Mode Effect or Question Wording? Measurement Error in Mixed Mode Surveys

Edith de Leeuw[1], Joop Hox[2], Annette Scherpenzeel[3]
[1] Utrecht University, POBOX 80140, NL-3508TC Utrecht, Holland
[2] Utrecht University, POBOX 80140, NL-3508TC Utrecht, Holland
[3]Tilburg University, POBOX 90153, NL-5000LE Tilburg, Holland

Abstract
Members of a high quality, probability-based Internet panel were randomly assigned to one of two modes: (1) computer assisted telephone interview or (2) web survey. Within each mode the same series of split ballot experiments on question format were conducted. We tested the effect of unfolding opinion questions in multiple steps vs. asking a complete question in one-step, and the effect of fully verbal labelling versus end-point labelling of response categories within and between the two modes. We found small direct mode effects, but no interaction. Unfolding (two-step question) had a larger effect. When using a mixed-mode design it is advisable to avoid unfolding formats in the telephone interview and use the complete (one step) question format in both modes. Full labelling is mildly preferred above labelling endpoints only. The absence of an interaction effect is an encouraging result for mixed-mode surveys.

Key Words: Mixed-mode, Labelling, Unfolding, Probability based panel

1. Introduction

We are now at a point in time were telephone surveys have to face many challenges (e.g., growing nonresponse, increasing number of cell phone only households), and where Internet surveys are not yet fully fit to replace them. Mixed-mode surveys are advocated as a solution to the coverage and nonresponse problems facing a single mode approach (e.g. Blyth, 2008). In a mixed-mode survey, survey designers try to combine the best of all possible worlds by exploiting the advantages of different modes to compensate for their weaknesses (de Leeuw, 2005), thereby achieving a good coverage of the intended population and a high response rate at affordable cost. However, when different subgroups of respondents are surveyed with different modes of data collection, mode effects may influence the resulting data and differential measurement error may threaten the validity of the results. One of the main challenges facing us is combining data from telephone and web surveys.

In the past, extensive mode comparisons have been made for the traditional data collection methods: face-to-face interviews, telephone surveys, and self-administered paper mail questionnaires. De Leeuw (1992, Chapter 3) performed a meta-analysis of 67 articles and papers reporting mode comparisons. The resulting overview showed consistent but usually small differences between methods, suggesting a dichotomy of survey modes in those with and without an interviewer. There are fewer comparisons of Internet with interview surveys, either telephone or face-to-face, and as a consequence

there are –as yet– no comprehensive meta-analyses, summarizing mode effects for Internet versus interview surveys. However, one effect is consistently found in the available studies: Internet surveys appear to give rise to less social desirability than interviews. In this sense, Internet surveys are indeed more like self-administered questionnaires and share their benefits as Couper (2008, 28) postulated. For instance, Link and Mokdad (2005) found more self-reported heavy drinkers in a web survey compared to those in a telephone interview. This result remained strong and significant after adjusting for different demographic characteristics of respondents in both modes. Krauter, Presser & Tourangeau (2008) confirmed and extended these findings. Internet administration increased reporting of sensitive information amongst university alumni. Krauter et al (2008) had also access to record data and found a higher accuracy in web surveys; they report that web surveys increased both the level of reporting sensitive information and the accuracy compared to CATI with the more private self-administered telephone survey (IVR) in between.

The studies summarized above were all well-conducted experimental comparison were equivalent questionnaires were used in all modes. However, because of different traditions in questionnaire construction or due to mode specific optimization of questions, researchers often change the question structure when changing the mode of data collection (Dillman and Christian, 2005). From past research (e.g. Rugg, 1941; Schwarz et al, 1985), we know that question wording and response format affect responses even within one mode; for an overview, see Sudman et al (1974). Therefore, changing the question format when changing mode may add to the net effect of mode and result in considerable overall differences between modes in mixed mode surveys (cf. Jackle, Roberts, and Lynn, 2010). To disentangle these effects, we experimentally investigated multiple question formats within and between modes.

