No. 2008–51

DESIGN EFFECTS IN WEB SURVEYS: COMPARING TRAINED AND FRESH RESPONDENTS

By Vera Toepoel, Marcel Das, Arthur van Soest

May 2008

ISSN 0924-7815

Design Effects in Web Surveys: Comparing Trained and Fresh Respondents

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

Abstract In this paper we investigate whether there are differences in design effects between trained and fresh respondents. In three experiments, we varied the number of items on a screen, the choice of response categories, and the layout of a five point rating scale. We find that trained respondents are more sensitive to satisficing and select the first acceptable response option more often than fresh respondents. Fresh respondents show stronger effects with regard to verbal and non-verbal cues than trained respondents, suggesting that fresh respondents find it more difficult to answer questions and pay more attention to the details of the response scale in interpreting the question.

Keywords: professional respondents, questionnaire design, items per screen, response categories, layout

JEL codes: C81, C93

[*] 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 and Netspar.

1. Introduction

Socio-economic panel surveys, where the same households or individuals are interviewed repeatedly at various points in time, have important advantages over independent cross-sections, such as efficiency gains in recruiting, reduced sampling variation in the measurement of change, and the possibility to analyze behavior at the individual respondent level (see, e.g. Baltagi, 2001). However, the fact that experienced panelists may respond differently than panelists without experience ("panel conditioning"), raises concern over survey quality. In particular, many researchers fear that online survey panels, where respondents are interviewed at a high frequency such as once a month or more, create trained respondents.[1] Brannen (1993) suggests that the issue of the effects of surveying on respondents has been more a matter of speculation than of empirical investigation. Suggestions on how to treat trained respondents are increasing rapidly on the Internet (as shown by, e.g., searching the web for 'professional respondents' or 'data quality'; see also, for example, www.comscore.com, www.quirks.com, www.hisbonline.com). Although commercial companies address the issue of trained respondents in web surveys, there appears to be little empirical research to date on the effect of prior survey participation on survey answers.

Trained respondents may answer questions differently than those with little or no experience as panelist. This can be due to changes in behavior or knowledge induced by previous surveys (e.g. because respondents acquire knowledge on topics addressed in a previous survey), as well as to changes in the question-answering process. Panel members may learn from taking surveys. They may prepare for future surveys and increase their knowledge on the topics addressed, or develop attitudes

[1] In this paper we speak of trained or experienced respondents rather than professional respondents. The term 'professional' implies incentives (money) as a stimulus to participate, while in this paper we consider the effect of prior survey experience (training).

towards certain topics. In addition, they may become familiar with the question-answering process, learn how to interpret questions, and make fewer errors than new respondents. Or, conversely: experienced respondents may answer strategically to avoid follow-up questions and to reduce the burden of their task or accelerate the completion of the survey, thereby making more errors than fresh respondents.

This paper addresses the issue of procedural learning from taking surveys: the question-answering process. Trained respondents may react differently to web survey design choices than inexperienced respondents. First, they may be able to process more information on a screen and, for example, make fewer errors when multiple items are presented on a single screen. Second, they may be less or more susceptible to social desirability bias and less or more reluctant to select a response category that seems unusual in the range of responses. Third, they may react differently to (changes in) question layout. The goal of this study is to explore differences in web design effects between trained and fresh respondents in these three aspects.

The remainder of this paper is organized as follows. Section 2 addresses the background of design effects and panel conditioning, while Section 3 discusses the design and implementation of our experiments. Section 4 presents the results. This section is divided into three subsections to separately discuss each of the three experiments (items per screen, answer categories, and layout). In each subsection we discuss whether a design effect is found, to subsequently compare trained and fresh respondents with regard to this effect.

