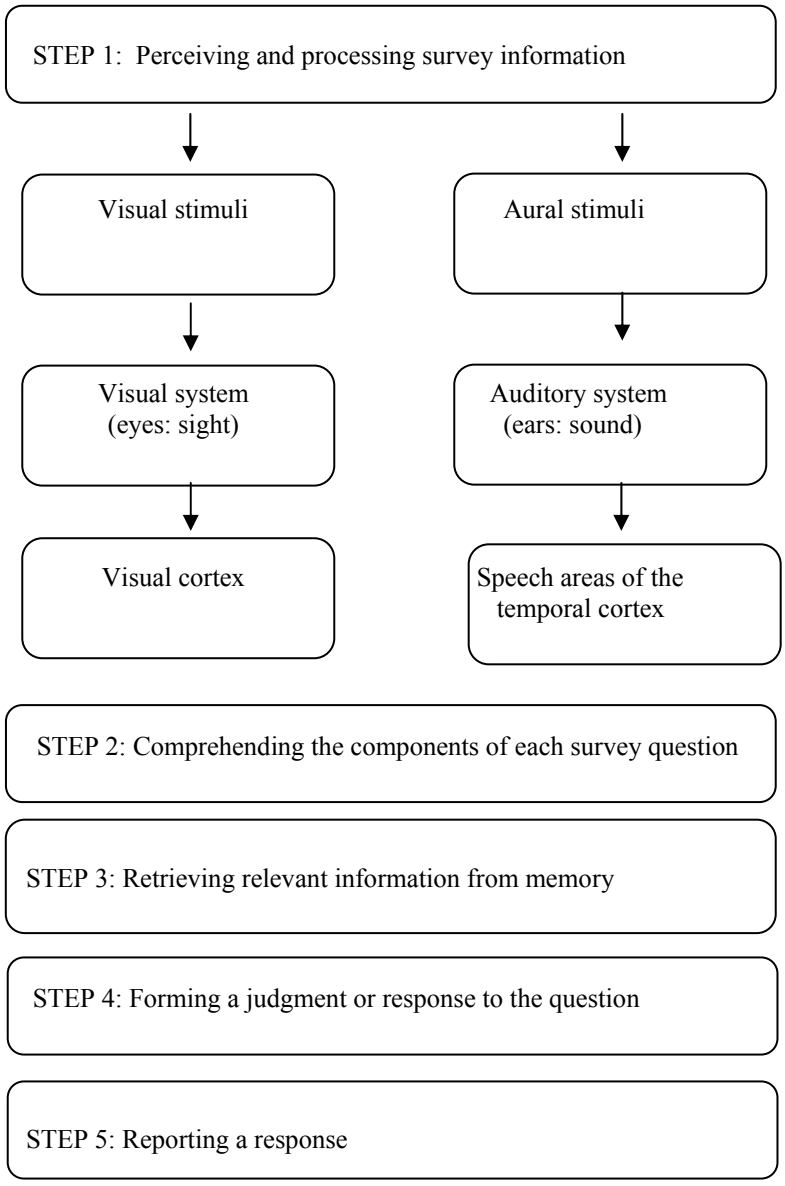
Figure 4.1: Forms of communication in each survey mode

SURVEY MODE	Visual communication	Aural communication
Web	Questionnaire screens	Recorded audio
Mail	Questionnaire pages	---
Telephone	---	Interviewer's voice
Face-to-face	Show cards Interviewer's body and face	Interviewer's voice

Many models of the survey response process include four steps: (1) comprehending the question, (2) retrieving relevant information from memory, (3) forming a judgment or response to the question, and (4) reporting a response (Tourangeau, Rips, and Rasinski, 2000). Before respondents can comprehend individual survey questions, they must first perceive, attend to, and process the visual and/or aural information presented to them during the survey conversation (see Figure 4.2). Information that is not perceived or processed cannot be used during the other steps in the response process. Whether information is presented to respondents visually or aurally during the survey is important because they are perceived using different sensory systems, the auditory versus the visual system. In addition, Ware (2004) discusses how people process aural and visual information using different neural centers in the brain, the speech areas of the temporal cortex versus the visual cortex. Understanding the influence of the form of communication on how respondents perceive and process survey information is crucial for

survey researchers because this initial step in the survey response process influences the manner in which people will comprehend and ultimately respond to survey questions (see Figure 4.2).

Figure 4.2: The five-step survey response process



Words are the building blocks of communication and are the most powerful source of meaning in both aural and visual communication. Spoken words and paralinguistic cues conveyed through the interviewer's voice (such as inflection, tone, and timing) are the primary channels through which information is conveyed aurally to respondents. Results from two surveys of Washington State University (WSU) students, the first conducted in the Fall of 2004 and the second in the Spring of 2006, demonstrate the importance of words in conveying meaning to respondents in telephone surveys. Changing the words in the question stem on the first survey significantly increased the number of respondents reporting both the month and the year.[4] More respondents reported the month and year when specifically asked "What month and year did you begin your studies at WSU" (84%) than when asked "What date" (50%; χ^2=84.6, p=.000) or "When" (13%; χ^2=316.9, p=.000). In the second survey, asking respondents to "Please provide your answer using two digits for the month and four digits for the year" significantly increased the percentage of respondents reporting their response in the desired format from less than 1% to 59% (χ^2=236.2, p=.000).[5] Significantly more respondents reported the month in two digits when provided the extra instruction (82% vs. less than 1%; χ^2=392.0, p=.000), but an equal percentage of respondents reported the year using four digits regardless of whether provided the instruction (72%; χ^2=0.0, p=.989). These two experiments demonstrate how respondents will make sense of the information presented to them aurally and how small changes in question wording can dramatically influence respondent behavior. After briefly exploring differences between aural and visual communication, the remainder of this paper will

[4] An expanded discussion of results from the first telephone survey are presented in: Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. 2005. "How to Instruct Web and Telephone Respondents to Report Date Answers in a Format Desired by the Surveyor." Social and Economic Sciences Research Center Technical Report #05-067.

[5] These data from the Student Experience Survey are unpublished but are available from WSU's Social and Economic Sciences Research Center. Don A. Dillman is the principal investigator.

focus on describing the three-stage process of how respondents visually perceive survey information and discuss how survey designers can strategically use visual design elements and their properties to influence respondent behavior.

