Public Opinion Quarterly, Vol. 75, No. 2, Summer 2011, pp. 349–365

RESEARCH SYNTHESIS
MORE POSITIVE OR MORE EXTREME? A META-ANALYSIS OF MODE DIFFERENCES IN RESPONSE CHOICE

CONG YE*
JENNA FULTON
ROGER TOURANGEAU

Abstract Some researchers have argued that respondents give more extreme answers to questions involving response scales over the telephone than in other modes of data collection, but others have argued that telephone respondents give more *positive* answers. We conducted a meta-analysis based on 18 experimental comparisons between telephone interviews and another mode of data collection. Our analysis showed that telephone respondents are significantly more likely than respondents in other modes to give extremely positive answers (for example, the highest satisfaction ratings in a customer satisfaction survey) but are not more likely to give extremely negative responses. This tendency to give highly positive ratings appears to be related to the presence of an interviewer, and it may reflect respondents' reluctance to express bad news, a tendency some social psychologists have dubbed the MUM effect.

Introduction

Many surveys use multiple modes of data collection in an attempt to reduce data-collection costs or boost response rates. This has led to increased concern about the comparability of the data collected through different modes (see, for example, de Leeuw 2005). Mode differences in responses are often observed in the literature. Mode comparisons show that question-order effects (Schwarz

Cong Ye and Jenna Fulton are graduate students at the Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA; Roger Tourangeau is a Research Professor at the Institute for Social Research at the University of Michigan, Ann Arbor, MI, USA, and the Director of the Joint Program in Survey Methodology at the University of Maryland, College Park, MD, USA. The authors are grateful for the helpful comments from the reviewers and the editors. *Address correspondence to Cong Ye, Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, MD 20742, USA; e-mail: cye@survey.umd.edu.

doi: 10.1093/poq/nfr009

1996), social-desirability biases (Tourangeau and Smith 1996; Tourangeau and Yan 2007), response-order effects (Krosnick and Alwin 1987), and satisficing behavior (Krosnick 1991) may all differ depending on the mode of data collection. For example, Tourangeau and Yan (2007) review evidence that modes of data collection in which the computer administers the questions directly to the respondent (such as audio computer-assisted self-interviewing, or audio-CASI) reduce socially desirable responding relative to modes of data collection featuring interviewer administration of the questions (such as computer-assisted personal interviewing, or CAPI). Similarly, Krosnick (1991) has shown that modes of data collection involving aural administration of the items (such as telephone surveys) foster recency effects (in which the last response option in the list of possible response categories is selected more often), whereas visual administration (such as a self-administered paper questionnaire) fosters primacy effects (in which the first option is selected more often).

Several researchers have suggested that there may be an additional mode difference in reporting that should be taken into account in planning mixed-mode surveys: Respondents are thought to be more likely to choose the most extreme response option in telephone surveys than in other modes of data collection. For example, in an early study, Groves and Kahn (1979) reported that "we have already noticed some tendency for the telephone respondents to choose the extreme satisfaction point" (p. 199). Similarly, Tarnai and Dillman (1992) noted that "telephone respondents more frequently selected the extreme response category than did self-administered respondents" (p. 124). More recently, Christian, Dillman, and Smyth (2008) and Dillman et al. (2009) have made similar claims.

Most of these prior investigations point to one or two studies demonstrating these mode differences in the tendency to select the extreme response options. The purpose of this article is to examine the evidence more systematically and to determine whether there is a consistent difference across modes of data-collection in the likelihood that respondents will select the extreme response options. In addition, it is not clear whether any mode difference involves a preference for the most extreme options at either end of the response scale or only the most extreme *positive* option. It is possible that telephone respondents display a positivity bias—that is, a preference for the most positive option—rather than an extremity bias. Tourangeau, Rips, and Rasinski (2000, pp. 240–41) describe this tendency to avoid giving others negative evaluations as a positivity or leniency bias. It has been demonstrated in employee performance appraisals (Landy and Farr 1980) and in course evaluations (Sears 1983).

Any heightened preference for the extreme response options (or at least the extreme positive options), if it exists, presumably reflects the characteristics of telephone interviews. One obvious difference between telephone and self-administered surveys is the presence of an interviewer. When an interviewer is involved, the interview becomes a social encounter. As a result, conversational norms or other social conventions may affect the response process. For example,

Rosen and Tesser (1970) argue that people are generally reluctant to communicate information that is likely to be undesirable for the recipient, dubbing this tendency the "Mum about Undesirable Messages (MUM) effect." If respondents exhibit a similar reluctance to give negative evaluations in interviewer-administered surveys, this may explain their heightened tendency to select the most positive answer in a telephone survey compared to a mail survey. According to Tesser and Rosen (1972), people are reluctant to communicate undesirable information for a variety of reasons—because they fear the recipient's emotional reaction (Rosen and Tesser 1970) or they feel guilty about not sharing the recipient's emotional state (Tesser and Rosen 1972). Weenig, Groenenboom, and Wilke (2001) suggest that this reluctance is greater when the recipients of the bad news are strangers (as in a telephone interview) than when they are close friends.

