The Effects of Mode and Format On Answers to Scalar Questions in Telephone and Web Surveys[1]

By: Leah Melani Christian, Don A. Dillman, and Jolene D. Smyth
lmchristian@wsu.edu, dillman@wsu.edu, jsmith@wsu.edu
Washington State University
Pullman, Washington 99164-4014 U.S.A.

ABSTRACT

The proliferation of mixed-mode surveys, where data is collected from respondents using different survey modes, raises concern about whether respondent characteristics are being measured equivalently across modes since the data are often combined for analysis. Previous research indicates that three types of factors differentiate survey modes: technical and cultural factors related to the mode or media itself, the impact of interviewer presence or absence, and how information is transmitted or conveyed during the survey (De Leeuw 1992). These differentiating factors can help us understand why variations in survey responses may arise based on the mode of data collection.

In addition to mode differences, survey designers often construct questions differently depending on the mode being used to survey respondents. For example, the advent of telephone surveying significantly influenced how scalar questions are asked of respondents. Since respondents to telephone surveys typically do not have any visual stimulus to remind them of the answer categories and to facilitate the telephone interviewing process, telephone surveyors often construct scalar questions where only the polar endpoints are labeled. However, fully labeled scales, where all categories of the scale are given a verbal label, are often used in face-to-face interviews using show cards and in mail and web surveys where the categories can be displayed as part of the questionnaire stimulus.

In this paper, we compare different scalar formats using both telephone and web modes to identify ways of asking scalar questions that present the same stimulus to respondents across telephone and web modes so that scalar questions can be constructed optimally for mixed-mode surveys. We designed a field experiment, collecting data from Washington State University students, using a total of 6 questionnaire versions (three telephone and three web). Since the survey was fielded simultaneously, using very similar procedures for both telephone and web respondents, we can estimate "pure" mode effects (Biemer 1988). We include several experimental manipulations designed to compare the *same scales across telephone and web modes* as well as *different scales within these modes*. These comparisons test the independent effects of survey mode and scale format as well as whether the interaction of mode and format influence responses to scalar questions. Our results from several multinomial logistic regression models indicate that mode and format independently influence responses to scalar questions; however, the interaction of mode and scalar format is not significant in any of the models.

[1] Analysis of these data was supported by funds provided to the Washington State University Social and Economic Sciences Research Center (SESRC) under Cooperative Agreement #43-3AEU-1-80055 with the USDA-National Agricultural Statistics Service, supported by the National Science Foundation, Division of Science Resource Statistics. Data collection was financed by funds provided to the SESRC by the Gallup Organization.

We include 26 mode effects comparisons of 5-point fully labeled and polar point labeled scales using various types of substantive verbal scales (i.e. satisfied/dissatisfied, agree/disagree, etc.) and 4 comparisons of 11-point polar endpoint labeled scales *across telephone and web modes*. Overall, we find that telephone respondents provide more positive ratings than web respondents (25 of 30 comparisons statistically significant) and are more likely to select the most positive extreme endpoint category. Our findings, in conjunction with previous research, suggest that the inability to visually present the response categories to telephone respondents and perhaps the faster pace of telephone interviews together seem to encourage more extreme positive ratings among telephone respondents when compared to mail, web, and face-to-face respondents (who receive a show card visually displaying the categories).

We also include comparisons of several types of scales *within telephone and web modes* testing the differential effects of verbal and numerical labeling and various "presentations" of the response scale. Our findings, in conjunction with other research, suggest that respondents answer fully labeled and polar endpoint labeled scales differently and that fully labeled scales are more desirable because they yield higher reliability and validity ratings. Further, construct-specific scales should yield higher reliability when compared to agree/disagree scales since respondents only have to map their answer to the underlying concept of the construct being asked in the question. If numerical labels are used in addition or instead of verbal labels, we suggest assigning higher numbers to more positive categories as this is consistent with respondents a priori expectations and because it is easier for respondents to select the category corresponding to their underlying attitude. Presenting the most positive versus negative category first does not seem to influence how respondents answer polar point labeled scales. Lastly, the use of an instruction "You may use any of the numbers 5, 4, 3, 2, or 1" or changing the direction of the instruction to begin with "1" do not significantly influence how telephone respondents answer polar point labeled scales.

