Design of Web Questionnaires

Toepoel, V.; Vis, C.M.; Das, J.W.M.; van Soest, A.H.O.

Publication date:
2006

Citation for published version (APA):
Toepoel, V., Vis, C. M., Das, J. W. M., & van Soest, A. H. O. (2006). *Design of Web Questionnaires: An Information Processing Perspective for the Effect of Response Categories*. (CentER Discussion Paper; Vol. 2006-19). Econometrics.

CentER

Discussion Paper

No. 2006–19

DESIGN OF WEB QUESTIONNAIRES: AN INFORMATION-PROCESSING PERSPECTIVE FOR THE EFFECT OF RESPONSE CATEGORIES

By Vera Toepoel, Corrie Vis, Marcel Das, Arthur van Soest

January 2006

ISSN 0924-7815

TILBURG UNIVERSITY

Working Paper, January 2006

Design of Web Questionnaires:
An Information-Processing Perspective for the Effect of Response Categories

Vera Toepoel[*], Corrie Vis[*], Marcel Das[*], and Arthur van Soest[**]

Abstract In this study we use an information-processing perspective to explore the impact of response scales on respondents' answers in a web survey. This paper has four innovations compared to the existing literature: research is based on a different mode of administration (web), we use an open-ended format as a benchmark, four different question types are used, and the study is conducted on a representative sample of the population. We find strong effects of response scales. Questions requiring estimation strategies are more affected by the choice of response format than questions in which direct recall is used. Respondents with a low need for cognition and respondents with a low need to form opinions are more affected by the response categories than respondents with a high need for cognition and a high need to evaluate. The sensitivity to contextual clues is also significantly related to gender, age and education.

JEL codes: C42, C81, C93

Keywords: web survey, questionnaire design, measurement error, context effects, response categories, need for cognition, need to evaluate

[*] CentERdata, Tilburg University, postal address: CentERdata, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Corresponding author: Vera Toepoel; e-mail: V.Toepoel@uvt.nl

[**] Tilburg University, Faculty of Economics and Business Administration, Department of Econometrics and Operations Research; and RAND, Santa Monica, USA.

1. Introduction

Judgments of the frequency of a person's behavior are one of the most commonly used questions in surveys. Range categories are used regularly and are often left to the knowledge or intuition of the researcher. Studies about the cognitive and communicative processes underlying question answering in surveys suggest that the choice of response categories can have a significant effect on respondent answers (Schwarz et al., 1985; Schwarz and Hippler, 1987; Strack and Martin, 1987; Krosnick and Alwin, 1987, Rockwood et al., 1997; Winter, 2002a; Winter, 2002b).

Based on a social information processing model proposed by Boudenhausen and Wijer (1987), Schwarz and Hippler (1987) argue that respondents use the response alternatives to determine the meaning of the question and use the frequency range suggested by the response alternatives as a frame of reference, extracting information about presumably common answers from the values stated in the scale.

This paper replicates pats of previous studies by Schwarz et al. (1985) and Rockwood et al. (1997) and adds to the existing literature in four directions. First, previous studies used paper and telephone as modes of administration, while we consider response category effects in an online web survey. Despite the enormous use of web questionnaires, the knowledge of what people read and comprehend and why, is still in its infancy (Redline et al., 2003). While a theory of web questionnaire design may draw from the principles for visual layout and design of paper questionnaires, it will also have new features and require independent testing and evaluation (Dillman et al., 1998). Therefore, it is important that response category effects are tested in an online survey.

In our experiment, we used high versus low answer ranges. A second contribution to the existing literature is the addition of an open-ended question format as a benchmark as suggested by Rockwood et al. (1997), who found that response category effects differ per question type: they found no differences in response category effects for salient and irregular questions (questions in which direct recall is used in response formatting and the occurrence of the event is episodic) but significant differences for mundane and regular questions (questions for which estimation is likely to be used in recall and the event occurs regularly). As a third addition to the literature, our experiment evaluates the full range of question possibilities: mundane and regular, salient and regular, as well as salient and irregular, and mundane and irregular.

Most of the previous studies used a convenience sample in their experiments (like a group of students). With a rather homogeneous sample, it is not possible to measure the effects of personal characteristics on survey responses. This paper adds to the existing literature in the sense that a representative sample of the Dutch population was used. This allows for testing in which way personal characteristics account for variance in survey responding. An indicator for respondent's need to think and evaluate is included in the analysis, as well as gender, age, and education variables.

2 Background

To find out if response categories influence respondent behavior, we need to know how respondents answer questions. Trying to understand how respondents comprehend survey questions leads inevitably to a more basic search for cognitive processes involved in answering questions. Interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, formatting a response, and editing it are the main psychological components of a process that starts with respondent's exposure to a survey question and ends with their report (Sudman et al., 1996). Performance of each of these steps is very context-dependent. Most of the answers that are recorded in surveys reflect judgments that respondents generate on the spot in the context of the specific interview. The words and visual stimuli are perceived as information. Respondents are influenced by all the information they perceive, so that their answers will be influenced by preceding questions as well as questionnaire and question format.

By using closed questions a respondent is asked to give his opinion by checking the appropriate value from a set of frequency response alternatives provided to him. Schwarz (1996) argues that this range may serve as a source of information to the respondent. A respondent assumes that the researcher constructed a meaningful scale that reflects his or her knowledge about the distribution of the behavior. Values in the middle range of the scale are assumed to reflect ‘average’ behavior, whereas the extremes of the scale are assumed to correspond to the extremes of the distribution. Therefore, giving a response is the same as locating one’s own position in the distribution. The more ambiguous the target behavior is defined, the more pronounced is the impact of the response alternatives. But even when the behavior at target is well defined, the range of response alternatives may affect respondents’ frequency estimates. Watching television, for example, is not presented in memory as a distinct episode but the various episodes go together in a more generic presentation of the behavior that lacks temporal markers. When asked how often a respondent watches television, respondents therefore cannot recall the episodes to determine the frequency of the behavior. Instead, they rely on estimation strategies. Respondents may not even try to recall how much they engage in a particular behavior, but rather use their biographical knowledge to locate themselves in the distribution suggested by the response scale. For example, a respondent who considers himself an ‘average TV viewer’ may select a response category in the middle part of the response scale without reviewing his actual TV consumption. Or a respondent may be reluctant to select a response category that seems unusual in the range of responses. This results in higher frequency estimates along scales that present high rather than low frequency response alternatives.