The MUM effect applies to situations in which the message is undesirable for the *receiver*. This is quite distinct from the concept of social-desirability bias, which refers to respondents' reluctance to reveal embarrassing information about themselves. In a survey context, the MUM effect means that respondents in a customer satisfaction survey may avoid giving negative evaluations of their local bank branch or of a fast-food restaurant they visited recently (see Tourangeau, Steiger, and Wilson 2002). By contrast, social-desirability bias refers to such survey behaviors as underreporting illicit drug use or overreporting having voted. Survey respondents have been shown to overreport socially desirable attitudes or behaviors and to underreport socially undesirable ones, and the presence of the interviewer magnifies this tendency (see DeMaio 1984 for an early review and Tourangeau and Yan 2007 for a more recent one). Tourangeau and Yan (2007) argue that social-desirability bias arises because respondents want to avoid embarrassing themselves. It seems implausible that respondents would find it embarrassing to complain about their banking experience or their most recent visit to a fast-food chain. Thus, it seems unlikely that this difference in socially desirable reporting by mode could account for any mode differences in the preference for extreme response options.

Still another difference between telephone interviews and other modes of data collection that may affect the propensity to select the most extreme response options is the communication channel. Telephone surveys rely on oral transmission of the questions, whereas other modes involve visual presentation; in face-to-face surveys, both methods of transmission may be involved, since show cards are often used in addition to having interviewers read the questions (and response options) aloud.[1] This difference in the channel of communication

1. Despite concerns about the declining importance of CATI, it continues to be the main data-collection mode for several major federal surveys, such as the Current Population Survey (CPS), the Behavioral Risk Factor Surveillance System Annual Survey, and the National Immunization Survey. (The CPS is a mixed-mode survey but, except for the first and fifth interviews, the bulk of the data are collected by telephone.)

is thought to explain why primacy effects predominate when questions are presented visually and recency effects predominate when questions are presented aurally. When the options are presented visually, the respondent is more likely to select one of the earlier response options because these options are likely to be subjected to deeper cognitive processing than the later ones (Galesic et al. 2008). When the options are presented aurally, respondents often do not have the opportunity to process the earlier options and are likely to begin processing the last option they hear (Holbrook et al. 2007; Krosnick and Alwin 1987). Because the most extreme option is the last to be presented to respondents in telephone surveys, this tendency to select the last option may lead to more extreme answers being chosen. In contrast, respondents in surveys that present information visually usually have more time for cognitive processing and can review the question as well as response options before selecting an answer; thus, the tendency may not be as strong.

Finally, the format of the response options often differs between telephone surveys and other modes, and this may contribute to any preference for selecting the most extreme response option. In telephone surveys, it is common to provide only the endpoints of the scale to the respondent ("on a scale ranging from 1 to 10. . ."); however, in visual modes, fully labeled scales are usually presented. Studies have found that the fully labeled scales result in more positive ratings than the endpoint-labeled scales (Krosnick and Fabrigar 1997; Schwarz and Hippler 1991), which may imply less extreme negative responses in the fully labeled scales and more extreme positive responses in the endpoint-labeled scales. When they do offer fully labeled scales, telephone surveys are more likely to use "unfolding" questions that present the answer categories in two parts (Groves and Kahn 1979; Krosnick and Berent 1993). For example, Sykes and Collins (1988) asked whether the respondents "agree strongly, agree, disagree, or disagree strongly" in a face-to-face survey, but asked half of the respondents in a telephone sample this same question in two stages. They first asked this group of telephone respondents whether they agreed or disagreed, and then asked them whether they agreed (or disagreed) strongly. This unfolding procedure may make the extreme options appear less extreme and may thus increase any preference for the extreme response categories in telephone surveys.

We present the results of a meta-analysis that examines a number of mode-comparison studies. Each of the studies compares results for items using scales from telephone surveys with similar items administered in some other mode or modes of data collection. We had three main goals for our study. First, we hoped to estimate the size of the extremity bias across studies; second, we hoped to clarify the nature of the effect (that is, whether the preference is for the scale endpoints or only for the positive endpoint); and finally, we hoped to identify the psychological basis for this mode difference.

Methods

Retrieval and coding of the papers. We searched for papers to include in our meta-analysis using the University of Maryland library online research portal and other search engines, such as Google Scholar; our keywords included "mode comparison," "mode experiment," "extremity bias," "mode effects," "telephone survey," "CATI survey," and "measurement error." Some of the papers we identified in this way cited other relevant papers; we included these in our analysis as well.

Of the studies we found through our literature search, we narrowed our attention to those that included CATI in the mode comparison and that had at least some questions in which the response options were rating scales. Our analysis is based on those studies for which the necessary data (specifically, the sample size and the percent responding in the most extreme response categories) either were given in the original paper or were made available to us by the authors after contacting them. In total, we located 12 studies that met these criteria, several of which reported customer satisfaction ratings (see table 1). Because multiple papers include more than one mode comparison, our meta-analysis includes a total of 18 comparisons of CATI with other modes: Two compare CATI with a Web survey; seven compare CATI with a mail survey; six compare CATI with face-to-face interviews; two compare CATI with interactive voice response (IVR); and one compares a conventional CATI interview with a CATI interview in which the respondents had a paper copy of the questionnaire. The analysis included a total of 199 items across the studies.