Overall we find significant independent effects of survey mode and scale format. The results from our research suggest caution when combining responses to scalar questions collected from telephone and web respondents for analysis, as is customarily done in mixed-mode surveys. Survey designers also need to pay careful attention to how constructing scales differently for particular modes can influence responses. Thus, careful research is needed to determine scalar formats that provide the same stimulus across telephone and web modes; however, in our current state of research, some effects of survey mode may still be unavoidable.

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 particular 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 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 sometimes 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 illuminates 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 the *same 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

Most surveys contain several scalar questions and their popularity and frequent use means that many ways of constructing scalar questions have been developed over the years. In this paper, we report the results of numerous experimental comparisons of several different

scalar formats to investigate various aspects of scalar construction and to determine which scalar formats translate optimally across modes. In this endeavor we analyze 74 experimental comparisons from one survey using 6 versions (3 telephone and 3 web) of 13 scalar questions. 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 include comparisons of the *same 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 responses to 5-point fully and polar point labeled scales as well as 11-point polar point labeled scales from telephone and web respondents (see Figure 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). The comparisons together provide insights into constructing scales optimally for mixed-mode surveys. Since we include a large number of comparisons in this paper, to facilitate connections between the theory and results we present general procedures here but reserve discussion of specific previous research, theoretical rationales, and detailed description of manipulations for presentation with the results of each set of comparisons. We group the comparisons into two

sections: mode effects and scalar format effects and have included a summary figure (Figure 3) of the main findings for the various experimental manipulations.

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 fifty-nine percent 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 10 callback attempts were made. Web respondents for whom we had an email address (about 2/3 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, which was provided in each contact to the respondent.

For 10 questions, we performed a series of multinomial logistic regression models 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 (2*2, df=1) and the associated p-values for both tests for independent survey mode effects 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 now only 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, Sangster, Tarnai, and Rockwood 1996) and in self-administered surveys, norms of silences are not an issue. Thus, telephone respondents usually take less time when answering, a behavior that cognitive psychologists have suggested results 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 chapter by Conrad, Schober, and Dijkstra 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 chapter by St-Pierre and Béland 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, Strack, Hippler, and Bishop 1991). 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, Couper, Tourangeau, and Galesic 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. 1991, 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 et al. chapter). 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 (i.e. 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 7-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 7-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, Phelps, Tortora, Swift, Kohrell, and Berck 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, it appears 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 5-point fully labeled and polar point labeled scales using various types of verbal labels (see Figure 1). First, we test for

mode effects across telephone and web among fully labeled scales with 5 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 1a). In all nine comparisons, we find that respondents to fully labeled scales provide higher ratings when surveyed by telephone than by web. Eight of the nine mean comparisons indicate that the telephone ratings are significantly higher (see Table 1), 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 5 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 1b). Again, we find that telephone respondents provide higher mean ratings than web respondents and that they are more likely to select the extreme category. In twelve of the thirteen comparisons, we find that telephone respondents provide higher mean ratings than web respondents; eleven are significant (see Table 2). Telephone respondents select the extreme positive category more frequently than web respondents for ten of the thirteen comparisons (7 of the 13 chi-square are significant).

We also include four comparisons of 11-category polar point labeled (worst possible/best possible) scales to test for mode effects across telephone and web (see Figure 1b). Similarly to the above findings, we find that respondents provide more positive ratings when surveyed by telephone than by web in all four comparisons but only three of the four comparisons are

statistically significant (Table 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 5 category response scales in 22 of the 26 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 11-category polar endpoint labeled scales also provide more positive ratings than web respondents in 3 of the 4 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 (i.e. satisfaction scales, agree/disagree, construct-specific, etc.).

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. 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 telephone interviews 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 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.