One of the most basic decisions a survey designer has to make is whether to use open or closed questions. From a cognitive perspective, open questions present a free-recall task to respondents whereas closed questions present a recognition task. In open questions respondents are unlikely to spontaneously report information that seems self-evident. Closed questions, on the other hand, may fail to provide an appropriate set of meaningful alternatives in substance or wording. Furthermore, respondents are influenced by the specific closed alternatives given. One can expect a more valid answer if the respondent must produce an answer himself (Schuman and Presser, 1981). Schwarz et al. (1985) and Schwarz (1996) recommend asking behavioral frequency questions in an open response format.

Research shows that frequency judgments, which by necessity rely on a person’s memories, contain difficulties that are not easy to correct. When respondents are asked to report how regularly they do something, they may use one of two strategies to arrive at an answer. If the question refers to discrete behaviors that occur with a low frequency, such as buying a new car, they may try to recall all instances of that behavior. In that case, the accuracy of their reports will depend on the accuracy of their memory. For more regular and mundane behaviors, such as watching TV, respondents have to provide an estimate of their behavior, using whatever information is available to them at the time of judgment. In computing this estimate, they may use the range of the response alternatives as a frame of reference. Subjects tend to overestimate the frequency of irregular events and to underestimate occurrence of events that happen regularly (Schwarz and Hippler, 1987; Strube, 1987). Menon et al. (1995) find that the range of response alternatives affect frequency reports of moderately regular and irregular behaviors, but not of very regular behaviors. They suggest relevant frequency information was inaccessible for the less regular behaviors, causing respondents to rely on response alternatives as a cue in computing a frequency estimate. Respondents may be more susceptible to context effects if relevant information is not accessible to them. Rockwood et al. (1997) conclude that response categories have a significant effect on response formulation in regular and mundane questions, whereas in irregular and salient questions the response categories do not have a significant effect. Although these studies show different results in relation to question types and response category effects, the argument that context effects are more likely to emerge if information is more difficult to process holds for both studies.

Krosnick et al. (1996) give a cognitive explanation of response effects. Their theory assumes that most respondents answer survey questions by choosing the first satisfactory or acceptable response alternative rather than select the true answer. The tendency to satisfy depends on three things: (1) the difficulty of the question/answer, (2) the respondent’s ability to retrieve, process and integrate information from memory, and (3) the respondent’s motivation. While the first is dependent of the question itself, the latter two depend on the respondent’s personal characteristics.

Whether or not the response scale influences a respondent may differ for the respondent’s cognitive activity in answering the survey. Cacioppo and

Petty (1982) developed a scale to measure the need for cognition. Need for cognition (NFC) represents the tendency for individuals to engage in and enjoy thinking. They reasoned that when respondents are motivated (such as when the topic is of high relevance to the respondent) respondents are more eager to think than when their motivation is low (such as when the topic is of low interest). According to them, not only situational factors determine how much thinking occurs. Individual differences in intrinsic motivation to engage in cognitive activity are also likely to affect the effort a respondent is willing to make. People with a high need for cognition (HNC) undergo different processes in formatting an answer than people with a low need for cognition (LNC). People with HNC tend to seek more information and think more carefully before making an evaluation than people with LNC, who are more easily influenced by peripheral cues.

Not only a person's need to think could affect the presence of context effects, a person's need to evaluate could also play a role. Jarvis and Petty (1996) developed a measure to assess individual differences in the propensity to engage in evaluation, the Need to Evaluate Scale (NES). One could expect that those with a High Need to Evaluate (HNE) are more likely to have formed attitudes toward objects or situations, and are therefore less sensitive to context effects in response scales, than people with a low need to evaluate (LNE). Evaluation by no means requires effortful thought. The relation between the NES and the NFC was tested by Jarvis and Petty and was found to be moderate and positive ($r=.35$, $p<.001$).

Research suggests that people differ in the extent to which they think about and evaluate issues. Petty and Jarvis (1996) suggest that need for cognition and need for evaluation are associated with a number of survey effects. LNCs and LNEs are expected to be more susceptible to various low effort biases than HNCs and HNEs, such as being influenced by cues in a survey that suggest one response over another. Whether or not a respondent formulates an answer based on retrieval (in memory) or construction (building an answer at the time of answering the survey) might also be influenced by the need for cognition and the need for evaluation.

Jarvis and Petty (1996) conclude that with including a respondent's score on the NES as a control variable, a researcher can account for unexplained variance in responses. Assuming that the NES will account for variance in the respondent's evaluative responses and will be uncorrelated with the independent variable, this reduction in error variance could make the test of an effect of any independent variable on responding more powerful. This benefit would be in addition to any potential for discovering informative interactions between the researcher's independent variables and the need to evaluate. One can expect the same effects for need for cognition.