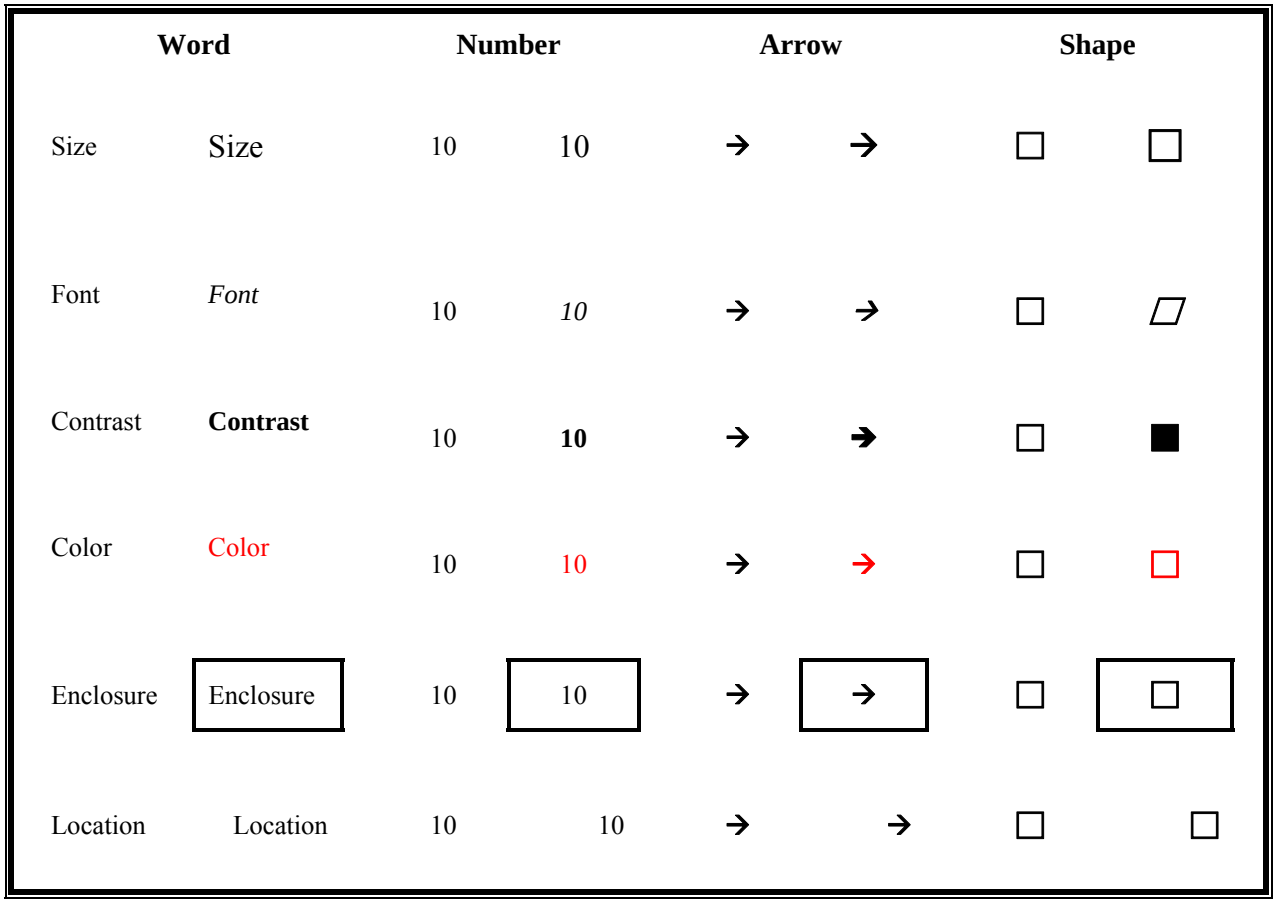
Visual design

Respondents determine meaning based on all of the visual information presented to them in the questionnaire. As mentioned earlier, words are extremely important in aural and visual communication; however, in visual communication other design features can convey additional information. In addition to letters and words, information is also communicated visually through various types of visual elements including lines and contours (| ⟩ ⌊), geometric shapes (□ △ ●), arrows (→), number forms (1, 2, iii, IV), punctuation marks (, . ?), mathematical (+ > = %) and currency ($ €) symbols. These elements convey information based on how they are used and on what they represent to respondents; therefore, the meaning assigned to particular elements may differ depending on the cultural expectations of respondents completing the questionnaire and on how these elements are presented to respondents. Surveyors can use each of these types of elements to convey meaning to respondents in each component of the question (the query, any definitions or instructions, the response categories, and the answer spaces) and throughout the survey questionnaire.

Respondents perceive variations in the form or presentation of visual elements and use them to interpret what is being asked of them; therefore, design features or properties can be used to modify the meaning respondents assign to visual information. Letters, words, numbers, arrows, shapes, and other visual elements can be presented in different ways: they can be small or big, in different fonts, light or dark, black or red, enclosed or not, or left or right justified (see

Figure 4.3). Similar to aural paralanguage conveyed through the speaker's voice, these properties act like a visual paralanguage to influence the meaning respondents assign to survey information. Visual design properties include size, font, contrast, color, enclosure, location, and other manipulations that modify how the elements are presented. Each of these properties can be used to make elements more or less visible to respondents as they complete the survey questionnaire. As discussed next, it is often these properties that are first recognized in the early stages of visual processing.

Figure 4.3: Visual design elements and their properties



These visual design elements and properties can be layered to create more complex visual images and graphics. For example, some corporate logos may only include one element, such as the white apple for Apple's company logo. However, many logos and other images include multiple visual elements such as the AT&T logo, which includes the name of the company and "Your world. Delivered." along with a graphical image of a globe to reinforce the text. The increased ease and affordability of including complex images and pictures in questionnaires has brought more attention to the influence they have on respondent answers. For example, respondents may interpret questions differently depending on the images included with the question (Couper, Tourangeau, and Kenyon, 2004). As more and more elements and properties are layered to create complex images, it is important to understand how the combination of different elements and properties influence how respondents assign meaning.

Most respondents are cooperative communicators and will try to make sense of the information in each survey question by drawing on all cues provided in the paper or web questionnaire (Schwarz, 1996). Thus, respondents use the visual design elements and properties that are included in the questionnaire (or the formal features of the questionnaire as discussed in Schwarz Grayson, and Knäuper, 1998) to guide them in completing the survey and help them understand what is being asked. Tourangeau, Couper, and Conrad (2004) discuss how visual design elements can be viewed as task or style elements. Visual design elements are included in questionnaires that are essential to the task of completing the survey (such as the words and numbers in the question stem and response options, symbols used to convey instructions, and graphics such as answer spaces). Visual design elements and their properties are also used to improve the overall style or look and feel of the questionnaire (such as logos and text to identify the survey sponsor). Respondents may interpret style elements as essential to completing the

survey so surveyors should use them carefully and understand their inclusion may influence how respondents complete the survey response process.

Respondents use the design properties to help them determine what elements to focus their attention on and to group or connect related information in the questionnaire. For example, placing components of the question in close proximity and within the same region can help respondents perceive them as related; locating them in the center of the screen and in a region with a colored background can help respondents focus their attention there and perceive that information as necessary to the task at hand. Including style information, such as logos or other graphics, in the header and footer and the right and left margins can help increase respondent motivation while not distracting from the information necessary to the task. After discussing how respondents process visual information, further examples will be presented of how visual design can be used to guide respondents through survey questionnaires.

Three-stage model of visual information processing

When people are presented with visual information, whether in a newspaper, website or survey questionnaire, many separate actions take place very quickly as the eye takes in the information and the human brain processes it to make sense of the page or screen. The ways in which people perceive and attend to survey information are determined by innate tendencies of visual information processing as described in recent work by Palmer (1999), Hoffman (2004), and Ware (2004). Respondents do not assign meaning to information all at once; instead, they process information in multiple steps with visual elements and their properties influencing each stage in the process. This paper applies concepts from the vision sciences to the survey context in

presenting a three-stage model for how respondents perceive and process visual information that is based on Ware's (2004) model of visual information processing.

The three stages of visual information processing are (1) *basic page layout*, (2) *information organization*, and (3) *task completion* (see Figure 4.4 for an overview of the three-stage model). In the first stage of visual processing, respondents rapidly perceive basic visual properties of the page or screen, such as changes in color or contrast, to understand the basic layout of information. In the second stage, respondents organize the information on the page by segmenting the page into regions, differentiating visual elements from the background, and grouping related elements based on their shared properties. Finally, in the third stage, respondents focus on completing the task at hand and sequentially process the components of individual survey questions.

Figure 4.4: Three-stage model of how survey respondents process visual information

Stage 1: Basic page layout	▪ Respondents quickly preattentively process basic visual properties of the page that "pop out" from other information based on their properties ▪ Respondents determine the basic layout of the page ▪ This stage is dominated by bottom-up or data driven processing
Stage 2: Information organization	▪ Respondents segment the page into regions ▪ Respondents differentiate visual elements from the background ▪ Respondents perceive groups among elements with shared properties ▪ This stage combines both bottom-up and top-down processing
Stage 3: Task completion	▪ Respondents focus on completing the task of responding to the survey ▪ Respondents sequentially process the components of individual questions ▪ Attention is restricted to 8 to 10 characters or the foveal view ▪ This stage is dominated by top-down processing

For many people, the three stages—understanding the basic layout of the page, organizing the information, and completing the task at hand—occur quite quickly. This is particularly true for surveys asking questions that can be anticipated and thought about by

respondents prior to their receiving them or when respondents are not highly motivated to respond. On the contrary, if the task is an unfamiliar one, going through these stages is likely to take longer because of more uncertainty about what should or should not be done. Now this process will be discussed in more detail to explain how visual elements and their properties are perceived by respondents at each of these three stages of information processing.