The studies we incorporated were conducted between 1967 and 2009 and covered a variety of subjects. Many of the studies collected satisfaction ratings of some sort; these ratings were collected from a range of target populations. For example, we included studies in which samples of college students rated their satisfaction with their campus facilities, consumers rated their satisfaction with various goods and services, and residents rated their satisfaction with their communities. The meta-analysis also includes studies involving ratings of health and happiness, neighborhood safety, symptoms of depression, and driving behavior. Only one study in our analysis includes questions with agree/disagree response options. The items we examined seem unlikely to be prone to social-desirability effects; most of the items we include elicit satisfaction ratings rather than information about the types of sensitive behaviors (e.g., drug use, voting, and sexual behavior) that the social-desirability literature focuses on.

We coded the responses as positive or negative based on the meaning of the responses, rather than the exact wording of the response categories themselves. For example, we defined "extreme positive" responses as those involving the most positive response category even when the response option read "very strongly agree." Extreme negative responses were defined similarly, based on the meaning of the response option. For each study, we calculated the proportion of respondents selecting the most extreme positive and negative

Table 1. Selected Study Characteristics, by Study

Authors	Year	Mode Compared with CATI	Subject
Aneshensel et al.	1982	FTF	Depression scale
Chan et al.	2004	Mail	Depression scale
Christian, Dillman, and Smyth	2008	Web	Student satisfaction with Washington State University
de Leeuw	1992	Mail, FTF	Satisfaction on five well-being questions
Dillman and Mason	1984	Mail, FTF	Satisfaction with neighborhood
Dillman et al.	2009	IVR, Web, Mail	Satisfaction with long-distance provider
Dillman and Tarnai	1991	Mail	Driving behavior
Groves	1979	FTF	Happiness and life satisfaction
Hochstim	1967	Mail, In-person	Health self-rating
Klecka and Tuchfarber	1978	FTF	Safety of neighborhood
Tarnai and Dillman	1992	Mail*	Student satisfaction with college facilities
Tourangeau, Steiger, and Wilson	2002	IVR	Customer satisfaction

*This study includes mode comparisons between CATI and mail, and ordinary CATI with a version of CATI in which the respondents also had a paper copy of the questionnaire.

response categories for each of the survey questions for each mode group. (For two of the comparisons, only information on the extreme positive responses was available; no information on negative responses was included for those comparisons.) When the survey included multiple questions using response scales, we computed the average proportion of extreme positive and extreme negative responses across the questions by mode.

Quantifying the mode effect. We calculated an effect size measure that allowed us to aggregate the mode comparisons across studies. Since the outcome measures are dichotomous (the respondent selected the extreme option or not), the effect size measure we used was the odds ratio (OR) for the mode effect within an individual study (Fleiss 1994); the ORs were based on the average proportion of extreme responses. An OR represents the odds of the event occurring in one condition (e.g., selected the extreme option in a mail survey) divided by the odds of the event occurring in the other condition (e.g., selected

the extreme option in a CATI survey). If there is no difference between the two conditions, the OR is equal to 1. If responding through the mail survey reduces the proportion of respondents endorsing the extreme option relative to responding by CATI, the OR is less than 1; if responding via mail increases the proportion of respondents endorsing the extreme option relative to responding by CATI, the OR is greater than 1. We calculated two odds ratios for each mode comparison—one for the average proportion of extreme positive responses and the other for the average proportion of extreme negative responses. Because CATI was included in each mode comparison and was the mode of primary interest, it was used as the reference group in each OR:

$$OR = \frac{p_{other}(1 - p_{CATI})}{p_{CATI}(1 - p_{other})},$$

in which p_{CATI} is the average proportion of extreme responses in CATI and p_{other} is the average proportion of extreme responses in the other mode of data collection.

Logged odds ratios (LORs) were used to perform all analyses, mainly because they simplify standard error equations relative to ORs and are centered around 0 (whereas the ORs are centered around 1). All results were then transformed back to ORs for ease of interpretation. Before aggregating the results across studies, we first calculated a weight for each LOR. The weight was the inverse of the variance of the difference between modes (that is, we summed the variances across CATI and the other mode and assumed a zero covariance term). We took the natural logarithm of each OR and applied the weight, and then used these LORs to calculate an overall (weighted) effect size, following methodology described by Lipsey and Wilson (2001):

$$\begin{aligned} w_{LOR} &= 1/se^2_{LOR} \\ &= 1/\Big(\frac{1}{p_{CATI}n_{CATI}} + \frac{1}{(1 - p_{CATI})n_{CATI}} + \frac{1}{p_{other}n_{other}} \\ &\quad + \frac{1}{(1 - p_{other})n_{other}}\Big) \end{aligned}$$

$$ES_{LOR} = \frac{\sum_{i=1}^{l} w_i LOR}{\sum_{i=1}^{l} w_i}$$

where ES_{LOR} is the estimated overall effect size and n refers to the sample sizes for the CATI group and the group administered the other mode. Studies were treated as clusters, meaning that multiple comparisons within one study belong to the same cluster. This analysis was carried out using SAS survey procedures that account for this clustering.