How strongly the scale biases answers will depend on how much the scale deviates from the respondent's actual behavior. A scale that matches a respondent's behavior increases the validity of the answers. However, the effect of a response scale may be different for different subpopulations. Because all respondents use the same scale, it may tend to reduce the differences between different subpopulations. The more information one has available in memory, the less susceptible one will tend to be to differences in the information that is immediately available in response alternatives. In their analysis of order effects, Krosnick and Alwin (1987) find that respondents with less cognitive sophistication are more likely to be influenced by changes in response order. Respondents with less education and more limited vocabularies are influenced more by manipulation of answer categories. Furthermore, Lynch et al. (1991) showed that context directly affects the people with less knowledge on the topic in question (novices) than experts. Thus, the more interested one is in a certain question topic, the less susceptible one is to differences in scale range.

3. Design and Implementation

Our study builds on previous studies by Schwarz et al. (1985) and Rockwood et al. (1997). Based on Rockwood et al. (1997) two question extremes are investigated: regular versus irregular and mundane versus salient. Rockwood et al. conclude that further research on response categories effects should investigate not only regular/mundane and irregular/salient question types, but also the full range of question possibilities: regular/mundane, regular/salient, as well as irregular/salient and irregular/mundane. They also conclude not only to use low and high answer categories, but also to introduce a third experimental condition in which the questions are asked open-ended. A study in which high versus low answer categories and open versus closed questions are investigated would greatly improve the understanding of the issues involved with context effects in answer categories.

Based on the literature, Figure 1 presents a process model for formatting a response. The response format influences the process in which a respondent formulates an answer. The response category effect is influenced by question type. The influence of question type on response category effects may depend on personal characteristics such as gender, age, and education. Also, a respondent's need to think and evaluate may play a role. Our study investigates whether there are response category effects in web surveys and if they differ per question type. We also investigate if personal characteristics cause stronger or weaker response category effects.

[figure 1]

The study was conducted in the CentERpanel, an online household panel consisting of more than 2,000 households. This panel is representative for the Dutch population and is administrated by CentERdata, Tilburg University (the Netherlands). CentERdata provides a set-top box to people who do not have a computer to make it also possible for them to complete the questionnaires online.

Four questions were asked in which the response scale was manipulated (see Table 1 for the questions asked). The topics of the questions were: hours per day watching television;[1] number of visits per year to a hairdresser; number of attended birthday parties per year, and days per year on holiday (away from home). An effort was made to find topics that vary in regularity and saliency of occurrence for most people. The question about the days per year on holiday is a question type in which direct recall is used in response formatting and the occurrence of the event is episodic. The number of attended birthday parties is a less episodic question type, while for visits to a hairdresser estimation strategy is likely to be used. Hours watching TV is a

[1] As used by Schwarz et al. (1985) and Rockwood et al. (1997).

question type that is not presented in memory as a distinct episode but the various episodes go together in a more generic presentation of the behavior that lacks temporal markers. Respondents therefore cannot recall the episodes to determine the regularity of the behavior, and have to rely on estimation strategies.

[table 1]

The response rate[2] was 81,8% (2924 persons were selected, 2393 participated). People were randomly assigned to format A (low response scale), format B (high response scale), or format C (open-ended question). See Table 2 for the response scales used.

[table2]

The existing literature suggests that response category effects are not the same for all question types. Watching TV is mundane and occurs regularly in most people's lives, so the actual amount of time is not likely to be remembered. As a result, one can expect that the effect of different response categories can have a significant effect on response formation in this type of question. On the other hand, questions about a respondent's holiday are well defined and response formation can be based on direct recall. Questions about regular and mundane behavior are likely to be more affected by the choice of response format than respectively mundane and irregular, salient and regular, and salient and irregular.

Need for cognition and need to evaluate are measure with questions on 34 and 16, respectively (see Appendices A and B). By counting the scores of the items, an overall cognition, respectively evaluation score, is derived. Using the mean score for both constructs, respondents were divided in a low and a high group. Based on Schwarz (1996) NFC and NES were combined into 4 quadrants (see Table 3). The first group consists of people who are low in their cognitive activity both in thinking and in evaluating. They are the most

[2] Response Rate 1 defined in the Standard Definitions of AAPOR (www.aapor.org)

likely to be affected by the choice of response format, because they are more easily influenced by peripheral cues. The second group consists of persons who don't like to think but do like to evaluate. They do form opinions but don't think them through. The third group consists of people who do like to think but do not like to evaluate. Their answers are constructed at the time they complete the survey, but they do think about their answers. The last group consists of people with a high need for cognition and a high need to evaluate. These people are expected to be the least sensitive to the response format.

[table 3]

Krosnick and Alwin (1987) find that respondents with less education and more limited vocabularies are more influenced by manipulation of answer categories. Lynch et al. (1991) showed that the context directly affected the people with less knowledge on the topic in question (novices) more than experts. Thus, the more interested one is in a certain question topic, the less susceptible one is to differences in scale range. As a result, one could expect women to be less sensitive to context effects in questions about a hairdresser than men are (assuming women are more interested in their hair than men).

4. Results

In Section 4.1 response scale effects are analyzed. Low and high response scales are compared, as well as respondents' reports in an open-ended format. Also, the influence of question type is taken into account. In Section 4.2 a respondent's need for cognition and need to evaluate in relation to response scale effects are discussed, and in Section 4.3 a closer look at gender, age, and education is taken.

4.1 Response scale effects

To assess the impact of the response scale on respondents' reports, the responses in the low response scale (see e.g. format A in Table 2) and the high response scale (see format B in Table 2) were summarized as either (a) two and a half hours or less, or (b) more than two and a half hours for the hours watching TV.[3] For birthday parties and days on holiday the low response scale and the high response scale were summarized as either (a) 17[4] or less, or (b) more than 17, and for visiting a hairdresser the low response scale and the high response scale were summarized as either (a) 9[4] or less, or (b) more than 9. We dichotomized the answer categories to remain consistent with previous research. The open-ended condition is seen as an unbiased benchmark since it does not provide any anchor to the respondent.