Basic page layout

Stage one begins immediately when the respondent glances at the survey page or screen; the respondent's eyes scan the entire visual scene simultaneously and work with the brain to extract simple visual properties, such as color, form, motion, orientation, and texture, and understand the basic layout of the page. This stage is dominated by *bottom-up processing* where visual information is quickly processed by the visual sensory system and only the visual scene itself influences how they are perceived (Ware, 2004). In bottom-up processing, meaning is easily determined by the senses alone so the respondent does not need to rely on his or her expectations, cultural knowledge, or the context of the situation to determine the meaning of the visual stimulus. For example, the respondent quickly perceives changes in color and contrast on the page or screen without applying information from long-term memory to help understand the meaning of the changes in color or contrast.

This stage of visual processing largely occurs prior to conscious attention, at the *preattentive level*, and influences which visual elements are attended to in later stages. During preattentive processing, one is gaining a general understanding of the broad visual field "at a glance" by noticing certain visual features or properties that stand out from the other visual information presented. Research has shown that when properties of visual design elements

deviate from other information on the page or screen, those elements are quickly recognized (almost instantaneously, ten milliseconds per item) and preattentively processed (Ware, 2004). Properties that stand out during preattentive processing include number, size, shape, contrast, enclosure, color hue and intensity, motion, convexity/concavity, and line orientation, length, and width. In Figure 4.5, the properties of the navigation buttons have been manipulated to distinguish them from the other written text on the page so respondents preattentively process them in stage one.

Figure 4.5: Manipulating the properties of the navigation buttons so that respondents Pre-attentively process them during stage 1 of visual processing

No visual difference	To move to the next question or to go backwards in the survey, please use the buttons at the bottom of the screen. Next question Back
Size	To move to the next question or to go backwards in the survey, please use the buttons at the bottom of the screen. Next question Back
Contrast	To move to the next question or to go backwards in the survey, please use the buttons at the bottom of the screen. Next question Back
Enclosure	To move to the next question or to go backwards in the survey, please use the buttons at the bottom of the screen. Next question Back
Contrast & Enclosure	To move to the next question or to go backwards in the survey, please use the buttons at the bottom of the screen. **Next question** **Back**

Visual elements and their properties are processed in relationship with the other information on the page or screen. As a greater number and variety of elements and their properties are manipulated, it becomes more difficult for the viewer to distinguish the ones that deviate from the other elements on the page. For example, if red is used to draw attention to a particular word, it will become more difficult for the respondent to preattentively process the color red as the number of other colors on the page or screen increases.

Information Organization

During the second stage, the visual system actively divides the visual scene into basic regions according to their shared visual properties, a step called ***segmentation.*** Once segmented into regions, the contours or boundaries are used to help respondents distinguish figure from ground. Respondents differentiate individual visual elements that are then used in further visual processing whereas those perceived as background recede from attention. In this paragraph, for example, the viewer differentiates the black text from the white background and then focuses on processing the words on the page.

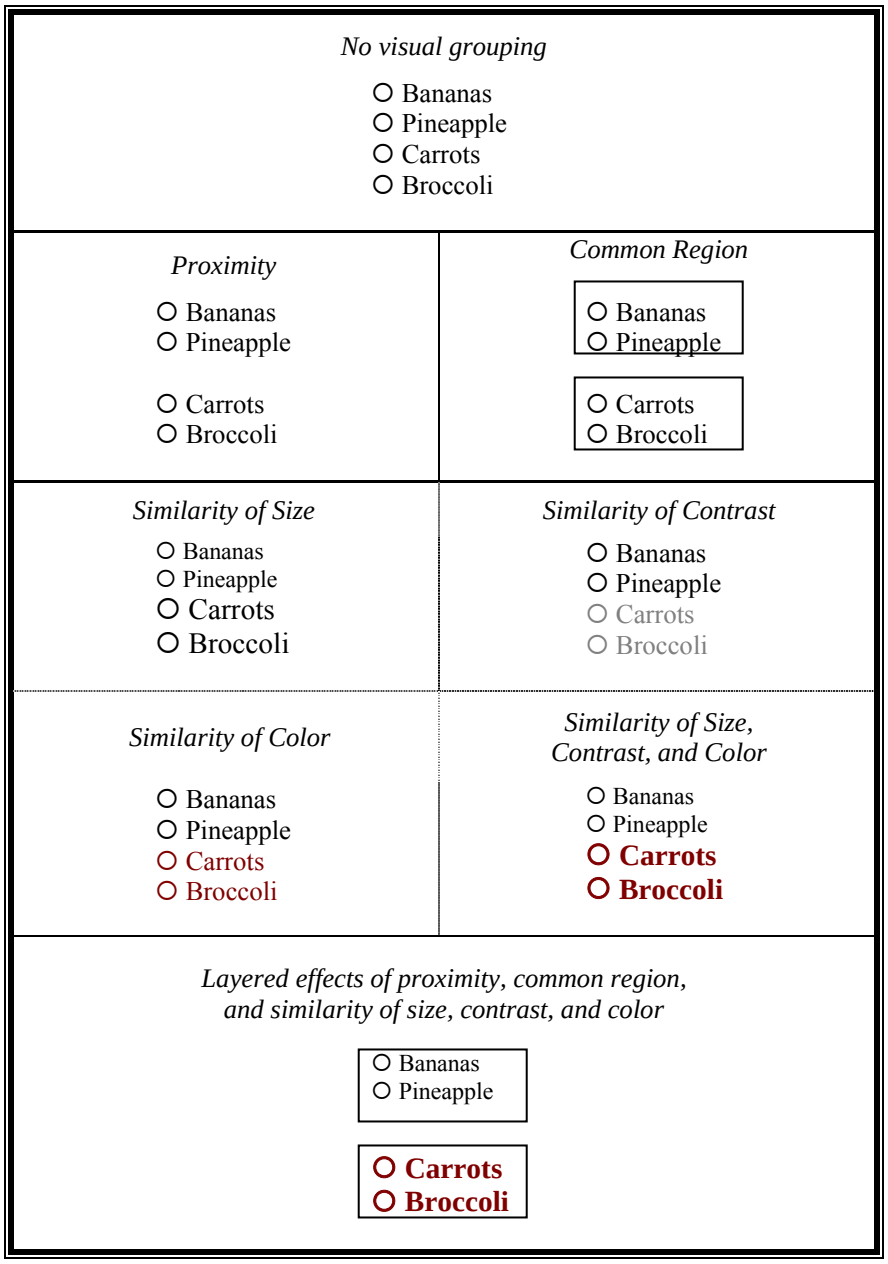
The second stage is an in-between stage or middle ground where the elements that are preattentively processed are now actively attended to and cognitively processed. This stage combines both bottom-up processing where meaning is determined by the visual scene and *top-down processing* where the context of the situation and the respondent's cultural knowledge, prior experiences, and expectations influence how they interpret visual elements. In this stage, respondents use prior information from completing surveys and other forms to identify the meaning of specific visual elements on the page or screen.

In this stage, the viewer also begins to perceive relationships or patterns among visual elements to help speed up processing and aid understanding. For example, respondents begin to perceive elements of similar contrast that are located close in proximity as a group, such as a set of response options. The Gestalt psychology principles of pattern perception can help surveyors understand how respondents perceive groups among visual elements with shared properties. The principles of proximity, common region, similarity, connectedness, and continuity can be used to encourage survey respondents to perceive elements as related and group them for use during the task completion stage. In Figure 4.6, different Gestalt principles can be applied to help respondents distinguish the fruits from the vegetables when processing a long list of response options. However, it is often the case that surveyors want respondents to process response options as one group and no visual grouping should be used, as in the top of Figure 4.6.