A main difference between web and telephone is the channel of communication used to present questions and provide answers. In telephone surveys, the auditory channel is used to convey information, while in web surveys the visual channel is mainly used (cf. De Leeuw, 1992, 2005); this influences the cognitive burden for the respondent (Tourangeau, Rips, and Rasinski, 2000). Aural-only transmission of information places higher demands on memory capacity than visual transmission (Schwarz, Strack, Hippler, and Bishop, 1991); with aural-only the respondents have to remember all information in stead of being able to repeatedly refer to the visual information. Especially when longer lists of response alternatives are presented, the telephone survey is at a disadvantage as all response alternatives have to be processed in the working memory of the respondents. However, in web surveys respondents always have the response alternatives on their screen and in face-to-face interviews visual show cards are often used to relieve the cognitive burden.

To accommodate the limitations of the auditory channel, researchers have designed special question formats for telephone interviews. Response scales are often simplified in telephone interviews by providing only the polar endpoint labels to ease the cognitive and memory burden placed on the respondents. In contrast, in web surveys the visual channel is used and response scales are often presented with all the scale points verbally labelled (Christian, 2007, chap 2). A second method to reduce cognitive burden in telephone interviews is ‘unfolding’ in which scalar questions are branched into two steps, instead of presenting all response categories directly in one step. The advantage of a two-step procedure is that respondents are asked to process a fewer number of categories at one

time, thereby putting less burden on the limited short term memory available when only aural stimuli are processed.

In a series of experiments we tested the effect of both 'endpoint' versus 'full labelling' of response categories and of 'unfolding' vs 'one step direct question' both within telephone and web mode and between telephone and web mode. We compared similar scales across modes and different scales within modes. This enabled us to investigate mode and question format effects independently and study potential interactions between mode and question format. While most question format experiments are usually done with selected or convenience samples, we had the great opportunity to implement these experiments using a random sample of the general population of the Netherlands

2. Method

2.1 Respondents

In this study members of the Dutch LISS panel (Longitudinal Internet Studies for the Social Sciences) were investigated. The LISS panel was established in autumn 2007 and consists of almost 8000 individuals that complete online questionnaires every month. It is based on a probability sample of households drawn from the Dutch population register by Statistics Netherlands. The sample includes people without an Internet connection. All people in the sample were approached in traditional ways (by letter, followed by telephone call and house visit) with an invitation to participate in the panel. Households that could not otherwise participate because they had no Internet access are provided with a computer and broadband Internet connection.

LISS panel members complete online questionnaires every month, for which they get an incentive of €7.50 per half hour of interview time. Each month, there can be several short or one or two longer questionnaires. It usually takes between 30 and 40 minutes to complete them all. Panel members have a personal login to the LISS panel website and can take the whole month to complete the questionnaires. They are invited by email and receive two reminders spread over the month when they do not complete the questionnaires.

We randomly assigned members of the LISS panel over mode (CATI vs CAWI). The CAWI-fieldwork was done by the regular staff of the LISS panel. The CATI fieldwork was done by the Dutch TNS-NIPO organization, but during the data collection a member of the LISS-research team was present as extra quality controller. For this experiment all respondents received a bonus incentive of 6.50 Euro, in addition to their regular incentive payment.

2.2 Experimental Design

We performed two question format experiments in Spring 2009. In the first experiment we investigated the effect of differential labelling of response categories: all response categories fully labelled versus only endpoints verbally labelled. In the second experiment, we investigated the effect of two-step branching or unfolding: all response categories offered directly in one step versus a two-step unfolding procedure. The question format experiments were embedded in a mixed-mode survey.

Firstly, LISS-panel members were randomly assigned to data collection mode**:** 2000 were assigned to the CATI-condition and 6134 were assigned to the CAWI (web)-condition. In the CATI-condition 1207 members of the 2000 members responded (60%), and in the

CAWI-condition 4003 of the 6134 members responded (65%). Secondly, within each mode respondents were again randomly assigned to experimental question format conditions in a two by two design (1: fully labelled vs. endpoint labelled and 2: one-step vs. two-step unfolding).

2.3. Questionnaire

Eight bi-polar opinion questions were used. Each question used a response scale with five response categories, ranging from "totally agree' to 'totally disagree'. "Do-not Know' was not explicitly offered, but was accepted when given. This to make the web-survey equivalent to the telephone survey, where it is always possible to spontaneously say 'do-not-know' to an interviewer.

The topic of the survey was the acceptability of usage of advanced medical technology. The questionnaire was balanced, in the sense that four questions were statements on acceptability and four on unacceptability. Examples are: 'If it will save a life, everything is permitted", and 'It is not desirable to utilize every medical invention just because it is technologically possible". These questions were part of a well-tested questionnaire that was used previously to investigate the acceptability by the Dutch population of medical technology (Steegers, Dijstelbloem, and Brom, 2008).