[table 4]

As expected, the range of the response scale affected respondents' behavior reports, as can be seen in Table 4. Only 22.0% of the respondents who were presented the low response scale reported watching TV for more than two and a half hours, while 53.6% of the respondents presented the high response scale did so. In comparison, 52.1% of the respondents that answered the question in the open-ended reported a TV consumption of more than two and a half hours. Comparing the different conditions, apparently the high response scale best matches the respondent's behavior; while the low response scale versus the high response scale and the low response scale versus an open-ended question show significantly different answers, the high response scale versus open-ended answers do not differ significantly (see Table 5).

With regard to birthday parties, all three conditions show significant differences; 25.6% of the respondents in the low response scale attend more than 17 birthday parties a year, compared to 44.6% of the high response scale and 39.4% of the respondents in the open-ended condition. With the open-ended question in the midst of the low and high response scale, the frequency ranges of the low and high scale both divert answers in relation to the free-recall task.

The question about visiting a hairdresser shows again statistically significant differences for the low and high scale conditions. But in this

[3] As used by Schwarz et al. (1985) and Rockwood et al. (1997).
[4] This number is based on a pilot study.

question, the answers on the low response scale are closer to the open-ended answers.

For the question on the number of days a respondent spent on holiday, only the difference between high and low response scale is significant.

In summary, the data provide strong support for the hypothesis that the range of response categories affects respondent's behavior reports. We found higher frequency estimates along scales that present high rather than low frequency response alternatives. This indicates an anchoring effect, as suggested by Schwarz (1996). All four questions show statistical differences in the high versus the low response scale. The open-ended condition is sometimes more similar to one response scale than to the other. How strongly the scale biases a respondent's answer, is influenced by how the scale relates to the population distribution. If the distribution of categories is closer to the population distribution, the influence of response categories is less pronounced.

[table 5]

The impact of response alternatives on behavioral frequency judgments is expected to depend on the regularity and the salience of the behavior. Questions about regular and mundane behavior are expected to be more affected by the choice of response format than mundane and irregular, salient and regular, and salient and irregular respectively. Table 5 shows an overview of significance and correlation between response formats for the different question types. With relation to the high versus the low response scale, the largest correlation is found in hours watching TV (mundane/regular), followed by the question on birthday parties (mundane/irregular), visiting a hairdresser (salient/regular), and days on holiday (salient/irregular). As expected, the impact of response categories differs across questions. Comparison of the open-ended question with the different response scales shows similar results, although not all comparisons reach statistical significance.

4.2 Need for Cognition and Need to Evaluate

Because the existing literature suggests that the need for cognition and the need to evaluate account for variance in survey responses, we include these in the analysis of response category effects. Table 6 shows the separate construct groups as well as the 4 quadrants in which we combine NFC and NES. This table indicates the significance and the strength of the deviation between response scales (high versus low scale) for each question type. Thus, the same analysis as in Section 4.1 was conducted, but then for each subgroup.

[table 6]

The difference in reports between respondents who were offered the low response scale and respondents who were offered the high response scale is greater in the mundane/regular question for respondents with a low need for cognition (NFC). Our hypothesis that respondents who score low on the NFC construct are more sensible for context effects is confirmed. However, we do not find evidence that NFC accounts for differences in response effects in the mundane/irregular question type. In the salient/regular question type, respondents who have a low NFC show a very similar deviation (eta=.201) between response scales to respondents with high NFC scores (eta=.217). Context effects in this question type influence respondents who do not like to think less. For the salient and irregular question type, there are only significant differences between answer scores in the high versus the low response scale for respondents with a low score on NFC. Respondents with a high NFC are not sensitive for response category effects in this question type. The same results are found for the Need to Evaluate construct.

Combining need for cognition and need to evaluate into 4 quadrants, similar results are found as the separate constructs in the first question. Table 6 indicates that for the regular and mundane question (hours watching TV), people with a low need for cognition and a low need to evaluate show the largest deviation between the low and high response scale (eta=.409). Because the first quadrant consists of people who are low in their cognitive activity both in thinking and in evaluating, they are the most likely to be affected by the choice of response format, because they are more easily influenced by peripheral cues. The second quadrant, consisting of persons who don't like to think but do like to evaluate, shows a lower correlation (eta=.353). The third quadrant with people who do like to think but do not like to evaluate have the smallest correlation between the different response scales (eta=.267). The people with a high need for cognition and a high need to evaluate are more affected by the response scale (eta=.306) than people in the third quadrant in the mundane/regular question type. In the mundane/irregular question type the deviation scores in the quadrants increase drastically compared to the separate constructs, indicating that the combination of NFC and NES increases the differentiation in context effects. Especially in the quadrants with a high need to evaluate (groups 2 and 4) the deviation between the high and low response scale is high. Apparently they evaluate on the spot, influenced by peripheral cues. In this question type, people with a high NFC and a low NES (group 3) have the most similar results in the different response scales alternatives: differences between the high and low scale now even do not reach statistical significance. Respondents, who score low on both constructs, have low variances as well. Looking at the quadrants in the salient/regular question type, especially the people who do not think things through well and who do evaluate a lot show more differences in results between the low response scale and the high response scale. For the salient and irregular question type, there are no significant differences between answer scores in the high versus the low response scale for respondents in the different quadrants. Again, the salient/irregular question is not very sensitive for context effects.

4.3 Personal characteristics

Context effects are not only different for people who differ in their need to think or evaluate. There are differences between gender, age groups, and education groups as well. Table 7 shows the results of a comparison between the high and low response scale for groups of respondents with different demographic characteristics.