- *Spatial proximity:* Placing elements that should be grouped closer to each other than to other elements.
- *Common region:* Enclosing the elements that should be grouped in a shared area or region of space such as locating them within larger circles or squares.
- *Similarity:* Making the elements that should be grouped visually similar to each other but distinct from other elements through the use of color, size, shape, and contrast.

Also shown in Figure 4.6 is how multiple principles can be layered to provide stronger distinctions than could be achieved using only one of these principles. Thus, the survey designer can send messages that are strong or ambiguous depending upon whether these mid-level aspects of processing are developed consistently or inconsistently.

Figure 4.6: Examples of using Gestalt principles to help respondents perceive groupings and organize information in stage two of visual processing



Task completion

In the final stage, after respondents perceive and organize the basic organization of information on the page or screen, they begin the task of answering individual survey questions. The task completion stage is dominated by top-down processing whereby the meaning of objects is determined by the context of the situation and depends on the respondent's expectations and prior knowledge. To complete the task of responding to the survey, respondents focus their active attention on sequentially processing the components of each survey question (the question stem, any additional instructions or definitions, the answer spaces and/or response options). Once respondents perceive and process the parts of the survey question, they can then complete the steps of the survey response process to provide an answer.

As respondents attend to completing the task of responding to the survey, the visual field is constricted to a smaller visual span of about two degrees or eight to ten characters, called the foveal view. While the field of view gets smaller when switching from pre-attentive to attentive processing, it can become even smaller yet in stressful or cognitively demanding situations (Norman, 2003). Although only a small portion of the visual field can be taken into consideration in attentive processing, the few elements and properties in the narrowed field are processed more deeply and are better remembered for future use. Survey designers can strategically use visual design elements and their properties to help respondents complete the task of responding to the survey as they move through each of the three stages of visual information processing.

Using visual design to help respondents during the three stages of information processing

To illustrate how visual design influences web respondents as they complete the stages of visual information processing, examples are drawn from a series of Student Experience web

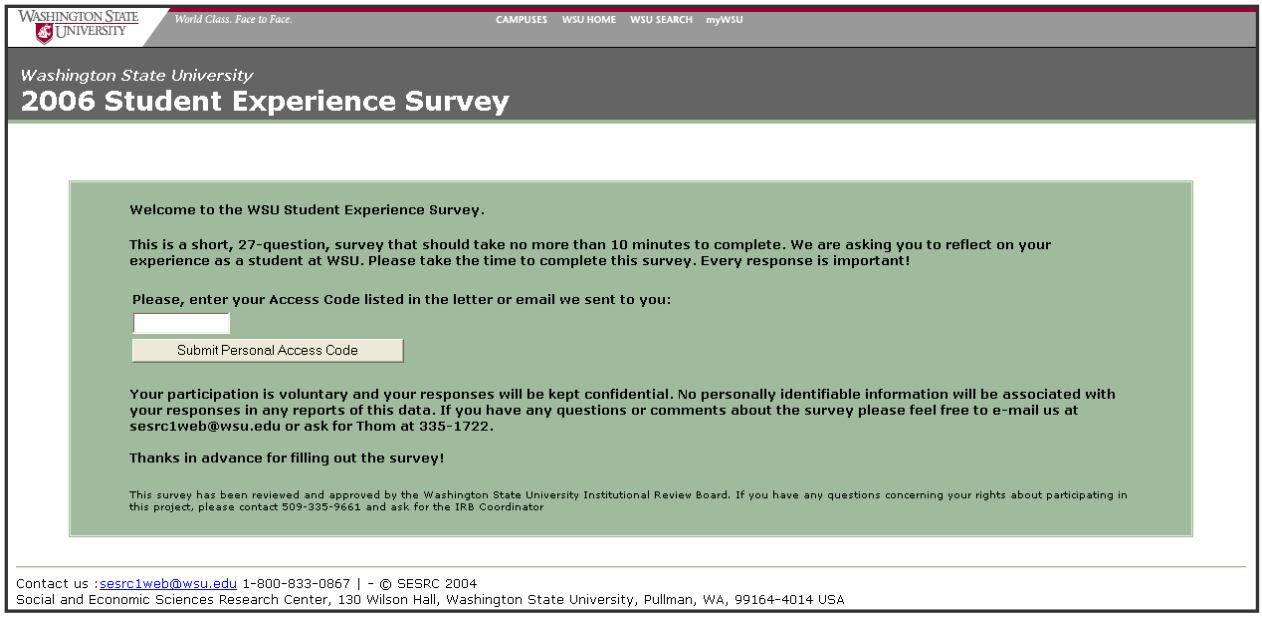
surveys conducted of Washington State University undergraduates from the Spring of 2002 to the Spring of 2006. Research on the influence of visual design in paper questionnaires has been summarized elsewhere (Christian and Dillman, 2004; Dillman, Gertseva, and Mahon-Haft, 2005; Jenkins and Dillman, 1997; Redline, Dillman, Dajani, and Scaggs, 2003) so the discussion here is focused on how visual design can be used to help guide respondents as they complete web surveys. These examples illustrate how surveyors can strategically vary the visual design elements and properties of web questionnaires to help respondents understand the basic layout and organization of information on the screen in the early stages of visual processing and then complete the task of responding to the survey questions in the third stage of processing.

When web survey respondents enter the survey address in their Internet browser or click on the link provided in the email to them, they first see the welcome page or opening screen (see example in Figure 4.7a). This page helps orient respondents by providing a description of the survey so they know they have found the correct page; surveyors also frequently place sponsorship and other information here, similar to what might be provided in a contact letter or email, to encourage respondents to begin the survey. At this screen, respondents are provided with additional information about their participation rights and are often asked to enter an individual access code before being able to proceed with the survey.

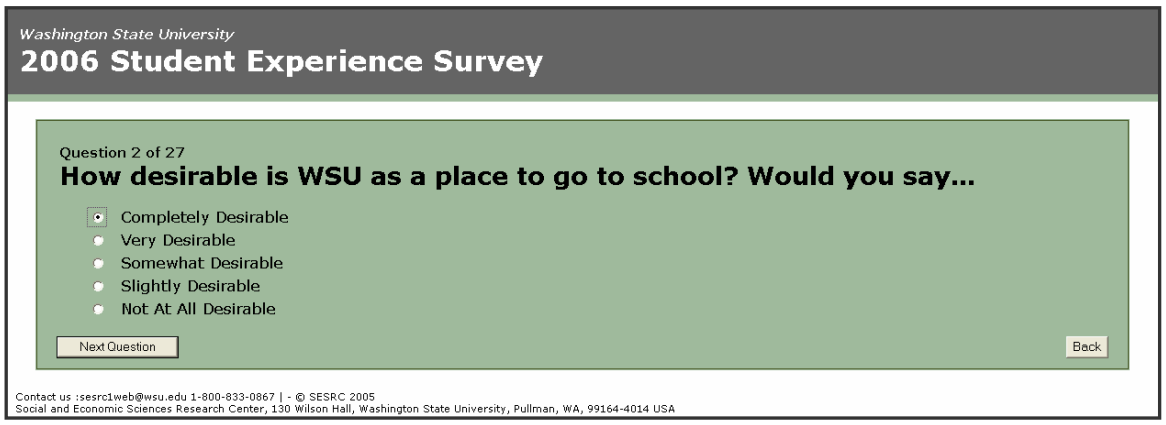
A consistent page layout across screens helps respondents understand the basic organization of information on the screen in the early stages of visual processing. Since respondents first notice basic visual properties such as color and form, the contours and different background colors help them segment the page into three regions: the header or banner region, the main question area, and the footer. The header and footer regions are consistent across the introductory and closing screens and in each of the question screens (see Figure 4.7).

Figure 4.7: Example web screens from the WSU Student Experience survey

a. Opening screen



b. First two question screens



The header region includes the name of the university and the title of the survey in white font against a gray background. The footer region contains contact information for respondents who have questions for the surveyor or who would like additional information about the survey. This information is located at the bottom of the page, similar to contact or designer information for many websites, with a white background and the text is in smaller black font to make it less visible since not all respondents will need this information.