2.4. Analyses

Because the response rates between the two modes differed slightly (CATI 60% and CAWI 65%) we first checked if differential nonresponse influenced the comparability of the two experimental mode conditions. For all the LISS panel members biographical as well as psychographical information is available. When we compare the respondents in the CATI-CAWI conditions, we find some small but significant differences. CATI and CAWI respondents differ in age (47.8 vs. 46.2 years, p=0.02), household size (2.9 vs. 2.8, p=0.00), (non)urbanicity (3.1 vs. 3.0, p=0.05), house-ownership (78.5% vs. 75.0%, p=0.0). In a multivariate analysis using logistic regression, only age and household size remained significant. Although the differences are very small, a propensity score weight was constructed based on the logistic regression using age and household size as predictors. The mean weight is 1.00, with an standard deviation of 0.04 and a range from 0.91-1.27. All analyses will be carried out both weighted and unweighted, if the differences are negligible, the unweighted results are reported.

Analysis of variance was used to investigate the effects of mode and of question-format. The main variable of interest was mean score on acceptability of using medical technology. In other words, are there differences on the outcome regarding the topic of interest? In addition, we investigated whether mode and question format had an effect on the extremeness of responses. For this, we calculated the proportion values 1 and 5 (totally disagree and totally agree) in the eight questions. Finally, we investigated the effect of mode and question format on the total response distribution obtained on the five-point response scale (1: 'totally disagree', 2: 'disagree', 3:' neither nor', 4 'agree', and 5: 'totally agree'). Items were recoded in such a way that a high score always indicated acceptability of modern medical technology. In all analyses, the independent variables are mode (CATI vs. Web), and the two question formats: Direct question vs. two-step unfolding and Fully labelled vs. endpoint-only labelled.

3. Results

3.1 Mean score on acceptability

Two-way analysis of variance showed a small, but significant main effect of mode on acceptance of new technology ($F(1,5185)=13.33$, $p<.001$, partial $eta^2=0.003$). Respondents reported a slightly lower acceptance of medical technology over the telephone (mean acceptance=2.93) than when answering through the Internet (mean acceptance=2.99). There was also a small, but significant effect of 'unfolding' ($F(1,5185)=11.523$, $p<.01$, partial $eta^2=0.002$). Respondents in the two-step unfolding condition reported a lower acceptance of medical technology (mean acceptance= 2.93) than respondents in the one-step direct question condition (mean acceptance=2.98).

We found a non-significant effect of answer category labelling (endpoints labelled (mean acceptance 2.95) vs. fully labelled (mean acceptance 2.96); $F(1,5185)=0.956$, $p=.328$). However, no significant interaction effects between mode and question formats were found. The interaction effect of mode and labelling was non-significant ($F(1,5185)=0.564$, $p=.453$) and the interaction effect of mode and unfolding was also non-significant ($F(1,5185)=0.045$, $p=.883$).

The results could not be attributed to self-selection of respondents in the different modes, since weighting for nonresponse did not change these results.

3.2. Extremeness

We found significant effects of data collection mode and question format on the average acceptance of medical technology (see section 3.1). What may cause these differences? Previous research showed indications of more extremeness when reporting over the phone and also more extremeness with certain question formats (De Leeuw, 1992; Christian, 2007). Since means are notoriously sensitive to extreme scores, the differences in means found may be caused by more extremeness in the telephone mode and when the two-step unfolding format is used.

Our data reveal that besides differences in means, we also observe differences in standard deviations between experimental conditions; see also Table 1. This might indicate that more extreme responses may have caused the shift in means.

Table 1: Acceptability of medical technology score.
Minimum is 1 (totally unacceptable, maximum is 5 (totally acceptable)
Mean and Standard Deviation and N of cases for Experimental Conditions

	Mean	*Standard Deviation*	*N of Cases*
Mode	*P<.001*		
CATI	2.93	0.44	1207
CAWI	2.99	0.49	3986
Unfolding	P<.01		
No: One-step	2.98	0.44	2605
Yes: Two-step	2.93	0.51	2588
Labelling	n.s		
Fully labelled	2.96	*0.46*	2601
Endpoint labelled	2.95	*0.50*	2592