[table 7]

From Table 7 it can be concluded that men are more affected by contextual cues than women. In the mundane/regular question, the salient/regular question type, and the salient/irregular question they show more differences in answer score between the low and the high response scale. Men are more distracted by peripheral cues in questions about regular behavior. Only in the mundane/irregular question women have a higher association between question type and scale type. The assumption that women are more interested in their hair than men and therefore are less distracted by category ranges was confirmed; women show a significantly smaller deviation between scales than men.

Table 7 also shows that respondents in the age of 15-24 are the least affected by the response scale offered in the mundane/regular question. Respondents in age group 25-34 show the highest relation between the scale presented and the reports of behavior. As of then, the influence of scale drops till the age of 65. With regard to the mundane/irregular question, respondents in the age of 15-24 show the highest difference (this in contrast to the mundane/regular question where they show the least difference). They apparently do not remember very well how many times they visited a birthday party. There is a U-shaped pattern in the effect, with a minimum response category effect at age 45. The same goes for the salient/regular question, but there the turning point is at age 35. It might be the case that visiting a hairdresser becomes less salient and/or regular in this stage of life. The elderly do not seem to remember the number of haircuts well either; again they have the highest correlation score. For the salient and irregular questions about behavior no significant age affect is found.

Based on the literature, we would expect that low educated respondents would be more likely to be susceptible to context effects. Our results are not so clear-cut. Table 7 shows that for the mundane and regular questions, the primary education level shows a reasonably high difference between response scales. The same goes for the mundane and irregular questions. For the other question types there were no significant differences in the first education group. We did not find that the highest education group is the least susceptible to context effects. The response scale influences the higher secondary education group the least.

5. Discussion and Conclusions

In this paper an information-processing perspective to explore the impact of response categories on the answers respondents provide in web surveys is used. This perspective focuses on the assumption that response scales influence the way in which respondents formulate answers in web surveys, as they do in other modes of administration. Respondents consider the information in the answer format as a reference to the ‘usual’ frequency of a specific behavior. In the present study we explored how the response scale affects behavior frequency reports. We looked at closed versus open-ended questions. The question type was taken into account in the analysis: the regularity and the saliency of the behavior at target. An examination of respondent’s behavioral reports indicates that respondents who were presented the low response scale report lower frequencies than those with a high response scale. We replicate the findings of Schwarz et al. (1985) and Winter (2002a) that response scales are perceived as informative. An extension of this study is that it also uses an open-ended format, avoiding the bias due to response scale anchors. The open-ended answers were sometimes more similar to one response scale than to another, depending on how the response scale relates to the population distribution.

Questions about regular and mundane behavior are more affected by the choice of response scale than irregular and mundane, regular and salient, and irregular and salient respectively. Comparison of the open-ended question with the different response scales shows similar results, although not all comparisons reach statistical significance. Our results are in line with Rockwood et al. (1997): response scales have a significant effect on response formulation in regular and mundane questions (response based on estimation strategies), whereas in irregular and salient questions (response based on direct recall) the response scales have a smaller effect. Because questions in which estimation strategies have to be used are most susceptible to measurement error, future research should focus on how to measure these kinds of question types best.

The hypothesis that response scale effects differ for respondents with different personal characteristics was confirmed. The Need for Cognition and the Need to Evaluate constructs account for variance in survey responding. In most question types, the deviation in reports between respondents who were offered the low response scale and respondents who were offered the high response scale is greater for respondents with a low need for cognition. The same goes for need to evaluate. Combining need for cognition and need to evaluate, we only found the expected higher deviations for respondents with low scores on the constructs in the mundane/regular question type, but not in the other question types. We did not find statistical differences for respondents who differ in their need to think and evaluate in the salient/irregular question.

Context effects are not only different for people who differ in their need to think or evaluate. We found differences between sexes, age groups, and education groups. Men are more affected by contextual cues than women. Only in the mundane/irregular question type women have a higher association between question type and scale type. The influence of scale drops till the age of 65. People in the age group of 65 and older show relatively large differences between response scales. For the salient and irregular questions about behavior there does not seem to be an age affect; differences between the high and low scale range are not statistically significant. The effect of education on response scale effects is not clear-cut.

This research has advanced our understanding of measurement error in web surveys. As in paper and telephone surveys, response category effects emerge in web surveys. Salient and irregular questions, in which respondents can use their recollection, are less sensible for response scale effects. On the other hand, questions about mundane and irregular behavior, which are difficult to remember, are very much affected by the response scale. These response scale effects are not the same for subpopulations of the NFC and NES constructs, gender, age, and education. When designing a web survey, a researcher should keep this in mind in order to reduce measurement error. An open-ended format is preferable in questions in which estimation strategies have to be used. If this type of answer format is not desirable, a careful strategy has to be used when using closed questions.

References

Bodenhausen, Galen V. and Robert S. Wijer, Jr (1987), “Social Cognition and Social Reality: Information Acquisition and Use in the Laboratory and the Real World“. In: Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman (eds.), *Social Information Processing and Survey Methodology,* Springer-Verlag, New York, 6-41.

Cacioppo, John T. and Richard E. Petty (1982), “The Need for Cognition,” *Journal of Personality and Social Psychology* 42, 116-131.

Dillman, Don A., Robert D. Tortora, and Dennis Bowker (1998), *Principles for Constructing Web Surveys,* SESRC Technical Report 98-50, Pullman, Washington. Retrieved 10-25-2005 on http://survey.sesrc.wsu.edu/dillman/papers.htm

Jarvis, W. Blair G. and Richard E. Petty (1996), “The Need to Evaluate,” *Journal of Personality and Social Psychology* 70, 172-194.

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, 201-219.

Krosnick, Jon A., Sowmya Narayan, and Wendy R. Smith (1996), “Satisficing in Surveys: Initial Evidence,” *New Directions for Program Evaluation* 70, 29-44.