Moderator analysis. To determine whether other study characteristics (for example, the order of the response options) affected the relationship between

endorsing an extreme response option and interviewing mode, we performed a series of moderator analyses. We gathered the following information for each study: whether response scales presented positive or negative response options first;[2] whether all scale points or only the endpoints were labeled;[3] and whether the questions asked about the respondents themselves or other individuals. This analysis was carried out using a random effects model in the commercial software Comprehensive Meta-Analysis.[4] We attempted to examine whether presenting a questionnaire or show card to respondents affects the relationship; however, having only five eligible studies in total (two studies in one group and three in the other) prevented us from including this analysis.

Results

Our analysis includes a total of 18 comparisons between CATI and another mode, derived from 12 studies. An additional study by Sykes and Collins (1988) compared data from face-to-face and CATI interviews and met each of our inclusion criteria except one—it did not include the sample sizes associated for each of the relevant questions. In order to include the Sykes and Collins results, we used the total sample sizes given for each mode in their paper, assuming there were no missing data for the specific questions we examined.

We first examined the issue of heterogeneity across studies—that is, whether the studies yield results that are more variable than would be expected based on sampling error alone. We calculated Cochran's Q-statistic (Cochran 1954), which assesses the level of heterogeneity, for the studies included in our analysis:

$$Q = \sum_{i=1}^{k} w_i (LOR_i - ES_{LOR})^2,$$

in which k is the number of studies in the analysis and the other terms are the weight (w_i) and effect size (LOR_i) for study i and ES_{LOR} is the weighted average of the effect sizes across all the studies, as defined previously. We calculated Q-statistics for the effect sizes on both the extreme negative responses and the extreme positive responses. For the extreme negative responses, we find no evidence of heterogeneity ($Q = 0.76$, n.s.). For the extreme positive responses, we do not find evidence of heterogeneity either ($Q = 16.8$, n.s.). As we demonstrate below, the tendency to give extreme positive responses does seem to vary across the different modes of data collection to which CATI is compared in these studies.

2. Items in some studies included both formats and therefore inflated the total sample size for these analyses.

3. In our analysis, there is only one study (Groves 1979), or two if including Sykes and Collins (1988), that included unfolding items, and thus we did not test the unfolding effect.

4. We also examined these differences in Stata, accounting for clustering, which provided the same results.

As we noted earlier, each comparison is based on the average proportion of extreme responses. For example, in Groves (1979), the average proportion of the CATI respondents selecting the extreme negative and extreme positive response options for the five target items were 1.0 percent and 13.9 percent, respectively, while the average proportions among respondents interviewed in person were 1.6 percent and 12.6 percent, respectively. The sample size for each comparison was calculated as the average sample size across the included items.

Overall, the weighted LOR for respondents choosing extreme *negative* responses is 0.001, corresponding to an OR of 1, which indicates that there is no difference between CATI and other modes. CATI respondents and respondents in other modes were overall equally likely to select the most negative response option. For the proportion choosing the extreme *positive* response, the weighted LOR is -0.21, corresponding to an OR of 0.81; this LOR has a t-value of -2.33 and is significant ($p < 0.05$). These results thus show stronger evidence for a positivity bias in CATI than for an overall extremity bias.[5]

Although there are only a few studies available for each alternative mode, comparing CATI responses with surveys administered in person, by mail, and by Web is still useful in understanding *why* there may be a mode difference in the tendency to select extreme positive responses. For example, some of the mode comparisons compare CATI to modes that do not involve an interviewer (mail, Web), while others do involve interviewers (face-to-face).

We found seven comparisons between mail and CATI surveys for our analysis. Using CATI as the reference mode, the average OR for selecting extreme *negative* responses is 1.02 ($t = 0.35$), which does not differ significantly from 1. In contrast, the OR for choosing extreme *positive* responses is 0.61 and is highly significant ($t = -7.17$, $p < .001$; see table 2). Thus, the comparisons between mail and CATI surveys confirm that there is a *positivity* effect in CATI surveys compared to mail surveys, but not an extremity effect. Six additional comparisons involve CATI and face-to-face interviews. Both of these modes of administration involve an interviewer, but the two modes may differ in how strongly they foster a sense of the interviewer's social presence. The physical presence of an interviewer in the face-to-face setting may create a stronger sense of social presence than a CATI interview does (see Groves 1989, p. 508, and Holbrook, Green, and Krosnick 2003 for somewhat differing views on this issue).