2. Background

Survey experience may influence responses to survey questions. In ongoing household panels, one could in principle test whether the time since respondents

entered the panel (the duration) or the number of surveys in which they have participated affects responses. However, in most panels almost none of the respondents are completely fresh, while the effect of panel experience may possibly be non-linear, with a noticeable difference between no and some experience, but much less or no effect when going from some to more experience. Bartels (1999) argues that panel surveys should routinely include parallel fresh cross-sections, to provide a solid basis to assess (and adjust for) biases arising from re-interviewing. In most panel surveys, comparable data from a fresh cross-section are not available.

Literature shows that answers to questions in (web) surveys are affected by design choices, such as the ordering of questions (see e.g. Couper, Traugott, and Lamias, 2001; Krosnick and Alwin, 1987; Toepoel, Das, and Van Soest, 2005), the categorical answers that the respondent can choose from (see e.g. Rockwood, Sangster, and Dillman, 1997; Schwarz et al., 1985), or the layout of the questions (see e.g. Christian, 2003; Christian and Dillman, 2004; Dillman and Christian, 2002; Toepoel, Das, and Van Soest, 2006; Winter 2002a, 2002b). Some studies have also analyzed whether such design effects vary with respondent characteristics such as age, gender, or education level (see e.g. Fuchs, 2005; Knauper, Schwarz, and Park, 2004; Krosnick and Alwin, 1987; Stern, Dillman, and Smyth, 2007; Tourangeau, Couper and Conrad, 2007), or attitudes such as a *need for cognition* or *need to evaluate* (see e.g. Toepoel et al., 2006). Despite the growing empirical support for (web) design effects, there exists virtually no reference to respondents' experience in answering surveys. As a result, empirical tests have not taken into account how experience may affect the question-answering process in web surveys. In this study, we analyze the differences in web design effects between experienced and fresh panel respondents.

2.1 Experience and the response process

Van der Zouwen and Van Tilburg (2001) find that panel conditioning effects sometimes arise and sometimes not, without a clear indication of the situations in which these effects occur. Trivellato (1999) concludes that panel participation mainly affects the way in which behavior is reported (response process), while it does not have pervasive effects on behavior itself. Coombs (1973) and Das, Toepoel, and Van Soest (2007) find that panel conditioning arises for knowledge questions, but not for other types of questions. Sturgis, Allum, and Brunton-Smith (2007) formulate a theory for panel conditioning: the cognitive stimulus hypothesis. Questions about certain topics may induce respondents to reflect on them after the survey has ended, to talk about them with friends and relatives, and to acquire additional information. Golob (1990) concludes that no panel conditioning effects exist in questions that require simple reporting tasks, suggesting instead that panel conditioning relates to the cognitive difficulty in answering questions. He finds no panel conditioning on car ownership variables that are measured using simple reporting requirements, but he does find panel conditioning effects for more cognitively demanding questions such as travel times for different modes of transport. Van der Zouwen and Van Tilburg (2001), on the other hand, conclude that panel conditioning does not take place through cognitive processes within the respondent's mind but through the task-related behavior of the interviewer.

Mathiowetz and Lair (1994) find evidence that respondents become familiar with the question-answering process and adjust their responses accordingly. They hypothesize that an improvement in daily life activities noted in a subsequent survey wave was a function of panel conditioning. Respondents learned in wave 1 that for every reported difficulty there was a series of follow-up questions, and they therefore

altered their responses in the subsequent wave to avoid the follow-up questions. Meurs, Van Wissen, and Visser (1989) also find that experienced respondents respond strategically, for instance after learning that answering "no" means evading follow-up questions, thereby reducing the burden of their task.

Trained respondents may be more sensitive to social desirability bias than fresh respondents. Sharpe and Gilbert (1998) find that repeated testing increases the scores on the Beck depression scale and attribute this to socially desirable response behavior, triggered by the first interview. Chan and McDermott (2007) and Wang, Cantor, and Safir (2000) find similar effects.