Lynch, John G., Dipankar Chakravarti and Mitra Anusree (1991), “Contrast Effects in Consumer Judgments: Changes in Mental Representations or in the Anchoring of Rating Scales?,” *The Journal of Consumer Research* 18, 284-297.

Menon, Geeta, Priya Raghubir, and Norbert Schwarz (1995), “Behavioral Frequency Judgments: An Accessibility-Diagnosticity Framework,” *The Journal of Consumer Research* 22, 212-228.

Petty, Richard E. and W. Blair G. Jarvis (1996), “An Individual Differences Perspective on Assessing Cognitive Processes”. In: Norbert Schwarz, and Seymour Sudman (eds.), *Answering Questions,* Jossey-Bass Publishers, San Francisco, 221-257.

Redline, Cleo D., Don A. Dillman, Lisa Carley-Baxter, and Robert Creecy (2003), *Factors that Influence Reading and Comprehension in Self-Administered Questionnaires,* paper presented at the workshop on Item-Nonresponse and Data Quality, Basel, Switzerland, October 10, 2003. Retrieved 10-25-2005 on http://survey.sesrc.wsu.edu/dillman/papers.htm

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

Strube, Gerhard (1987), “Answering Survey Questions: The Role of Memory”. In: Hans-J. Hippler , Norbert Schwarz and Seymour Sudman (eds.), *Social Information Processing and Survey Methodology,* Springer-Verlag New York, 86-101.

Schuman, Howard and Stanley Presser (1981), *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording and Content.* Academic Press: Quantitative Studies in Social Relations, New York.

Schwarz, Norbert (1996). *Cognition and Communication. Judgmental Biases, Research Methods, and the Logic of Conversation.* Lawrence Erlbaum Associates, Publishers, New Jersey.

Schwarz, Norbert and Hans-J. Hippler (1987), “What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives“. In: Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman (eds.), *Social Information Processing and Survey Methodology,* Springer-Verlag, New York, 163-178.

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

Strack, Fritz and Leonard L. Martin (1987), “Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys”. In: Hans-J. Hippler, Norbert Schwarz and Seymour Sudman (eds.), *Social Information Processing and Survey Methodology,* Springer-Verlag New York, 123-148.

Sudman, Seymour, Norman Bradburn, and Norbert Schwarz (1996), *Thinking About Answers,* Jossey-Bass Publishers, San Francisco.

Winter, Joachim K. (2002a), “Design Effects in Survey-Based Measures of Household Consumption,” Discussion Paper No. 02-34, Sonderforschungsbereich 504, University of Mannheim.
Retrieved 16-01-06 on
http://www.sfb504.uni-mannheim.de/publications/dp02-34.pdf

Winter, Joachim K. (2002b). “Bracketing Effects in Categorized Survey Questions and the Measurement of Economic Quantities”, Discussion Paper No. 02-35, Sonderforschungsbereich 504, University of Mannheim.
Retrieved 16-01-06 on
http://www.sfb504.uni-mannheim.de/publications/dp02-35.pdf

Appendix A

The Need for Cognition Scale (Cacioppo and Petty, 1982) is a scale designed to measure the tendency for individuals to engage in and enjoy thinking. The list of 34 items is presented below.

1. I really enjoy a task that involves coming up with solutions to problems.
2. I would prefer a task that is intellectual, difficult, and important to one that's somewhat important but does not require much thought.
3. I tend to set goals that can be accomplished only by extending considerable mental effort.
4. I am usually tempted to put more thought into a task than the job minimally requires
5. Learning new ways to think doesn't excite me very much.*
6. I am hesitant about making important decisions after thinking about them.*
7. I usually end up deliberating about issues even when they do affect me personally.
8. I prefer to let things happen rather than try to understand why they turned out that way.*
9. I have difficulty in thinking in new and unfamiliar situations.*
10. The idea of relying on thought to get my way to the top does not appeal to me.*
11. The notion of thinking abstractly is not appealing to me.*
12. I am an intellectual.
13. I only think as hard as I have to.*
14. I don't reason well under pressure.*
15. I like tasks that require little thought once I've learned them.*
16. I prefer to think about small daily projects to long-term ones.*
17. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
18. I find little satisfaction in deliberating hard and for long hours.*
19. I more often talk with other people about the reasons for and possible solutions to international problems than about gossip of tidbits of what famous people are doing.
20. These days, I see little chance for performing well, even in 'intellectual' jobs, unless one knows the right people.*
21. More often than not, more thinking just leads to more errors.*
22. I don't like to have the responsibility of handling a situation that requires a lot of thinking.*
23. I appreciate opportunities to discover the strengths and weaknesses of my own reasoning.
24. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.*
25. Thinking is not my idea of fun.*
26. I try to anticipate and avoid situations where there is a likely chance I'll have to think in depth about something.*
27. I prefer watching educational to entertainment programs.
28. I think best when those around me are very intelligent.

29. I prefer my life to be filled with puzzles that I must solve.
30. I would prefer complex to simple problems.
31. Simply knowing the answer rather than understanding the reasons or the answer to a problem is fine with me.*
32. It's enough for me that something gets the job done; I don't care how or why it works.*
33. Ignorance is bliss.*
34. I enjoy thinking about an issue even when the results of my thoughts will have no outcome on the issue.

*=item is reverse worded

Answer format:

1	2	3	4	5
extremely uncharacteristic				extremely characteristic

Appendix B

The Need to Evaluate Scale (Jarvis and Petty, 1996) is a scale designed to measure individual differences in the propensity to engage in evaluation. The list of 16 items is presented below.