We defined extremeness as either answering 1 or a 5 on a five-point response scale. We then calculated the proportion extreme answers (either 1 or 5) over the eight acceptability

questions, and investigated whether this differed for data collection mode and the two question formats**.**

We found a small, but significant main effect of mode on extremeness of answers ($F(1,5202)=15.17$, $p<.001$, partial $eta^2=0.003$). Respondents in the telephone mode had a slightly higher proportion of extreme answers (mean proportion=0.28) than those in the web condition (mean proportion=0.22). There was also a small, but significant effect of 'unfolding' ($F(1,5202)=11.605$, $p<.01$, partial $eta^2=0.002$). Respondents to the two-step unfolding format had a higher proportion of extreme answers (mean proportion=0.28) than those answering to the one-step direct question (mean proportion=0.22). Although we did not find a significant effect of labelling on the mean score (see 3.1), we do find a significant effect of the way answer categories were labelled ($F(1,5202)=8.798$, $p<.01$,partial $eta^2=0.002$). Respondents who were presented with a full verbally labelled response scale gave less extreme answers (mean proportion of extreme answers=0.23), than respondents who were presented with end-point only labelled response scales (mean proportion of extreme answers=0.28).

Again, we found no significant interaction effects between mode and question formats. The interaction effect of mode and labelling was non-significant ($F(1,5202)=0.193$, $p=.661$) and the interaction effect of mode and unfolding was also non-significant ($F(1, 5202)=0.376$, $p=.540$).

The results could not be attributed to self-selection of respondents in the different modes, since weighting for nonresponse did not change these results.

3.3. Response Distribution

We found lower means in the telephone mode and when using a two-step unfolding format (see section 3.1), we also found a tendency to choose the most extreme response categories in the telephone mode, and when two-step unfolding and partially (endpoint-only) labelled question formats were used (see section 3.2). Finally, we investigated if mode and question format have differential effects on the response distribution. Since half of the questions were positively formulated and half of them negatively, we recoded all responses in such a way that a higher number always indicated more acceptability of modern medical technology. Next, we calculated the response proportion on each scale point, averaged over the eight questions. The results are summarized in Table 2.

Table 2: Effect of Mode, One step vs. Two step (Unfolding), and Fully vs. Endpoint Labeling
Difference in proportion on 8 questions: Range Five Point Scale

	CATI – CAWI	*Onestep – Twostep unfolding*	*Fully – Endpoint labelling*
Proportion 1	+ 0.02	- 0.06	- 0.03
Proportion 2	+ 0.07	+ 0.02	+ 0.00
Proportion 3	- 0.13	+ 0.04	+ 0.04
Proportion 4	+ 0.04	+ 0.06	+ 0.01
Proportion 5	- 0.00	- 0.05	- 0.02

Table 2 shows that in the telephone mode, respondents more often chose the extreme response categories 2 and 4 and less often the real endpoints, while they clearly chose the neutral midpoint far less often than web respondents. Furthermore, respondents to the two step unfolding format far more often chose the extreme endpoints, while respondents to the one-step direct question format more often chose the less extreme and middle categories. Finally, labelling of the response categories had the expected effect, that is, respondents who are presented with the endpoint-label-only format more often chose the extreme response categories, while respondents who were offered the fully labelled response format more often opted for the middle category.

4. Summary and Discussion

In a large study of the Dutch population we found small but consistent mode effects and question format effects, but no interaction effects between data collection mode and question format. These results can not be attributed to self-selection of respondents in the different modes, since weighting for selective nonresponse did not change the results.

Independent of scale format, telephone respondents provided lower mean ratings than web respondents. However the differences were small, the average difference between telephone and web was only 0.06 on a five-point scale. We also found that telephone respondents more often chose the extreme response categories than web respondents, a finding earlier reported by Christian, Dillman, and Smyth (2008) for a student population in the USA.

Regardless of mode, respondents to the two-step unfolding format provided lower mean ratings than respondents to the one-step format. These differences were also small: the average difference was 0.05 on a five-point scale. Although small, the difference is in accordance with earlier findings for web and telephone surveys among student populations in the USA (Christian, 2007). Respondents to a two-step unfolding format also more often chose the extreme response categories than respondents to a full, one-step question. These findings from the Dutch general population confirm earlier findings from the USA (Christian 2007, Groves, 1979). Furthermore, respondents who reacted to a partially (end-point only) labelled question format had a tendency to choose the more extreme response category when compared to respondents to a fully labelled question format. Labelling had an influence on proportion extremes chosen and on the standard deviation of a score on acceptability, but not on the means. Still, it is advised to fully label all scale point if possible, as previous research (Krosnick and Fabriger, 1997) shows that fully labelled scales have better psychometric properties.