Coen, Lorch, and Piekarski (2005) compare frequent and non-frequent respondents. They find evidence that responses of frequent responders are more in line with actual consumer behavior than responses of less frequent responders. This finding is in contrast to the conventional view that past experience is not desirable with regard to measurement errors (Bartels, 1999; Brannen, 1993; Golob, 1990; Mathiowetz and Lair, 1994; Meurs et. al, 1989; Sharpe and Gilbert, 1998; Sturgis et al., 2007; Williams, 1970; Williams and Mallows, 1970). Coen et al. (2005) find no evidence that frequent responders try to speed through the survey. In fact they find a relatively high number of marks on check-all-that-apply questions. Inexperienced panelists more often choose socially desirable answers. This is in line with the results of Dennis (2001). Coen et al. also demonstrate that experience (number of surveys completed) is more associated with response behavior than duration (the length of time on the panel).

2.2 Experience and web survey design

There is a growing literature that suggests that the design of a web survey has a significant impact on measurement error (see e.g. Christian and Dillman, 2004; Couper et al., 2001; Dillman, 2007; Dillman and Christian, 2002; Tourangeau, Couper, and Conrad, 2004, 2007). Design may be more important in web surveys than in other modes of administration, because there are many tools available and because of potential variation in how the survey appears on a screen. Couper (2000) concludes that more work is needed to determine the optimal designs for different groups of people, emphasizing the need for research on panel conditioning and web page design.

Despite the widespread use of online panels, there appears to be no empirical research to date on the difference in response effects between trained and fresh respondents. There are some papers offering suggestions on questionnaire design in relation to prior survey experience in general. Trivellato (1999), for example, offers a number of strategies with regard to initial and follow-up sampling, panel length and number of waves, and to tracking and tracing techniques to locate respondents to maintain high participation rates. Moreover, he outlines questionnaire design strategies such as the sequence of questions, probing, skip patterns, and consistency checks to limit response errors. He also recommends a low-frequency measuring of variables that are reasonably stable over time, preferably in the first interview. Web surveys are particularly suited to implementing Trivellato's suggestions thereby improving the longitudinal consistency of the data. This paper addresses three design issues in which trained and fresh respondents may differ.

2.2.1 Items per screen

For web questionnaires, interface design varies in terms of the distribution of questions on the screen and the navigation methods used. At one end of the design continuum are form-based designs that present questionnaires as one long form in a scrollable window, at the other end are screen-by-screen questionnaires that present only a single item at a time (Norman, et al., 2001). Presenting questions in a matrix is somewhere in between, reducing the number of screens without the need for scrolling.

The grouping of related items on a single screen is likely to lead respondents to view the items as related entities, thus increasing the correlation among them (Dillman, 2007; Schwarz and Sudman, 1996; Strack, Schwarz, and Wanke, 1991; Sudman, Bradburn, and Schwarz, 1996; Tourangeau et al. 2004, 2007). Couper et al. (2001) conclude that correlations are consistently higher among items appearing together on a screen than among items separated across several screens. However, the overall effect is not large, and none of the differences between pairs of correlations reach statistical significance. Tourangeau et al. (2004) replicate the above findings. Respondents seem to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. Peytchev et al. (2006) find few differences between paging and scrolling designs.

Non-response and time to complete the interview can also be indicators for the optimal number of items per screen. Lozar Manfreda, Batagelj, and Vehovar (2002) find that a one-page design results in higher item non-response. Couper et al. (2001), Lozar Manfreda et al. (2002), and Tourangeau et al. (2004) find that a multiple-item-per-screen design takes less time to complete than a one-item-per-screen design. Evaluation questions can show whether respondents are comfortable with a particular survey design. Toepoel et al. (2005) find that placing more items on a screen negatively influences the respondent's evaluation of the layout.

We are not aware of any studies on the optimal number of items on a screen in relation to survey experience. Our conjecture is that trained respondents can process more information on a screen, thus showing less item non-response when more items are placed on a single screen than fresh respondents. We expect them to complete the survey faster than fresh respondents, especially if many items are placed on a screen. We also expect them to better evaluate a large number of items on a screen than fresh respondents.