Thus, in addition to the across mode comparisons discussed earlier, we also include several comparisons of different types of response scales *within* both telephone and web modes to identify question format effects. 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 the same 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 (i.e. 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 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 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 also 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 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.

Agree-Disagree vs. Construct-Specific scales

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. In addition, respondents answer agree/disagree scales in 2/3 of the time that it takes them to answer questions with construct-specific scales (Krosnick. 2005 AAPOR Short Course – requesting citation information). However, research on acquiescence suggests that people have a tendency to agree, particularly in the social presence of someone else (i.e. an interviewer) so respondents are more likely to select agree than disagree, yes than no, and true than false. Krosnick (2005 AAPOR Short Course – requesting citation information) has suggested that using construct-specific response options where verbal labels are designed using the underlying concept or construct being measured in the scale should decrease acquiescence. Additionally, the process of respondents mapping their answer to the appropriate response option should be more accurate when the response options are construct-specific because they only have

to think along one dimension, the underlying construct, instead of the underlying construct *and* whether they agree/disagree. To test whether respondents to agree/disagree scales are more likely to acquiesce, we include 3 comparisons of agree/disagree labeled scales (i.e. strongly agree, agree, neutral, etc.) to construct specific scales (i.e. very accessible, somewhat accessible, etc.) for both telephone and web respondents (see Figure 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 construct-specific category than to select the most positive agree/disagree category (see Table 5). Respondents to the agree/disagree scales are most likely to report that they "agree" whereas responses to the construct-specific scales are more evenly distributed over the two most positive categories (i.e. very and somewhat accessible). However, the pattern for the third comparison is slightly different. Here, responses to the agree/disagree scales yield significantly more positive ratings within both telephone and web modes. 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, it seems that when respondents are more positive on the topic of the question, the construct-specific scales yield more positive ratings than the agree/disagree scales; however, when respondents have more negative viewpoints, they yield more negative ratings.

Overall, for two of the three comparisons, respondents seem to avoid the "strongly agree" category and instead most respondents choose the "agree" category. Telephone and web respondents seem to acquiesce by selecting "agree" but seem to avoid the "strongly agree" category when answering agree/disagree than construct-specific scales. The across mode results

reported earlier in this paper suggest that telephone respondents provide more positive ratings regardless of whether agree/disagree or construct-specific labels are used.

Assigning 5 versus 1 to the most positive category

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. 1991). 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 one 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 5-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 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 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 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 2d). Within

the telephone mode the means are slightly higher when the positive (versus the negative) category is presented first, but within 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[2]). 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 that response differences might also occur across modes when the scale is presented visually in one mode (i.e. web) and not at all in another mode (i.e. 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

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

"top of the head" answers. We discuss six comparisons where telephone respondents are asked 5-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 2e). Interviewers were instructed to read the instruction slowly, pausing between each number. We also include the results from 2 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 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 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 5-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 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 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.

DISCUSSION 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 74 experimental comparisons indicate that survey mode and scalar format independently influence

respondents' answers (see Figure 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 5-category scales (22 of 26 difference of means comparisons significant) and to 11-category scales (3 of 4 mean tests significant). This mode effect finding appears quite robust as telephone respondents provide more positive ratings to both 5 and 11-point scales, regardless of whether all or only the endpoint categories are labeled, and across various types of substantive scales (i.e. 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 "agree." However, respondents to construct-specific scales are more likely to select the "very" and "somewhat" positive categories. 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 (i.e. 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. Since our experimental comparisons were tested using students, a population with typically higher levels of education and younger in age (18-25), 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 well 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 questionnaire), suggesting that visually presenting 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 (i.e. satisfied or dissatisfied) and another question asking respondents to indicate the strength or intensity of their attitude (i.e. 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 same *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 2006. "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*.

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." Draft Paper?

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 Groves, R., D. Dillman, J. Eltinge, and R. Little (eds.), *Survey Nonresponse*. New York: John Wiley.

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

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.

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 2006. "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.