5. As we noted earlier, Sykes and Collins (1988) do not provide sample sizes for each item; in addition, this study uses a quasi-experimental mode comparison in which the same questions were administered to one sample by mail and to a different sample by telephone. If we nonetheless include these results in our analyses, the OR for the proportion of respondents choosing the most extreme negative responses in other modes relative to the proportion choosing that option in CATI is 0.91 ($t = -0.99$, n.s.), and the OR for respondents choosing the most extreme positive responses is 0.80 ($t = -2.54$, $p < 0.05$). Therefore, including the findings from this study does not change our overall conclusions.

Table 2. Summary Statistics from the Meta-Analysis, by Mode Group

Comparison	Response Option (*n*)	Log OR	Odds Ratio	95% CI-	95% CI+	*t*-value
Mail	Negative End (6)	0.02	1.02	0.92	1.12	0.35
	Positive End (7)	−0.50	0.61	0.52	0.71	−7.17**
Face-to-face	Negative End (6)	−0.04	0.96	0.74	1.23	−0.38
	Positive End (6)	0.02	1.02	0.93	1.11	0.49
All studies, except for face-to-face	Negative End (10)	0.08	1.09	0.98	1.20	1.98
	Positive End (12)	−0.45	0.64	0.59	0.69	−12.81**
All studies	Negative End (16)	0.001	1.00	0.81	1.23	0.01
	Positive End (18)	−0.21	0.81	0.66	0.99	−2.33*

NOTE.—MCATI is used as the reference category in computing all odds ratios. *$p < 0.05$; **$p < 0.001$.

As shown in table 2, the OR for face-to-face respondents choosing extreme *negative* responses relative to CATI is 0.96 (compared to 1.02 in the mail vs. CATI comparison); on the other hand, the OR for respondents choosing extreme *positive* responses is 1.02 (compared to 0.61 in the mail vs. CATI comparison). While neither OR is significant, they are in reverse directions and seemingly tell a different story from the comparison between CATI and mail surveys. In particular, the positivity bias in CATI interviews relative to mail surveys disappears in the comparison of face-to-face and CATI interviews.

Besides the seven studies comparing results from CATI and mail surveys and the six comparing CATI and face-to-face surveys, we found two studies comparing CATI and IVR, two comparing CATI and Web surveys, and one comparing CATI with a telephone interview in which respondents were also given a paper copy of the questionnaire. The conclusions from these studies are in the same overall direction as the comparisons between the CATI and mail surveys mentioned above. When we include all of these studies, except those involving face-to-face interviews, we have a total of 12 comparisons. A common factor across these 12 surveys is that they lack interviewer involvement. With CATI as the reference group, the OR for selecting extreme negative responses is 1.09 ($t = 1.98$, $p < 0.09$) across the 12 comparisons and the OR for selecting extreme positive responses is 0.64 ($t = -12.81$, $p < .001$).

In summary, the available studies show strong evidence of a differential positivity bias by mode.[6] Respondents are more likely to endorse the most extreme positive response category in CATI than in mail, IVR, or Web surveys. In

6. There is significant heterogeneity in the effect size values for studies comparing CATI and mail surveys in the proportion of extreme positive responses ($Q = 27.3$, $p < 0.001$). More specifically, the positivity bias in the telephone mode is considerably higher in the Dillman et al. (2009) study than in the other studies that compare CATI and mail responses.

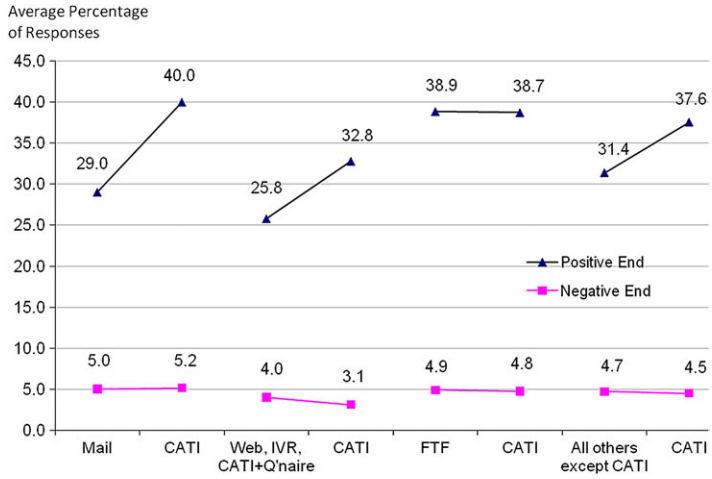

Figure 1. Average Proportion of Extreme Responses by Mode Comparison.

addition, they are more likely to select the extreme positive option in face-to-face interviews than in CATI. Respondents in face-to-face surveys are also prone to avoid extreme negative responses relative to CATI respondents. This overall pattern of results is illustrated in figure 1. The findings are similar if we include the results from Sykes and Collins (1988).