2.2.2 Response categories

Studies on the cognitive and communicative processes involved in answering survey questions suggest that the choice of response categories can have a significant effect on the answers. Toepoel et al. (2006) and Winter (2002a; 2002b) find response category effects in web surveys, while Krosnick and Alwin (1987), Rockwood et al. (1997), Schwarz et al. (1985), Schwarz and Hippler (1987), and Strack and Martin (1987) find effects in other modes of administration. Schwarz and Hippler (1987) argue that respondents use the response alternatives to determine the meaning of the question and use the frequency range as a frame of reference, presuming the values stated in the scale to be commonly held values. In other words, a respondent may be reluctant to select a response category that seems unusual in the range of responses. This results in higher estimates along scales that present high rather than low ranges. The literature suggests that response categories have a significant effect on responses to questions for which estimation is likely to be used in recall, whereas in questions in which direct recall is used in response formatting the response categories do not have a significant effect.

Choquette and Hesselbrock (1987) suggest that respondents attempt to present themselves more favorably in later waves. This would lead to the conjecture that trained respondents are more prone to social desirability bias and more reluctant to select a response category that seems unusual in the range of responses. On the other hand, Coen et al. (2005) and Dennis (2001) find that inexperienced panelists more often choose socially desirable answers. Survey experience may also make the respondents less uncertain and thus less susceptible to social desirability bias. The second experiment in this paper assesses the impact of a response scale on both trained and fresh respondents.

2.2.3 Layout

Differences in question layout can lead to detectable differences in responses to survey questions (see, e.g. Christian, 2003; Christian and Dillman, 2004; Dillman and Christian, 2002; Schwarz and Hippler, 1987; Toepoel et al., 2006; Tourangeau et al., 2004). A question format contains verbal and nonverbal cues that influence respondent behavior. Nonverbal cues include graphical, numerical and symbolic languages that convey meaning in addition to the verbal language (Dillman and Christian, 2002). Jenkins and Dillman (1997) have developed a conceptual framework to explain how visual languages may influence respondent behavior.

Redline et al. (2003) confirm that the visual and verbal complexity of information in a questionnaire affects what respondents read, the order in which they read it, and ultimately, their comprehension of the information. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales (graphical manipulation) in paper questionnaires do not elicit the same responses. Schwarz et al. (1985) show that respondents gain information about the researcher's

expectations using numerical labels as frames of reference. Schwarz et al. (1991) find that changing the numerical values attached to scales changes the answers, and that respondents hesitate to assign a negative score to themselves in a face-to-face interview: a scale with numbers 0-10 results in lower scores than a –5 to 5 format.

We expect trained panelists to be more sensitive to layout choices than fresh panelists. They may be used to a particular question format so that changing that format (e.g. from disagree-agree to agree-disagree) may not be noticed. In addition, we expect them to be more sensitive to added numerical labels and signs than fresh respondents.

3. Design and implementation

To study design effects on trained and fresh respondents, we used two online household panels administered by CentERdata. The first, the CentERpanel (see also http://www.centerdata.nl/en/CentERpanel), has existed for 17 years. Panel members fill out questionnaires every week. Panel duration of respondents ranges from seventeen years to a few months. The second panel is the LISS-panel (see http://www.centerdata.nl/en/LISSpanel). Our experiments were the very first questionnaire for this panel. Both panels are designed to be representative of the Dutch population. Thus, the CentERpanel consists of trained respondents (varying in panel duration, with a mean duration of 6 years and 8 months, standard deviation equals 4 years), while the LISS-panel consists of completely fresh respondents.

We fielded the questionnaire in June 2007. In the CentERpanel, 1356 panel members were selected to fill out the questionnaire; 981 respondents (72.3%) responded. In the LISS-panel, 4530 panel members were selected; 2809 respondents

(62.0%) filled out the questionnaire. To correct for differences due to non-response, we used weights based on gender, age and education.