The main question area contains the elements respondents need during the task completion stage to answer individual questions and complete the survey so this area is located in the center of the page with a green background color to focus their attention on the information in this region. Locating the components of each individual question in the same region (defined by the contours and the green background color) encourages respondents to group the question stem, any additional instructions or definitions, and the answer categories or boxes (see Figure 7b). Similarly, the visual properties of the components of each question can be manipulated to assist respondents as they sequentially process each part of the question. A larger size and darker contrast are used to distinguish the question stem from the answer categories. In some questions, additional instructions or definitions can be presented in italics (or another property can be manipulated) to differentiate them from the main body of the question stem (not shown in figure). White is used to help respondents distinguish where they should select a response or type in their answer. To encourage respondents to perceive the answer categories for closed-ended questions as a group, they are presented in the same size, font, and contrast and are located closer in proximity to each other than to the question stem or the navigation buttons.

In many web surveys, survey designers provide navigation buttons for respondents to use when moving through the survey, which are different from the back and forward buttons in their

Internet browser that people typically use to navigate between web pages. These navigation buttons allow greater control over how answers are saved as respondents proceed through the survey and the surveyor can collect paradata to record how respondents move forward and backward through the survey, how long they spend answering each question, and whether they change their answers to questions. As Figure 4.7 illustrates, the navigation buttons are also located in the central question area because respondents will need them to move through the survey; however, they are differentiated from the questions by using a light gray fill color (lighter than the gray used in the header region). The location of the navigation buttons is also consistent across screens; they are always located in the lower left and lower right of the green question region because surveyors want to encourage respondents to move forward and complete the questions in the survey. This layout is different from most web browsers but it eases the task for the respondents because the next question button is located directly below the answer categories and/or answer boxes so they can easily click the next button without having to scroll across the page (respondents can also use the tab key). The back button is located on the same horizontal line but on the right side since fewer respondents will need to move backwards in the survey.

To help respondents in the early stages as they process visual information, surveyors should use a consistent visual layout to aid respondents as they move through individual pages or screens in the questionnaire. In addition, designers should use lines and contours, contrast and color to help respondents segment the page into regions. Light background colors with sufficient contrast from the visual elements will help focus respondents' attention on the necessary task elements. Applying Gestalt principles can help respondents perceive groups and subgroups of information and identify related material. Survey designers should manipulate the properties of

visual design elements to make more important information stand out to respondents in the early stages of visual processing and to make information that respondents do not need or that will be used infrequently less visible to respondents. Lastly, it is important to use visual design elements and properties consistently to differentiate particular types of information (e.g. using white to delineate where respondents should report their responses and locating response formats in a similar location or along the same vertical or horizontal plane).

Examples of using visual design to help respondents answer individual survey questions

Several survey methodologists have been conducting experiments to test whether and how visual elements and properties influence respondents as they complete the cognitive steps in the survey response process and answer individual survey questions. This paper presents two examples to demonstrate how visual design can help guide respondents in providing answers to individual survey questions. These examples are from a question asking for a numerical response and from scalar questions asking respondents to select a response from a list of answer categories that lie along an implied continuum.

An experiment was embedded in the most recent Student Experience survey conducted in the Spring of 2007. This experiment tested the independent and combined effects of several manipulations on the type and amount of information respondents provide to a question asking for a month and year response in numeric format. Changing the size of the answer boxes, including a verbal instruction in the question stem, and providing a symbolic instruction located with the answer spaces were tested to see if they encouraged respondents to provide their answers in a desired format (see Figure 4.8 for an example of the comparisons and the results of the experiments). Providing respondents with a month box half the size of the year box, rather

than equal size month and year boxes, encouraged significantly more respondents to provide their response in the desired format, two digits for the month and four digits for the year (57% vs. 46%; χ^2=6.7, p=.010). As shown in Figure 4.8, a fewer percentage of respondents to the version with the half-size month box provided a word month (2% vs. 20%; χ^2=55.6, p=.000) and a greater percentage reported the month in numbers. However, more respondents provided the month using one-digit rather than the desired two-digit format (37% vs. 16%; χ^2=35.2, p=.000). In addition, more respondents reported the year using four digits (91% vs. 77%%; χ^2=24.0, p=.000).

Figure 4.8: The influence of sequentially manipulating the properties of the answer spaces, a verbal instruction included with the answer boxes, and a symbolic instruction provided with the answer spaces on responses to date questions

Question stem *verbal instruction	**What month and year did you begin your studies at WSU? *Please provide your answer using two digits for the month and four digits for the year.**			
				
	Equal size boxes	Half size month box	Half size month box, verbal instruction*	Half size month box, verbal instruction* & symbolic instruction
Word month	20.1	1.8	0.3	--
One-digit month	16.0	36.7	15.8	3.7
Two-digit month	63.9	61.5	83.9	96.0
Two-digit year	23.0	8.9	7.1	1.9
Four-digit year	77.0	91.1	92.6	97.8
Desired format Two-digit month & four-digit year	**46.3%**	**56.6%**	**78.0%**	**94.4%**

Chi-square tests comparing differences in respondents use of the desired format:
Influence of half-size month box 46% to 57% – χ^2=6.7, p=.010
Influence of verbal instruction with half-size month box 57% to 78%– χ^2=33.6, p=.000
Influence of symbolic instruction with half-size month box and verbal instruction 78% to 94% – χ^2=36.4, p=.000
Combined influence of half-size month box, verbal instruction, and symbolic instruction 46% to 94% – χ^2=176.6, p=.000

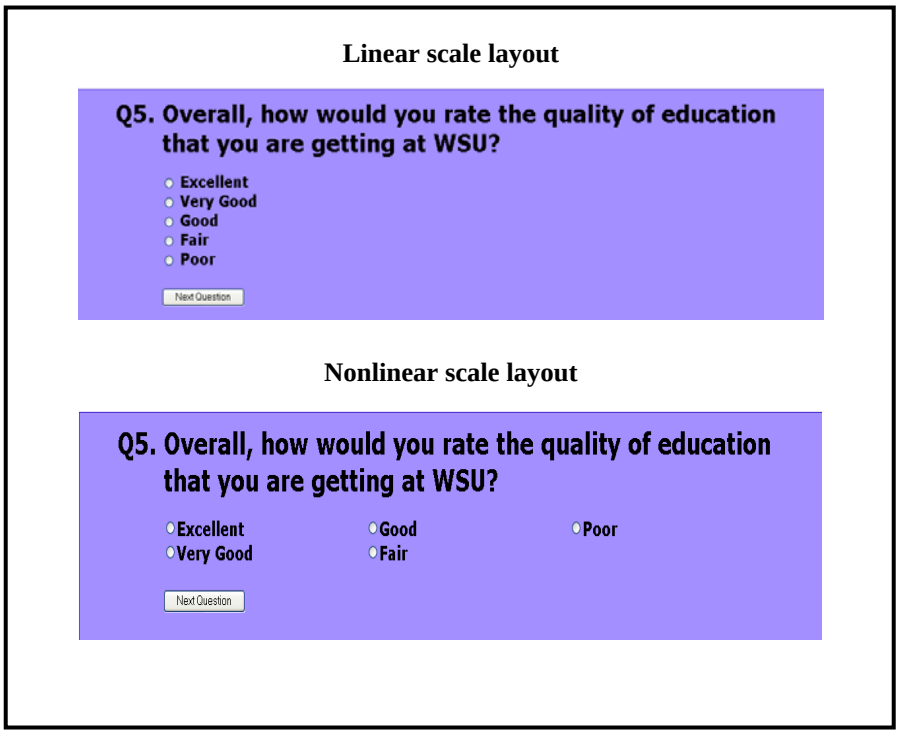
Adding the verbal instruction "Please provide your answer using two digits for the month and four digits for the year" significantly increased the percentage of respondents reporting their answer in the desired format from 57% to 78%, an increase of twenty-one percentage points (χ^2=33.6, p=.000). A significantly greater number of respondents reported the month in two digits rather than one when provided with the verbal instruction (84% vs. 62%; χ^2=40.8, p=.000); however, there was no significant difference in how respondents reported the year (93% vs. 91%; χ^2=0.4, p=.510). The addition of the symbolic instructions MM YYYY, where the number of letters represents the number of digits respondents should use in their response, significantly increased the number of respondents reporting their answer in the desired format from 78% to 94% (χ^2=36.4, p=.000). Providing respondents with additional information located where they provide their response (below the answer boxes) significantly increased the percentage of respondents reporting a two-digit month from 84% to 96% (χ^2=25.9, p=.000) and the percentage of respondents reporting a four-digit year from 93% to 98% (χ^2=9.7, p=.002).