We note that these analyses do not include any face-to-face studies conducted after 1992. To determine whether this could affect the overall results, we recalculated the effect sizes comparing mail and CATI that occurred in or prior to 1992 (no studies carried out during this period involved IVR or Web). As a result, two of the seven studies were excluded. However, with the remaining five comparisons, we arrived at the same conclusions as before. Using CATI as the reference mode, the average OR for selecting extreme *negative* responses is 1.01 ($t = 0.11$, n.s.), compared to 1.02 ($t = 0.35$, n.s.) with all seven comparisons. The OR for choosing extreme *positive* responses is 0.65 ($t = -9.87$, $p < .001$), compared to 0.61 ($t = -7.17$, $p < .001$) with all studies. Thus, the comparisons between mail and telephone surveys conducted in or prior to 1992 also confirm that there is a *positivity* effect in CATI surveys, but not necessarily an extremity effect.

We graphed the ORs of telephone responses and responses in other modes by year (figure 2), which demonstrates that there are no obvious trends over time. The seemingly declining line for responses on the positive end reflects the fact that all face-to-face and CATI comparisons were conducted in or prior to 1992. Therefore, these studies are clustered to the left. The data points toward the right represent other modes and exclude face-to-face and CATI comparisons.

The moderator analyses demonstrate that very few of the study characteristics are significantly correlated with the observed positivity effect. As shown in table 3, there are no significant differences in ORs between fully labeled

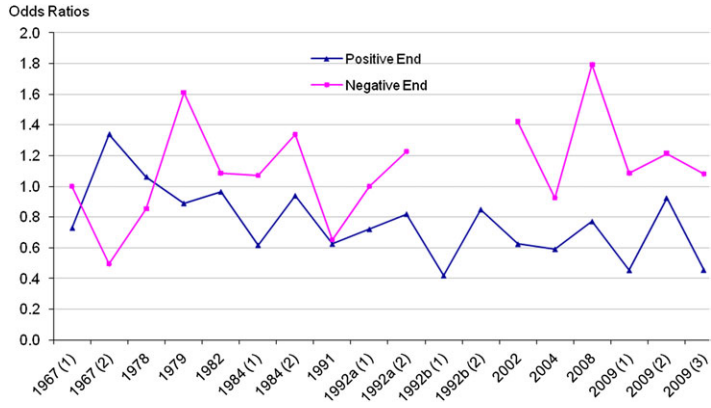

Figure 2. Odds Ratios for Extreme Responses in CATI versus Other Modes by Year. (The numbers in parentheses after the year of the study refer to the different comparisons within a single study.)

scales and scales in which only the endpoints were labeled, and there are no effects on ORs for the negative end of the scale (regardless of whether positive or negative response options were presented first).

The tendency for CATI respondents to select the most positive response option more often was most obvious when the most positive option was presented last, although this effect is only marginally significant ($Q = 3.01$, $p = 0.08$). This provides evidence for a mild recency effect in CATI surveys for the most positive response option. Overall, we observed this positivity effect regardless of response option order.

In addition, we classified items as to whether or not they asked about the respondents themselves or about someone else, and tested for potential differences. We found that the positivity effect was somewhat stronger if the questions did not ask about the respondents themselves, but this difference was not significant. This provides evidence for the MUM effect because if the underlying mechanism is social desirability, then we would expect to observe more positivity in CATI surveys when the question asks about respondents themselves.

Discussion

We carried out a meta-analysis of the results from 18 mode experiments. Overall, we find that telephone respondents are more likely to select the most extreme positive option than respondents to Web, mail, or IVR surveys but not respondents to face-to-face interviews. Although most of the studies we examined were conducted using CATI, we think that the important variable is the presence of a telephone interviewer rather than administration by computer. We find no evidence for a mode difference in preference for the extreme

Table 3. Results of Moderator Analysis

	Positive End				Negative End			
	N	Odds Ratio	95% CI−	95% CI+	*N*	Odds Ratio	95% CI−	95% CI+
Label								
Full	17	0.79	0.66	0.93	15	1.07	0.86	1.32
Endpoints only	10	0.67	0.55	0.82	8	1.21	1.02	1.44
			$Q = 1.40$	$p = 0.24$			$Q = 0.84$	$p = 0.36$
Response option order								
Negative first	10	0.67	0.56	0.80	8	1.16	0.89	1.51
Positive first	12	0.83	0.71	0.97	10	1.03	0.89	1.21
			$Q = 3.01$	$p = 0.08$			$Q = 0.55$	$p = 0.46$
Subject of questions								
Respondents	6	0.83	0.65	1.05	6	0.96	0.68	1.36
Others	12	0.71	0.58	0.86	10	1.01	0.89	1.15
			$Q = 0.95$	$p = 0.33$			$Q = 0.08$	$p = 0.77$

negative response options. There are at least three possible explanations for the differential positivity bias by mode that we find in our meta-analysis.