The questionnaire consisted of three different experiments. In the first, we used the Marlowe-Crowne Social Desirability Scale (the 10-item version of Strahan and Gerbrasi, 1972) and varied the number of items per screen. We used three groups, with 1, 5, and 10 items per screen. We added some questions to determine whether respondents react differently to the number of items displayed per screen. In the second experiment we varied the answer scale in four questions. We used the same questions as Toepoel et al. (2006), varying in cognitive difficulty. We used a low response scale, a high response scale, and an open-ended format. In the third experiment we varied the layout of a five-point rating scale. The first group was presented answer categories in a linear vertical format from positive to negative (excellent, very good, good, fair, and poor). Five other groups were presented with different manipulations. The second group answered from negative to positive, the third in a horizontal format, for the fourth group we added numbers 1 to 5 to the response categories, for the fifth group numbers 5 to 1, and for the sixth group numbers 2 to –2.

4. Results

In this section we discuss the results of the three experiments. For each experiment, we first discuss the response effect and then compare the answers of trained and fresh respondents.

4.1.1 Response effect: items per screen

We found differences in inter-item correlations when the items were presented (1) one-item-per-screen (Cronbach's alpha of .473 for the trained panel and .528 for the fresh panel), (2) 5-items-per-screen (alpha of .602 for the trained panel and .516 for the fresh panel), and (3) 10-items-per-screen (alpha of .515 for the trained panel and .498 for the fresh panel).

In principle, the web survey software can force the respondent to give a response. If a respondent fails to give an answer, he/she would then be presented with an error message indicating a need to choose an answer. We deliberately did not program this feature, so that respondents could proceed without filling in answers. We found no significant differences in item non-response when more items were placed on a single screen in the trained panel. In the fresh panel, the more items were placed on a single screen, the lower the item non-response ($F=3.795$, $p=.023$). This is contrary to the findings of Lozar Manfreda et al. (2002).

If more items are placed together on one screen, fewer physical actions (keystrokes or mouse clicks) are required than when items are presented separately. Therefore, we expected that placing more items on a single screen would reduce the time needed to complete the questionnaire. However, we found no significant differences in mean duration[2] between formats (1, 5, and 10 items per screen) in both panels.

Respondents answered some evaluation questions about the social desirability questions:

1. How interesting did you find the questions?
2. How would you evaluate the duration?
3. How clear did you find the wording of the questions?

[2] Means were calculated after deleting outliers with more than 2 times the standard deviation (28 respondents in the fresh panel and 4 respondents in the trained panel).

4. How easy was it to answer the questions?
5. What did you think of the layout?
6. What is your overall opinion of these questions?

These questions were asked on a ten-point scale ranging from 1 ('very poor'/'not at all') to 10 ('very good'/'very much'). In the trained panel we found a significant effect of format in question 4, with the 5-items-per-screen format receiving the highest rating ($F=3.32$, $p=.037$). This suggests that respondents found that the 10-items-per-screen format contained too much information, while the 1-item-per-screen format contained too many screens. The fresh panel also preferred the layout of the 5-items-per-screen format to other formats ($F=3.816$, $p=.022$).

The counting of all ten social desirability items resulted in an overall score of social desirability. Neither the trained nor the fresh panel showed differences in social desirability scores between the 1, 5, and 10-item-per-screen format.

4.1.2 Comparison of trained and fresh respondents: items per screen

Trained respondents had higher inter-item correlations for multiple-items-per-screen formats, while fresh respondents showed the highest inter-item correlation in the one-item-per-screen version. Trained panelists seem to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. Fresh panelists may be triggered by the new experience of participating in a survey and therefore read each item more carefully.

We found no significant difference in item non-response between trained and fresh respondents; 1.2% (12 out of 981 respondents) had one or more items missing in the trained panel, compared to 1.5% (42 out of 2809 respondents) in the fresh panel. Linear regression of item non-response on the number of items per screen, a dummy

for panel (trained versus fresh), and the interaction between these two showed no significant interaction effect.