The combined effects of the half-size month box and the verbal and symbolic instructions significantly increased the percentage of respondents reporting a response in the desired format from 46% to 94% (χ^2=176.6, p=.000). This experiment illustrates how sequentially manipulating the visual design of different components of the question can influence respondent behavior; varying the visual elements and their properties to convey a consistent message can encourage web respondents to provide responses in the format desired by the researcher. This often reduces respondent burden in web surveys where error messages can occur if respondents do not provide answers in the requested format.

Additional research has demonstrated how the visual design and layout of response scales influences how respondents interpret and respond to the scale, particularly with ordinal scales

where the categories are ordered along some underlying continuum. Presenting answer categories in one column versus two columns (see examples of linear and nonlinear formats in Figure 4.9) has been shown to influence how respondent answer scalar questions in both paper (Christian and Dillman, 2004) and web surveys (Christian, 2003). When response options are presented in one vertical column, most respondents perceive and process the options in the intended order. However, when options are presented in two or more columns, respondents may process the options horizontally, going across the page first, or vertically, going down the list first. Thus, placing all the response options in one vertical (or horizontal) column can help ensure that respondents process the categories in the intended order.

Figure 4.9: Examples of linear and nonlinear response scale layouts



Tourangeau, Couper, and Conrad (2004) further discuss how respondents expect the response scale to begin with the category in the top and left position (this category should represent one of

the endpoints) and expect more positive categories to be on top in a vertical list of response options. Their research supports other findings that respondents associate higher numbers with more positive categories when no visual display of the response scale is provided (Christian and Dillman, 2004; Dillman and Christian, 2005).

Additional research on response scales demonstrates how respondents expect the visual midpoint or middle position on the scale to represent the typical or average response and how respondents may use the visual midpoint as an anchor when forming their own judgment. Smyth, Dillman, and Christian (2007) report that students' ratings of the number of hours they study, watch TV, and use the computer are influenced by the hours presented within each category range, including the range of hours in the middle category. Their results on the web reinforce earlier findings on paper surveys (Schwarz, Hippler, Deutsch, and Strack, 1985; Rockwood, Sangster, and Dillman, 1997). Schwarz, et al. (1998) and Smith (1993) found that respondents perceive the average or middle position of the scale differently when the scale is displayed using a ladder versus a pyramid shape; respondents assume the middle or average position is lower on the pyramid scale than the ladder scale. Tourangeau, et al. (2004) also conclude that non-substantive options (e.g. don't know or prefer not to answer) should be visually separated from substantive options with a divider line or additional space so the visual midpoint aligns with the conceptual midpoint of the scale.

These research findings demonstrate that the visual design and layout of response scales can significantly impact how respondents interpret the scale and the responses they select. Decisions about how response scales are presented visually should be carefully evaluated and based on empirical findings where previous research has been conducted. Research now suggests three aspects surveyors should consider when designing response scales. First, surveyors should

present ordinal response scales in a consistent vertical layout (or a horizontal layout as in a series of items asked in a grid) to encourage respondents to process the options in the intended order. In addition, the conceptual distances between category labels should be roughly equal because respondents will interpret equivalent distances or separation between categories (i.e. response scales should be roughly balanced without large gaps between some categories and smaller gaps between other categories). Lastly, response scales should be presented where the visual midpoint of the scale represents the middle or average position on the scale.

Overall, research is now accumulating that demonstrates how visual design impacts respondents as they complete the task of responding to individual survey questions. These two examples illustrate how instructions to respondents should be provided where they are needed and in a format so they will be interpreted in the intended manner and that response scales should be designed to help respondents understand the order and meaning of the categories. The visual design of various components of the question influences respondents as they sequentially process the components of each question. Survey designers can selectively manipulate visual elements and their properties to encourage respondents to process information in the intended order during the task completion stage.

Conclusions and future directions

The type of communication available in each survey mode is important because how information is communicated influences the survey response process. It is imperative that survey methodologists understand how respondents perceive and process information because this initial stage influences whether respondents comprehend the question as the surveyor intends, the information they retrieve from memory, the judgment they form, and how they report a response.

Ware's (2004) three-stage model of visual information processing has been adapted to the survey context to explain how survey respondents perceive and process information presented to them in the questionnaire. This chapter has also illustrated how visual design can help web respondents to understand the basic layout and organization of information on the page, navigate the questionnaire, and sequentially process the components of the question as they complete the task of responding to individual questions.

Research on how visual design influences survey respondents has advanced significantly over the past ten years and many surveyors are applying these concepts when designing questionnaires. Prior to the 1990s, most survey methodologists did not contemplate how the visual design of survey questionnaires could impact respondent behavior. However, this growing area of research is requiring survey methodologists to learn and apply new concepts to the practice of survey design. This paper provides an initial framework to further our understanding of how the visual design of web surveys influences respondents as they complete the survey and have focused on the practical application of these concepts to help surveyors make informed design decisions.

The visual design of questionnaires is still a relatively new area of research that will continue to develop, particularly with the influence of the Internet and other computer technologies on survey methodology practices. Survey researchers should continue to apply concepts from the vision sciences to advance our theoretical understanding of how survey respondents perceive and process visual information. In addition, methodologists should continue to conduct experiments testing the potential impacts of various visual design elements and properties on respondent behavior at each of the three stages of visual processing.

References

Christian, Leah Melani. 2003. *The influence of visual layout on scalar questions in web surveys.* Master's Thesis: Washington State University, Pullman, WA.

Christian, Leah Melani and Don A. Dillman. 2004. "The Influence of Symbolic and Graphical Language Manipulations on Answers to Paper Self-Administered Questionnaires." *Public Opinion Quarterly* 68: 57-80.

Christian, Leah Melani**,** Don A. Dillman, and Jolene D. Smyth. 2007. "Helping Respondents Get it Right the First Time: The influence of words, symbols, and graphics in web surveys." *Public Opinion Quarterly* 71: 113-125.

Couper, Mick P., Roger Tourangeau, and Kristin Kenyon. 2004. "Picture This! Exploring Visual Effects in Web Surveys." *Public Opinion Quarterly* 68: 255-266.

Dillman, Don A., Arina Gertseva, and Taj Mahon-Haft. 2005. "Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles." *Journal of Official Statistics* 21:183-214.

Dillman, Don A. and Leah Melani Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17(1): 30-52.

Hoffman, Donald D. 2004. *Visual Intelligence*. New York: Norton.

Jenkins, Cleo and Don Dillman. 1997. "Towards a Theory of Self-Administered Questionnaire Design." In L. Lyberg, P. Biemer, M. Collins, et al. (Eds.), *Survey Measurement and Process Quality.* New York: Wiley-Interscience. 165-196.

Norman, Donald A. 2003. *Emotional Design: Why We Love (or Hate) Everyday Things*. New York: Basic Books.

Palmer, Stephen E. 1999. *Vision Science: Photons to Phenomenology*. London: Bradford Books.

Redline, Cleo, Don A. Dillman, Aref N. Dajani, and Mary Ann Scaggs. 2003. "Improving Navigational Performance in US Census 2000 by Altering the Visual Administered Languages of Branching Instructions." *Journal of Official Statistics* 19: 403-419.

Rockwood, Todd H., Roberta L. Sangster, and Don A. Dillman. 1997. "The Effect of Response Categories in Questionnaire Answers: Context and Mode Effects." *Sociological Methods and Research* 26: 118-140.