First, it is possible that respondents prefer to keep "Mum about Undesirable Messages," as Rosen and Tesser have argued (Rosen and Tesser 1970; Tesser and Rosen 1972). That is, they are more reluctant to communicate bad news to an interviewer—in this case, negative evaluative ratings—than good news. The MUM effect seems especially strong in situations in which the messenger and recipient are mere acquaintances or strangers; people are more willing to transmit messages that convey negative consequences to close friends (Weenig, Groenenboom, and Wilke 2001; Tesser and Rosen 1972). Since the interviewer is likely to be a stranger to the respondent, the respondent may be especially unlikely to impart negative evaluations in an interviewer-administered survey than in a self-administered one (for a similar argument, see Tourangeau, Steiger, and Wilson 2002).

The overall pattern of a higher proportion of extreme positive responses in telephone interviews was not observed in the comparison with face-to-face interviews. This finding is also more or less consistent with the view that the positivity bias is a product of the MUM effect, since the presence of a stranger—the interviewer—is presumably more salient to the respondent in a face-to-face interview than in a telephone interview. However, while the strength of the relationship between the messenger and the recipient directly affects the likelihood of expressing bad news, it does not seem to affect the willingness to share good news (Weenig, Groenenboom, and Wilke 2001). As figure 1 demonstrates, most of the mode effects we found involve an increased likelihood of giving answers at the extreme positive end.

A related account is based on the notion of impression management (e.g., Schlenker and Weingold 1989). People may form more favorable impressions of people who are themselves more positive about others; thus, to help make a favorable impression on the interviewer, respondents may give highly positive responses when an interviewer is present (as in a telephone or face-to-face survey), as opposed to when an interviewer is not present (as with a mail or Web survey).

A second possible account for the findings is more cognitive. Modes of data collection that present the questions aurally may impose greater burdens on a respondent's working memory than do modes that present the questions visually. Because of this increased burden on working memory, respondents may be more susceptible to response-order effects in, say, a telephone interview than they are in a mail survey. In telephone surveys (and in other modes of data collection that administer the questions aurally), response-order effects typically take the form of recency effects—that is, a heightened probability of selecting the last option presented (e.g., Holbrook et al. 2007; Krosnick and Alwin 1987). Thus, if the most positive answer was also the last answer option read to the respondent, this could explain many of our findings.

There are a couple of reasons that we reject this line of argument. First, Tarnai and Dillman (1992) found that providing a paper copy of the questionnaire to telephone respondents did not eliminate the differences between telephone and

mail surveys. Similarly, face-to-face respondents who were provided with a paper questionnaire were still more likely than mail respondents to select the most positive response options (Dillman and Mason 1984). The paper questionnaire should have eliminated the mode differences in working memory burden. Thus, the key variable affecting whether respondents exhibit a positivity bias seems to be the presence of an interviewer rather than the absence of a paper questionnaire. Second, in the telephone studies we analyzed, the most positive response was not always the last option presented. Therefore, there was not any systematic relationship between the position of the most positive response option and the mode difference in the proportion of respondents selecting that option. Although the positivity bias in telephone surveys is heightened when the most positive option is the last one, it is still apparent when the most positive option is the first one presented (see table 3).

A third potential explanation for our results is that question format can cause acquiescence or yea-saying, thus leading to the apparent positivity bias when the survey involves interviewers. However, our results do not provide evidence supporting this. Only one study in our analysis includes questions with agree/disagree response options that would be subject to acquiescence or yea-saying; we do not see much potential for an acquiescence effect in the other questions included in our analysis, most of which ask for satisfaction ratings.

Most of the findings on which our conclusions are based involve evaluative scale ratings (such as ratings of customer satisfaction). It is not clear what other types of items might show similar effects. The best account for these findings seems to be one based on the MUM effect. Respondents seem to exhibit the same general reluctance originally found by Tesser and Rosen about transmitting bad news to a stranger and, as a result, they choose highly positive answers when a stranger administers the questions. Thus, one possible approach for reducing this preference for extreme positive answers might be to rely on forms of self-administration such as mail or IVR. It is possible that wording changes or other strategies might minimize this tendency to select the most positive answer option. In the original demonstrations of the MUM effect, it is not clear whether the participants were avoiding giving bad news or just preferred giving good news. Thus, it may difficult to disentangle the MUM effect from the impression management account, which emphasizes the self-presentation advantages of giving positive feedback to others. Whatever the right theoretical account, our findings indicate that survey respondents are more prone to give extreme positive ratings than they are to avoid extreme negative ones. We do not find evidence for systematic mode difference in the selection of the most extreme negative options, only for the most extreme positive options. At the same time, as figure 1 plainly demonstrates, respondents (regardless of the mode of data collection) are often quite reluctant to give negative evaluations. This tendency to put a positive spin on things in interviews may reflect a habitual self-presentation strategy carried over from everyday life.

References

Aneshensel, Carol S., Ralph R. Frerichs, Virginia A. Clark, and Patricia A. Yokopenic. 1982. "Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews." *Public Opinion Quarterly* 46:110–21.

Chan, Kitty S., Maria Orlando, Bonnie Ghosh-Dastidar, Naihua Duan, and Cathy D. Sherbourne. 2004. "The Interview Mode Effect on the Center for Epidemiological Studies Depression (CES-D) Scale: An Item-Response-Theory Analysis." *Medical Care* 42(3):281–89.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. 2008. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In *Advances in Telephone Survey Methodology*, eds. James M. Lepkowski, N. Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, 250–75 New York: Wiley-Interscience.