There was a difference in mean duration of the entire survey[3] between panels (t=-2.4, p=.016): 436 seconds for the trained panel and 576 seconds for the fresh panel. The mean duration to complete just the ten social desirability items did not differ significantly between panels. Linear regression of the duration of the survey on the number of items per screen, a dummy for panel, and the interaction between these two showed no significant interaction effect either.

Although this paper discusses design effects, we also looked at the mean score of the Social Desirability Scale used for the items-per-screen experiment. In contrast to Choquette and Hesselbrock (1987), we found no evidence for social desirability bias for trained respondents. The mean scores of the Social Desirability Scale in the two panels were not significantly different (F=2.16, p=.642).

4.2.1 Response effect: response categories

To assess the impact of a response scale on respondents' answers, we asked four questions on the frequency of various activities with a randomized answering format: a low response scale, a high response scale, and an open-ended format. See Appendix A for the questions and response scales used. We dichotomized answers to compare the results.

We found a scale range effect (see Tourangeau, Rips, and Ransinki, 2000, p. 249): the range of the response scale affected respondents' frequency reports. Table 1 shows that 20% of the trained respondents who were presented the low response scale reported watching TV for more than two and a half hours, compared to 51% of the

[3] The questionnaire consisted of all experiments discussed in this paper.

trained respondents who were presented the high response scale. In comparison, 46% of the trained respondents who were presented the open-ended question reported watching TV for more than two and a half hours. Similar results were found for the fresh panel.

[Insert table 1 about here]

Table 2 shows an overview of the correlations between answer score (1 if more than the reference level, 0 otherwise) and response format for the different question types. A higher correlation coefficient (eta) between the answer score and the scale used indicates a larger effect of the response scale. With the high versus low response scale, the largest correlation between the answer score and the scale is found in hours watching TV (difficult to process), the lowest for days on holiday (easy to process). As expected, the effect of response scales depends on how well a behavior is presented in memory. More details of this experiment on response category effects can be found in Toepoel et al. (2006).

[Insert table 2 about here]

4.2.2 Comparison of trained and fresh respondents: response categories

We found an effect of response categories on answers, but this effect is not significantly different for trained and fresh respondents. For none of the questions we found a significant interaction effect between format and panel. Our conjecture that trained respondents are more prone to social desirability bias and more reluctant to select a response category that seems unusual in the range of responses was not

confirmed. The conjecture that survey experience may make the respondents less uncertain and thus less susceptible to social desirability bias was not confirmed either.

4.3.1 Response effect: layout

In our third experiment, we manipulated the layout of a five-point rating scale using verbal and non-verbal manipulations. Appendix B presents the question that was asked and shows the answer distributions for all formats for both panels. Table 3 shows that the distributions of the answers in a negative-positive format differ significantly from those in a positive-negative format (verbal manipulation: 1 versus 2). Respondents selected the response option 'very good' less often when it was presented as a fourth alternative. No significant differences were found for the graphical manipulation (changing the layout from vertical to horizontal), i.e., comparing format 1 versus 3. Comparing adding numbers 1 to 5 to the scale did not lead to significant differences in answer scores either, suggesting that respondents take a numbering beginning with 1 as a kind of default labeling that does not convey much information about the meaning of the scale points. Adding the numbers 5 to 1 to 1 to 5 (formats 4 and 5) did produce significant differences, indicating that respondents react to numbers as well as words in a numerical ordering not beginning with 1. The strongest effect was found when numbers 2 to -2 were added. This manipulation showed significantly different answer scores compared to all other manipulations. Respondents are apparently reluctant to assign negative scores. Negative numbers might be interpreted as implying more extreme judgments than low positive numbers (scale label effect, see Tourangeau et al., 2000, p.248; see also Tourangeau et al., 2007, who make a similar argument and provide additional evidence for the added attention that negative signs receive).

A Chi Square test and a difference of means test showed significant differences for all non-verbal manipulations (all formats except format 2), indicating that the layout of the answer categories influences the answers. Also, the overall test comparing all six formats showed significant differences between formats.