1. I form opinions about everything.
2. I prefer to avoid taking extreme positions.*
3. It is very important to me to hold strong opinions.
4. I want to know exactly what is good and bad about everything.
5. I often prefer to remain neural about complex issues.*
6. If something does not affect me, I do not usually determine if it is good or bad.*
7. I enjoy strongly liking and disliking new things.
8. There are many things for which I do not have a preference.*
9. It bothers me to remain neutral.
10. I like to have strong opinions even when I am not personally involved.
11. I have many more opinions than the average person.
12. I would rather have a strong opinion than no opinion at all.
13. I pay a lot of attention to whether things are good or bad.
14. I only form strong opinions when I have to.*
15. I like to decide that new things are really good or really bad.
16. I am pretty much indifferent to many important issues.*

*=item is reverse worded

Answer format:

1	2	3	4	5
extremely uncharacteristic				extremely characteristic

Table 1.
Types of questions used in the experiment

Questions	Regular	Irregular
Mundane	How many hours per day do you typically watch TV?	How many birthday parties do you typically attend per year?
Salient	How many times did you go to the hairdresser last year?	How many days did you leave your home (have a holiday) last year?

Table 2.

Response scales used in the experiment

Response scales	Format A	Format B	Format C
Hours watching TV			
1	½ hour or less	2½ hour or less	open-ended question
2	½ - 1 hour	2½ - 3 hours	
3	1 - 1½ hours	3 - 3½ hours	
4	1 ½ - 2 hours	3½ - 4 hours	
5	2 - 2 ½ hours	4 - 4 ½ hours	
6	more than 2 ½ hours	more than 4 ½ hours	
Birthday parties			
1	9 or less	17 or less	open-ended question
2	9 - 11	17 - 19	
3	11 - 13	19 - 21	
4	13 - 15	21 - 23	
5	15 - 17	23 - 25	
6	more than 17	more than 25	
Visiting a hairdresser			
1	1 or less	9 or less	open-ended question
2	1 - 3	9 - 11	
3	3 - 5	11 - 13	
4	5 - 7	13 - 15	
5	7 - 9	15 - 17	
6	more than 9	more than 17	
Days on holiday			
1	9 or less	17 or less	open-ended question
2	9 - 11	17 - 19	
3	11 - 13	19 - 21	
4	13 - 15	21 - 23	
5	15 - 17	23 - 25	
6	more than 17	more than 25	

Note: answer categories 1 to 5 in Format A match answer category 1 in Format B. Answer category 6 in Format 1 matches answer categories 2 to 6 in Format B.

Table 3.
Different groups in the experiment for Need for Cognition (NFC) and Need to Evaluate (NES) and combination of NFC/NES into four quadrants

	Low	**High**
NFC	(NFC<112*) N=688	(NFC>111*) N=638
NES	(NES<52*) N=663	(NES>51*) N=633

	Low NFC	**High NFC**
Low NES	Group 1 (N=490)	Group 3 (N=173)
High NES	Group 2 (N=198)	Group 4 (N=465)

*Counting scores on the 34 NFC items and the 16 NES items yield the overall score per person. With a minimum of 53 and a maximum of 157, 111 is the mean score for NFC, and with a minimum of 27 and a maximum of 80, 51 is the mean score for NES.

Table 4.

Overview of frequencies of the results from different response formats

	Low Response Scale		High Response Scale		Open-Ended	
	X* or less	more than X*	X* or less	more than X*	X* or less	more than X*
Mundane and Regular						
Hours watching TV	78.0%	22.0%	46.4%	53.6%	47.9%	52.1%
Mundane and Irregular						
Birthday Parties	74.4%	25.6%	55.4%	44.6%	60.6%	39.4%
Salient and Regular						
Visiting a Hairdresser	84.7%	15.3%	72.1%	27.9%	81.5%	18.5%
Salient and Irregular						
Days on Holiday	53.9%	46.1%	46.6%	53.4%	49.8%	50.2%

*X=2.5 for hours watching TV and listening to the radio, 9 for visiting a hairdresser, and 17 for birthday parties and days on holiday.

Table 5.
Overview of significance and association between response formats per question type

	High Response Scale vs. Low Response Scale		Low Response Scale vs. Open-Ended		High Response Scale vs. Open-Ended	
	p<	eta	p<	eta	p<	eta
Mundane and Regular						
Hours watching TV	.01	.325	.01	.311	n.s.	n.s
Mundane and Irregular						
Birthday Parties	.01	.199	.01	.148	.05	.052
Salient and Regular						
Visiting a Hairdresser	.01	.152	n.s.	n.s.	.01	.112
Salient and Irregular						
Days on Holiday	.01	.073	n.s.	n.s.	n.s.	n.s.

Note: A higher correlation coefficient (eta) between the answer score and the scale that was used indicates greater differences between response formats.
n.s.=non significant: there are no statistically significant differences between formats in this question type.

Table 6.
Overview of significance and association between the low and high response scale per question type for different subgroups of NFC, NES, and NFC/NES quadrants

	Mundane and regular Hours watching TV		**Mundane and irregular** Birthday Parties		**Salient and regular** Visiting a Hairdresser		**Salient and irregular** Days on Holiday	
	p<	eta	p<	eta	p<	eta	p<	eta
NFC								
1 low	.01	.389	n.s.	.068	.01	.201	.05	.108
2 high	.01	.294	n.s.	.048	.01	.217	n.s.	.087
NES								
1 low	.01	.359	n.s.	.077	.01	.136	.05	.109
2 high	.01	.322	n.s.	.043	.01	.278	n.s.	.087
NFC – NES								
Group 1*	.01	.409	.05	.136	.01	.151	n.s.	.102
Group 2	.01	.353	.01	.355	.01	.257	n.s.	.111
Group 3	.01	.267	n.s.	.133	n.s.	.122	n.s.	.102
Group 4	.01	.306	.01	.249	.01	.149	n.s.	.081