Cochran, Williams G. 1954. "The Combination of Estimates from Different Experiments." *Biometrics* 10:101–29.

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

de Leeuw, Edith D. 2005. "To Mix or Not to Mix Data-Collection Modes in Surveys." *Journal of Official Statistics* 21:233–55.

DeMaio, Theresa J. 1984. "Social Desirability and Survey Measurement: A Review." In *Surveying Subjective Phenomena*, eds. Charles E. Turner and Elizabeth Martin, 257–81, Vol. 2. New York: Russell Sage Foundation

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, Wisconsin.

Dillman, Don A., Glenn Phelps, Robert D. Tortora, Karen Swift, Julie Kohrell, and Jodi Berck. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet." *Social Science Research* 38:1–18.

Dillman, Don A., and John Tarnai. 1991. "Mode Effects of Cognitively Designed Recall Questions: A Comparison of Answers to Telephone and Mail Surveys." In *Measurement Errors in Surveys*, eds. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 73–93 New York: Wiley.

Fleiss, Joseph L. 1994. "Measures of Effect Size for Categorical Data." In *The Handbook of Research Synthesis*, eds. H. Cooper and L.V. Hedges, 245–60 New York: Russell Sage Foundation.

Galesic, Mirta, Roger Tourangeau, Mick P. Couper, and Frederick G. Conrad. 2008. "Eye-Tracking Data: New Insights on Response-Order Effects and Other Cognitive Shortcuts in Survey Responding." *Public Opinion Quarterly* 72:892–913.

Groves, Robert M. 1979. "Actors and Questions in Telephone and Personal Interview Surveys." *Public Opinion Quarterly* 43:190–205.

———. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Groves, Robert M., and Robert L. Kahn. 1979. *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.

Hochstim, Joseph R. 1967. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association* 62:976–89.

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social-Desirability Response Bias." *Public Opinion Quarterly* 67:79–125.

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. "Response-Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71:325–48.

Klecka, William R., and Alfred J. Tuchfarber. 1978. "Random Digit Dialing: A Comparison to Personal Surveys." *Public Opinion Quarterly* 42:105–14.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

Krosnick, Jon A., and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201–19.

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* 37:941–64.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." In *Survey Measurement and Process Quality*, eds. Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, 141–64. New York: Wiley-Interscience.

Landy, Frank J., and James L. Farr. 1980. "Performance Rating." *Psychological Bulletin* 87(1):72–107.

Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis* (Applied Social Research Methods Series, Vol. 49). Thousand Oaks, CA: Sage Publications.

Rosen, Sidney, and Abraham Tesser. 1970. "On Reluctance to Communicate Undesirable Information: The MUM Effect." *Sociometry* 33:253–63.

Schlenker, Barry R., and Michael F. Weingold. 1989. "Goals and the Self-Identification Process: Constructing Desires and Identities." In *Goal Concepts in Personality and Social Psychology*, ed. Lawrence A. Pervin 243–90. Hillsdale NJ: Erlbaum.

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

Schwarz, Norbert, and Hans Hippler. 1991. "Response Alternatives: The Impact of Their Choice and Presentation Order." In *Measurement Errors in Surveys*, eds. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, 41–56 New York: Wiley.

Sears, David O. 1983. "The Person-Positivity Bias." *Journal of Personality and Social Psychology* 44(2):233–50.

Sykes, Wendy, and Martin Collins. 1988. "Effects of Mode of Interview: Experiments in the U.K." In *Telephone Survey Methodology*, eds. Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, James T. Massey, William L. Nicholls II, and Joseph Waksberg, 301–20 New York: Wiley.

Tarnai, John, and Don A. Dillman. 1992. "Questionnaire Context as a Source of Response Differences in Mail versus Telephone Surveys." In *Context Effects in Social and Psychological Research*, eds. Norbert Schwarz and Seymour Sudman, 115–29 New York: Springer Verlag.

Tesser, Abraham, and Sidney Rosen. 1972. "Similarity of Objective Fate as a Determinant of the Reluctance to Transmit Unpleasant Information: The MUM Effect." *Journal of Personality and Social Psychology* 23:46–53.

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

Tourangeau, Roger, and Tom W. Smith. 1996. "Asking Sensitive Questions: The Impact of Data Collection, Mode, Question Format, and Question Context." *Public Opinion Quarterly* 60:275–304.

Tourangeau, Roger, Darby Miller Steiger, and David Wilson. 2002. "Self-Administered Questions by Telephone: Evaluating Interactive Voice Response." *Public Opinion Quarterly* 66:265–78.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–83.

Weenig, Mieneke W. H., Arieneke C. W. J. Groenenboom, and Henk A. M. Wilke. 2001. "Bad News Transmission as a Function of the Definitiveness of Consequences and the Relationship Between Communicator and Recipient." *Journal of Personality and Social Psychology* 80:449–61.