We should emphasize that no interaction effects were found for mode and question formats. These results are encouraging for mixed-mode surveys; both mode and question format effects were small. We found a slightly lower acceptance score in telephone surveys and also when using the two-step unfolding format. However, in practice, two-step unfolding is often used in the aural telephone mode, while the one-step direct question is used in visual modes, such as web. The two small effects then sum up, and the total effect between the two mode systems is clearly larger. We therefore advise, not to use differential question formats in mixed-mode studies, but adhere to a unified mode design (Dillman, 2000) and use equivalent questions whenever possible.

Acknowledgements

The authors thank Stephanie Stam (CentERdata, Tilburg University) and Robbert Zandvliet (TNS-NIPO) for their valuable assistance during the data collection phase.

References

Blyth, Bill (2008). Mixed-mode: the only 'fitness' regime? *International Journal of Market Research, 50, 2,* 241-266.

Christian, Leah Melani (2007). *How mixed-mode surveys are transforming social research: the Influence of survey mode on measurement in web and telephone surveys*. Ph.D. dissertation, Washington State University.

Christian, Leah Melani, Dillman, Don A., and Smyth, Jolene D. (2008). The effect of mode and format on answers to scalar questions in telephone and web surveys. In James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster. *Advances in Telephone Survey Methodology*.New York: Wiley.

Couper, Mick P. (2008). *Designing Effective Web Surveys.* New York: Cambridge University Press

Dillman, Don A. (2000). *Mail and Internet surveys: The tailored design method* 2nd ed. New York: Wiley.

Dillman, Don A. and Melani Christian, Leah (2005). Survey mode as a source of instability in responses across surveys. *Field Methods, 17, 1,* 30-52.

Groves, Robert M. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly, 43,* 190-205.

Jackle, Annette, Roberts, Caroline, and Lynn, Peter (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review, 78, 1,* 3-20.

De Leeuw, Edith D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics (JOS), 21*, 233-255. Also available at www.jos.nu (Retrieved September 10, 2010)

De Leeuw, Edith D. (1992). *Data Quality in Mail, Telephone, and Face-to-Face Surveys.* Amsterdam: TT-Publikaties. Available online: www.xs4all.nl/~edithl (Retrieved September 10, 2010).

Krauter, Frauke, Presser, Stanley, and Tourangeau, Roger (2008). Social desirability bias in CATI, IVR, and Web Surveys. The effect of Mode and Question Sensitivity. *Public Opinion Quarterly, 2008, 72, 5,* 847-865. This issue is freely available online at http://poq.oxfordjournals.org/content/vol72/issue5/#ARTICLES (Retrieved June 21, 2010)

Krosnick, Jon A. and Fabrigar, Leandre R. (1997). Designing rating scales for effective measurement in surveys. In Lars lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. *Survey measurement and process quality.* New York: Wiley.

Link, Michael, W., and Mokdad, Ali H. (2005). Effects of survey mode on self-reports of adult alcohol consumption: A comparison of mail, web and telephone approaches. *Journal of Studies on Alcohol,* March 2005, pp239-245.

Rugg, D. (1941). Experiments in wording questions. *Public Opinion Quarterly, 5,* 91-92.

Schwarz, Norbert, Hippler, Hans-Juergen, Deutsch, Brigitte, and Strack, Fritz (1985). Response Categories: Effects on behavioural reports and comparative judgements. *Public Opinion Quarterly, 4*9, 388-395.

Schwarz, Norbert, Hippler, Strack, Fritz, Hans-Juergen, and Bishop, George (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*, 193-212.

Steegers, Chantal, Dijstelbloem, Huib, and Brom, Frans, W.A. (2008). *Meer dan status alleen. Burgerperspectieven op embryo-onderzoek.* [In Dutch: Assesment and points of view of citizens on technology and embryo-research]. The Hague: Rathenau-Instituut, TA rapport 0801.

Sudman, Seymour, Bradburn, Norman, and associates (1974). *Response effects in surveys.* Chicago: Aldine.

Tourangeau, Roger, Rips, Lance J., & Rasinski, Kenneth (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.