Schwarz, Norbert. (1996). *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation.* Mahwah, NJ: Lawrence Erlbaum.

Schwarz, Norbert, Carla Grayson, and Barbel Knäuper. 1998. "Formal Features of Rating Scales and the Interpretation of Question Meaning." *International Journal of Public Opinion Research* 10: 177-183.

Schwarz, Norbert, Hans J. Hippler, Brigitte Deutsch, and Fritz Strack. 1985. "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments." *Public Opinion Quarterly* 49: 388-395.

Smyth, Jolene D., Don A. Dillman, and Leah Melani Christian. 2007. "Context Effects in Web Surveys: New Issues and Evidence." In the *Handbook of Internet Psychology*. Oxford University Press.

Tourangeau, Roger, Mick Couper, and Fred Conrad. 2004. "Spacing Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68: 368-393.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Oxford University Press.

Ware, Colin. 2004. *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufmann.

CHAPTER FIVE

THE FUTURE OF MIXED-MODE SURVEYS

The use of survey data in sociological analyses is being confronted by major changes in how such information is collected and analyzed. These changes in survey methodology are requiring new knowledge and tools to understand their impact on data quality. The development of new technologies, particularly computers and the Internet, have had profound effects on all aspects of survey research, including data collection. Coverage problems associated with the increased use of cell phones and declining response rates to telephone surveys have led survey researchers to question surveying respondents only by landline telephone. We have already witnessed the profound impact of web surveys; however, instead of replacing other modes, they are still limited because many people remain "offline" and it is difficult to contact those who are "online." Interest in mixed mode surveys has increased dramatically over the past decade as surveyors are developing various types of mixed-mode survey designs where some respondents are surveyed by one mode and others by another mode. The flexibility of mixed-mode survey designs allows researchers to combine modes of data collection in new and innovative ways.

As mixed-mode surveys continue to proliferate, survey researchers need to analyze their effects on coverage, sampling, nonresponse, and measurement error. Strategically combining different modes of data collection can help improve coverage and sampling error from noncoverage bias in web and telephone surveys. In addition, sequentially employing less expensive modes of data collection before more expensive modes can decrease costs over using an expensive mode alone, while often helping to improve response rates and reduce nonresponse bias. Understanding the effects of survey mode on how questions are measured is especially

important as researchers are increasingly comparing and combining responses to surveys conducted by different modes. Since various social, cultural, and technological features of the survey mode can impact the survey conversation, it is essential to understand how these features of different modes influence respondents as they complete the survey response process.

The challenge of addressing measurement issues in mixed-mode surveys involves seeking answers to many questions. In this dissertation, I focused on exploring how one feature of web and telephone surveys, the form of communication during the survey conversation, influences the response process and the answers respondents provide to web and telephone surveys. I have addressed whether people provide different answers to aural versus visual surveys and whether those differences can be mitigated by the use of different question formats. I have also examined the multiple steps respondents perform to process visual information presented to them in the questionnaire and the influence of the visual presentation of survey information on their responses.

Chapters Two and Three demonstrated that respondents surveyed by telephone consistently provide more positive responses than those surveyed by web across various types of response scales and a variety of different question topics and scale labels. Both of these chapters also showed that different ways of presenting response scales significantly influenced the answers respondents provide. Web and telephone respondents are impacted by what labels are provided, whether all or only the endpoint categories are labeled, and whether the categories are presented all at one time or in two-steps. Respondents provided more positive responses to fully labeled scales than to scales where only the endpoints are labeled. In addition, more respondents endorsed the most positive category when the response scale is presented in two-steps (the first asked respondents to report the direction, positive or negative, and then the computer skipped

them to the appropriate follow-up question that asked respondents to evaluate the strength of their position) than when all of the categories are presented at one time. Lastly, respondents had more difficulty responding to scalar questions in paper and web surveys when no visual display of the scale was presented and respondents expected higher numbers to be associated with more positive categories. Overall, the presentation of the response scales significantly influences how web and telephone respondents interpreted the scales and the answers they reported.

Since respondents provide different answers when surveyed using alternate modes, the results that sociologists and other social researchers report may be influenced by features of the survey mode. It is important for researchers to continue to examine how these methodological effects are influencing their findings, particularly in longitudinal surveys where the goal is to measure change over time so when the mode changes between waves it can have significant effects on the ability to accurately isolate the effects of time from the effects of survey mode (Dillman and Christian, 2005). Even though research has shown that some question formats can translate effectively across modes that rely on visual or aural communication, such as multiple item questions asking for yes or no responses (Smyth, Christian, and Dillman, Forthcoming), we have not yet identified a format for asking rating scales that translates effectively across modes. Survey researchers may need to consider statistical adjustments to help mitigate these mode differences between telephone surveys and other modes that can present scales visually.

Because of the importance of response scales as measurement tools and the impact of survey data on social research, it is important to conduct further research to evaluate why respondents provide different responses when surveyed by telephone. Since telephone surveys tend to be conducted at a quicker pace than surveys by other modes, slowing the pace of the telephone interviews might encourage respondents to spend more time carefully considering

each category, forming a judgment, and perhaps resulting in less positive responses. In addition, the lack of visual communication in telephone surveys may require new strategies for supplementing aural communication. As the growing research in visual design suggests, the presentation of survey information can be manipulated to help respondents understand the meaning of survey information, including individual questions and response options. The reliance on aural communication in telephone surveys necessitates creative ways of helping respondents comprehend the meaning of each question as the researcher intends.

The fourth chapter of this dissertation provided a conceptual framework to help researchers understand generally how people process visual information and specifically how survey respondents visually perceive and process information presented to them in the questionnaire. Respondents first understand the basic layout and organization of information on the page, including segmenting the page into regions, differentiating specific visual elements, and grouping elements based on their shared properties. Then, respondents attend to the task of completing the survey and sequentially process the individual questions and their component parts. Generally, this chapter explored how visual design can be used to encourage respondents to process the information in the order intended by the researcher, to read and apply instructions by providing them where respondents will need them, and to help respondents understand the meaning of the questions and how to report their answer.

Survey designers need to continue to research how the visual design of survey questionnaires can help respondents process and comprehend survey information in the order and manner intended by the researcher. Using visual design to guide respondents as they complete self-administered surveys, similar to the benefits of using interviewers in face-to-face and telephone surveys, could bring more equivalency in how responses are measured in different

modes. Further research is needed on the impact of varying the visual design of the survey questionnaire, and the visual design used in contacts to respondents, such as pre-notice letters and postcard reminders. As technological capabilities continue to change and more surveyors are conducting web surveys, even more research is needed that applies the science of visual perception to the design of survey questionnaires. This area of research is greatly expanding as the potential for manipulating the presentation of visual information in both paper and web questionnaires is becoming cheaper and more accessible.

Further research is needed on how survey researchers can best design questionnaires that can translate effectively across modes. Questionnaire design for mixed-mode surveys where most of the responses are expected by one mode should maximize the effectiveness of the design for the primary mode and allow it to inform the design of the questionnaires for the secondary or supplementary modes. However, optimal design of survey questionnaires where each mode of data collection is equally important should focus on presenting an equivalent stimulus to respondents across different modes to reduce measurement differences. Questionnaire design in mixed-mode surveys should recognize how differences in meaning may depend on the presence or absence of an interviewer, the cultural and technological features of each mode, and how information is communicated with respondents.

References

Dillman, Don A. and Leah Melani Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17(1): 30-52.

Smyth, Jolene D., Leah Melani Christian, and Don A. Dillman. Forthcoming. "Does 'Yes or No' on the Telephone Mean the Same as 'Check-All-That-Apply' on the Web?" *Public Opinion Quarterly.*

METHODOLOGICAL APPENDIX

This appendix contains additional information about the analyses of the survey data presented in Chapters Two and Three, and the methodology of the survey data discussed in Chapter Four.