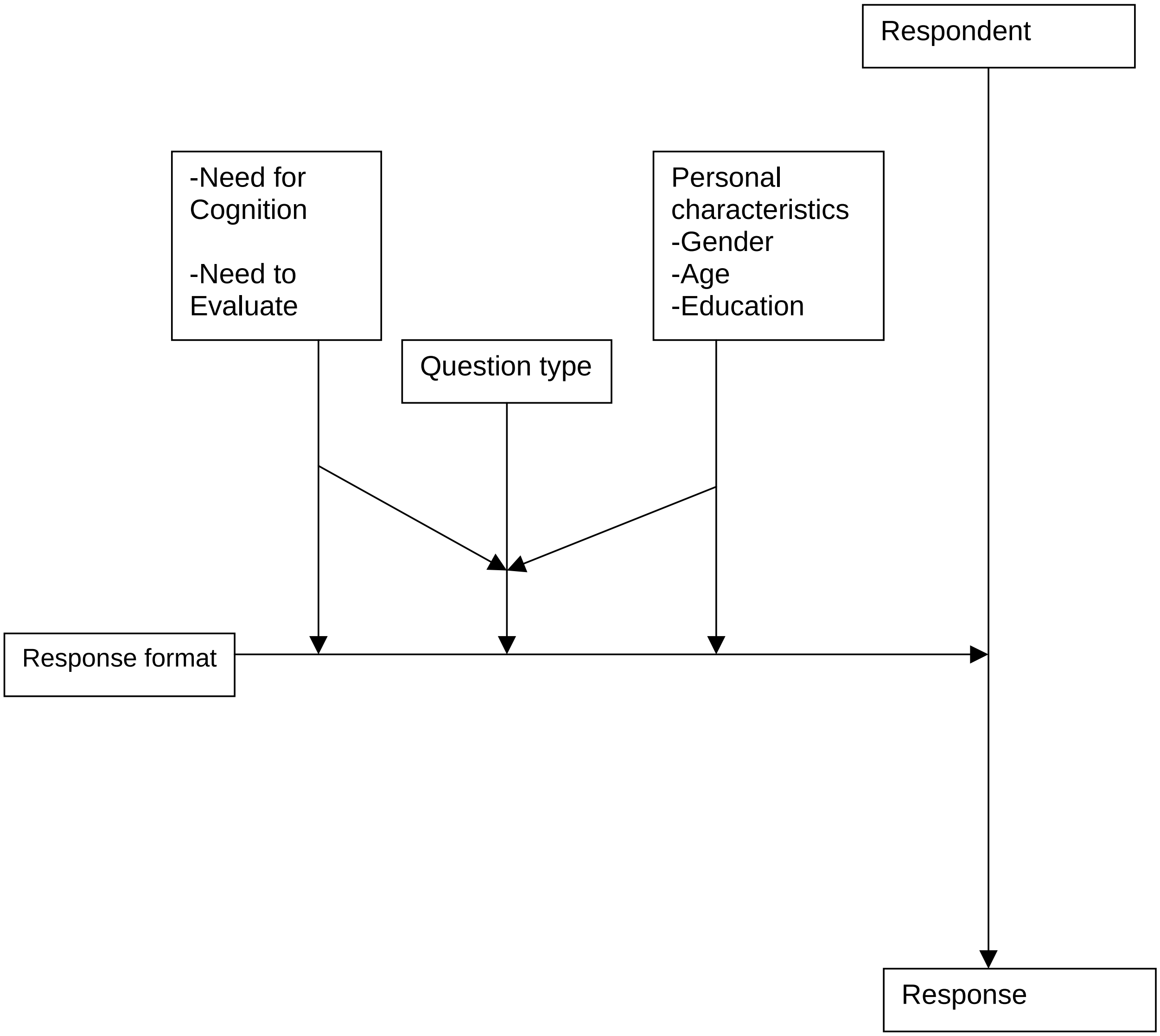
Note: A higher correlation coefficient (eta) between the answer score and the scale that was used indicates greater differences between response scales.
n.s.=non significant: there are no statistically significant differences between formats in this question type.
*See Table 3 for the definition of groups.

Table 7.
Overview of significance and association between the low and high response scale per question type for different subgroups of sex, age, and education

	Mundane and regular		**Mundane and irregular**		**Salient and regular**		**Salient and irregular**	
	Hours watching TV		Birthday Parties		Visiting a Hairdresser		Days on Holiday	
	p<	eta	p<	eta	p<	eta	p<	eta
Sex								
Male	.01	.331	.01	.191	.01	.165	.01	.099
Female	.01	.316	.01	.222	.01	.138	n.s.	.046
Age								
15-24	.01	.289	.01	.268	.05	.161	n.s.	.130
25-34	.01	.378	.01	.208	.05	.133	n.s.	.040
35-44	.01	.333	.05	.144	.01	.162	n.s.	.072
45-54	.01	.322	.01	.197	.05	.108	.05	.135
55-64	.01	.297	.01	.184	n.s.	.105	n.s.	.005
>64	.01	.313	.01	.225	.01	.241	n.s.	.066
Education								
Primary	.01	.341	.01	.336	n.s.	.072	n.s.	.079
Lower Secondary	.01	.326	.01	.194	.01	.178	.05	.115
Higher Secondary	.01	.285	.05	.159	n.s.	.071	n.s.	.047
Intermediate Vocational	.01	.395	.01	.146	.01	.189	.05	.126
Higher Vocational	.01	.344	.01	.264	.01	.141	n.s.	.015
University	.01	.294	.05	.171	.05	.171	n.s.	.096

Note: A higher correlation coefficient (eta) between the answer score and the scale that was used indicates greater differences between response scales.
n.s.=non significant: there are no statistically significant differences between formats in this question type.

Figure 1. Process model for formatting a response


Respondent
-Need for Cognition
-Need to Evaluate
Question type
Personal characteristics
-Gender
-Age
-Education
Response format
Response