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

	Telephone example	**Web example**
2a. Fully labeled vs. Polar point labeled scales *6 telephone comparisons* *6 web comparisons* 3 satisfied/dissatisfied 3 agree/disagree	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 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? Very Satisfied Somewhat Satisfied Neutral Somewhat Dissatisfied Very 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
2b. 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

	Telephone example	Web example
2c. 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.	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
	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? 1 Very Satisfied 2 3 4 5 Very Dissatisfied
2d. 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.	How likely are you to continue to attend WSU until you finish your degree? 5 Extremely likely 4 3 2 1 Not at all likely
	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? 1 Not at all likely 2 3 4 5 Extremely likely

2e. <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 (beginning with 5 vs. 1) *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.
2f. <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 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	8 of 9 comparisons significant	5 of 9 comparisons significant
Polar point 5-point scales	11 of 13 comparisons significant	7 of 13 comparisons significant
Polar point 11-point scales	3 of 4 comparisons significant	No significant differences (4 comparisons)
Scale effects		
Fully labeled vs. polar point	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	No consistent pattern 2 of 3 comparisons significant on both the telephone and web	
Most positive category 5 vs. 1	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 (2 comparisons)	
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)	(web only) More positive ratings when 5 (vs. 1) is positive - 2 of 2 comparisons significant	

Table 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		$5+ \chi^2$ 2*2 df=1	
		% of R selecting category …																	
Q	**Scale type**	**5+[1]**	**4**	**3**	**2**	1	**n**	**Mean**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**		**p**		**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	1.0		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

Table 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 t-test		5+ χ^2 2*2 df=1	
		% of R selecting category …																	
Q	**Scale type**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**	**5+**	**4**	**3**	**2**	**1-**	**n**	**Mean**		**p**		**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 possible	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		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 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

		% of respondents selecting category …													Diff. Means		5+ χ^2 2*2	
Q	**Format**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**n**	**Mean**	**t-test**	**p**	df=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

Table 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		χ^2 2*2 df=1	
			% of R selecting category …							% of R selecting category …										
	Q	Scale type	5+	4	3	2	1-	n	Mean	5+	4	3	2	1-	n	Mean		p		p
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

Table 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		χ^2 2*2 df=1	
		% of R selecting category …							% of R selecting category …										
	Q	5+	4	3	2	1-	n	Mean	5+	4	3	2	1-	n	Mean		p		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

Table 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-							Polar point 1+ 5-							Diff. Means t-test	χ^2 2*2		
			% of R selecting category …							% of R selecting category …										
	Q	Scale type	5+	4	3	2	1-	n	Mean	1+	2	3	4	5-	n	Mean		p	df=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

Table 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 t-test	χ^2 2*2		
		% of R selecting category …							% of R selecting category …										
Q	Scale type	5+	4	3	2	1-	n	Mean	1+	2	3	4	5-	n	Mean		p	df=1	p
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

Appendix A: List of survey questions with topic

Question	Scale Labels	Topic
2	Satisfied/Dissatisfied	Washington State University (WSU) as a place to go to school
4	Satisfied/Dissatisfied	Quality of Instruction in the classes taken at WSU
5	Agree/Disagree and Construct-specific	Accessibility of instructors outside of class
6	Satisfied/Dissatisfied	Satisfaction with classes taking this semester
7	Satisfied/Dissatisfied	Satisfaction with the quality of advising received as a WSU student
9	Worst/best possible	Worst to best possible advisor
16	Satisfied/Dissatisfied	Satisfaction with decision to attend WSU
17	Worst/best possible	Quality of education at WSU
18	Extremely/ Not at all likely	To continue to attend WSU until you finish your degree
19	Extremely/ Not at all likely	To recommend WSU to your friends and family
21	Agree/Disagree and Construct-specific	Cougar varsity sports central to campus life
24	Satisfied/Dissatisfied	Satisfaction with housing living in while attending WSU
25	Agree/Disagree and Construct-specific	Pullman is a desirable place to live while going to school