Chapter Two -- multivariate analysis

In addition to the bivariate analyses presented in Chapter Two, I performed a series of ordinal and multinomial logistic regression models for nine comparisons where the effects of survey mode, scale format, and their interaction could be modeled simultaneously (for six of the comparisons, the scale formats being compared were the fully-labeled to the polar point labeled and for three of the comparisons, the agree-disagree scale was compared to the construct-specific scale). I first ran ordinal logistic regression models because the dependent variables were responses to scales with ordered categories, where the distances between categories are not equal as in interval level data. The ordered logistic regression model assumes that the slope coefficients are equal across each regression equation (Long and Freese, 2003). In the models for this data, the proportional odds (or parallel regression) assumption was consistently violated suggesting that the multinomial logistic regression model might be more appropriate.

The multinomial logit model simultaneously estimates coefficients for all comparisons among the dependent categories, where the effects of each independent variable are reported relative to the base category or comparison group (Long and Freese, 2003). For each comparison, I added the two independent variables and the interaction variable one at a time as well as examined the overall combined model that included the independent and interaction variables. The interaction of survey mode and scale format was not significant in any of the nine

comparisons and because the independent effects of mode and format mirror the bivariate results, those results were presented in Chapter Two to simplify the analysis and make the findings more accessible to readers of varying statistical backgrounds. However, I include a brief discussion of the results from the multinomial models in the next two paragraphs.

For the comparisons of the fully labeled and polar point labeled scales across web and telephone, the interaction of mode and scale format was not significant in any of the six tests. Independently, survey mode had a significant influence on responses in four of the six comparisons, where telephone respondents were more likely to select the most positive category than web respondents. Scale format also independently influences responses in four of the six comparisons, where respondents to the fully labeled format were more likely to select the most positive (fifth) category than respondents to the scale where only the endpoints were labeled.

The interaction of scale format and survey mode was also not significant in any of the three comparisons of the agree-disagree and construct-specific scales across web and telephone modes. Similar to the comparisons of the fully labeled and polar endpoint labeled scales, mode had independent effects on responses, where telephone respondents were more likely to select the most positive category. The effect of labeling the scales with agree-disagree versus construct-specific labels was significant in all three comparisons but in opposite direction in one of the comparisons. Respondents to the construct-specific scales were more likely to select the most positive category in two of the three comparisons.

Chapter Three -- additional multivariate analysis

In addition to the bivariate and multivariate analyses presented in Chapter Three, I also performed a series of ordinal logistic regressions to test whether these data can be analyzed using this model without violating the proportional odds assumption. However, this assumption was violated in nearly all of the models ran. The multinomial logit model was used as an alternative since this assumption was violated and to provide symmetry with the analyses conducted on the data from the Fall 2004 survey. The results from the multinomial models are presented in Chapter Three.

Many researchers use Analysis of Variance (ANOVA) models to analyze responses to ordinal scales even though most rating scales are not measured at the interval level and are frequently positively skewed rather than normally distributed. Because other researchers prefer the simplicity of presenting ANOVA results, even though rating scales are more appropriately analyzed using ordinal and multinomial logit models, I also compared the results from the multinomial models presented in Chapter Three to results for a series of ANOVA models for the same comparisons.

For the Analysis of Variance (ANOVA) and multinomial logistic regression models, the lowest three categories (1 through 3) are combined because none or only a few respondents selected those categories. Table A.1 compares the results from the ANOVA and multinomial logistic regression models for these ten questions. For the (ANOVA) models (df=3), I report the adjusted R^2 and F-tests for the overall model and for each variable (survey mode, scale format, and the interaction of mode and format). For the multinomial logistic regression models (df=4), I report the pseudo R^2 and likelihood-ratio chi-square tests for the overall model and for the same three variables (mode, format, and the interaction).

Table A.1: Comparing the ANOVA and Multinomial Logistic Regression Models

	q#	Scale labels	n	Model	R^2	Overall model	Survey Mode	Scale Format	Mode* Format
Fully labeled one-step	2	Desirable/ Undesirable	2214	ANOVA	.210	196.62	41.84	358.62	5.68
				Logistic	.093	652.92	81.38	166.94	16.40
	4	Satisfied/ Dissatisfied	2213	ANOVA	.062	49.88	40.87	65.85	1.63
				Logistic	.046	307.32	54.81	95.22	14.10
	7	Satisfied/ Dissatisfied	2215	ANOVA	.140	120.96	40.64	211.24	4.14
				Logistic	.066	456.94	49.03	140.95	21.63
	8	Satisfied/ Dissatisfied	2210	ANOVA	.119	100.17	22.53	139.62	0.12
				Logistic	.062	418.02	27.26	180.21	39.86
	5	Accessible/ Inaccessible	2208	ANOVA	.115	96.56	27.39	178.67	4.30
				Logistic	.058	412.04	43.03	111.76	11.36
	9	Accessible/ Inaccessible	2200	ANOVA	.209	194.20	45.05	363.82	7.92
				Logistic	.084	581.09	61.04	143.01	12.89
Polar point labeled one-step	19	Desirable/ Undesirable	2210	ANOVA	.006	5.14	1.96	11.93	1.32
				Logistic	.013	91.43	13.47	46.30	10.47
	23	Satisfied/ Dissatisfied	2204	ANOVA	.014	11.40	10.97	0.95	2.12
				Logistic	.023	156.72	36.14	56.43	5.45
	25	Satisfied/ Dissatisfied	2211	ANOVA	.019	15.37	6.97	0.62	7.69
				Logistic	.026	153.03	19.30	67.49	6.42
	26	Confident/ Unconfident	2206	ANOVA	.008	6.90	6.32	.08	1.98
				Logistic	.018	112.11	11.84	49.69	12.86

Notes:
For the Analysis of Variance (ANOVA) models the df=3 and the adjusted R^2 is reported; F-tests are reported for the overall model, the survey mode and scale format variables, and for the interaction variable (mode*format).
For the Multinomial logistic regression models the df=4 and the pseudo R^2 is reported; Likelihood-ratio chi-square tests are reported for the overall model, the survey mode and scale format variables, and for the interaction variable.
Significant p-values ($p<.05$) for the F-tests and likelihood-ratio chi-square tests are shown in black and p-values ≥ .05 in gray

Since the multinomial logit results are presented in Chapter Three, I will only briefly summarize the similarities and differences in the results from the multinomial and ANOVA models. The results of both models consistently suggest that the survey mode (web versus telephone) and the scale format (fully labeled one-step versus fully labeled two-step and polar endpoint labeled one-step versus fully labeled two-step) independently influence the answers respondents provide to rating scales. The survey mode and scale format independent variables are significant for all ten comparisons across both models (see Table A.1). The difference between the models stems for what is happening with the interaction variable. The interaction of mode and format is significant in eight of the comparisons in the multinomial models but in only five of the ten comparisons in the ANOVA models (see Table A.1). The multinomial models are more sensitive to the interaction because all of the categories of the dependent variable are estimated simultaneously.

Chapter Four – methodological information about the survey data reported

The methodological details for the Fall 2004 and Spring 2006 telephone surveys are not repeated in Chapter Four because they are presented in Chapter Two and Chapter Three, respectively. The data presented in Chapter Four about how the wording of questions requesting date information influences the amount and type of information respondents provide (pg. 94) are from the telephone versions of the mixed-mode surveys discussed in these two chapters.

The data presented later in the chapter regarding the visual design of questions requesting date information (pgs. 110-112) are from experiments embedded in the most recent Washington State University Student Experience Survey, conducted in the Spring of 2007. The implementation procedures for the web survey were similar to the other surveys where students were initially contacted by mail and provided a $2 incentive as well as their unique access code to verify they were eligible to the survey. Subsequent contacts were sent via email and postal mail. The response rate was 55% with 1,967 of the 3,577 students sampled completing the survey. Six version of the survey were developed and at least 300 respondents completed each version of the survey; results from only four of the six versions are presented in Chapter Four.

Reference

Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*, *revised edition*. College Station, Texas: Stata Corporation.