[Insert table 3 about here]

4.3.2 Comparison of trained and fresh respondents: layout

Although the third response option 'good' has the same number (3) in formats 4 and 5, fresh respondents selected this answer significantly more often in format 4 (numbers 1 to 5: 53.4%) than in format 5 (numbers 5 to 1: 44.3%). The effect for trained respondents was much smaller. Apparently, fresh respondents extract information not only from the number itself but also from the ordering of numbers added to the verbal labels.

Although changing the layout from vertical to horizontal did not change the answer distributions significantly (see Table 3: 1 versus 3), trained respondents selected the second response 'very good' more frequently than fresh respondents. The fresh respondents selected the response option 'fair' more often in the horizontal format. This indicates a primacy effect for trained respondents and a recency effect for fresh respondents.

Combining all six formats and looking at the distribution of all answers, independent of the layout manipulations, we found a similar result: trained respondents more easily selected one of the first options, while fresh respondents more often selected one of the last options (χ^2=14.93, p=.01). A possible interpretation of this difference is that trained respondents are more sensitive to

satisficing and therefore select the first satisfying response category more often (cf. Krosnick and Alwin, 1987; and Tourangeau et al., 2000).

Linear regression explaining the answer to the question by dummies for the five format manipulations (with format 1 as reference level), a panel dummy, and interaction terms between the panel dummy and the five formats showed no significant interaction effect between panel experience and the five formats. However, the interaction effect between the panel dummy and the graphical manipulation (horizontal format) almost reached significance (t=1.83, p=.07).

5. Discussion and Conclusions

Despite the growing empirical support for (web) design effects, there exists virtually no reference to respondents' experience in completing surveys. This means that empirical tests have not taken into account how experience may affect the question-answering process in web surveys. We have tried to gain more insight into the response processes of trained and fresh respondents. We did so by conducting three experiments on web survey design issues with two different panels: a new panel of fresh respondents, and a panel that has been in place for seventeen years now, thus consisting of respondents that have extensive experience. The web survey design issues we considered were the number of items per screen, response category effects, and layout effects.

First of all, the social desirability scale used to assess the impact of a 1, 5, and 10-item-per-screen format showed no difference in social desirability scores between the trained and fresh panel. A small effect with respect to inter-item correlations for multiple-items-per-screen formats was found, indicating that trained panelists use the proximity of the items as a cue to their meaning more than fresh panelists do. We did

not find evidence that trained respondents are able to process more information on a screen, that is, that they show less item non-response when more items are placed on a single screen. They did complete the survey in less time than fresh respondents. Our analysis showed no interaction effect between the number of items per screen and panel experience on item non-response, time to complete the survey, and evaluation questions. We did not find evidence that the number of items per screen influences the answers respondents provide, but it does have an influence on respondents' evaluation of the questionnaire. Both the trained and the fresh panelists appreciated the 5-items-per-screen format the most. Keeping the respondent satisfied is important for panel maintenance, and therefore it is important to place more than one item of a battery on a screen, but not too many.

With regard to response category effects, we found no significant interaction effect between web survey design and panel experience either; our conjecture that trained respondents are more prone to social desirability bias and more reluctant to select a response category that seems unusual in the range of responses is not confirmed, but neither is the conjecture that survey experience may make the respondents less uncertain and thus less susceptible to social desirability bias.

Fresh panelists showed stronger effects than trained respondents with regard to the verbal and non-verbal cues in a five-point scale. We found no significant interactions between panel experience and layout manipulations. Our results show a primacy effect for trained respondents and a recency effect for fresh respondents, suggesting that trained respondents more often select the first acceptable response option than fresh respondents.

In summary, we found some evidence that survey experience influences the question-answering process. Trained respondents seem to be more sensitive to

satisficing. The advantage of using trained respondents is that they are less sensitive to visual cues. Fresh respondents show stronger effects for details of the response scales than trained respondents, even though some features may simply be a matter of style rather than adding any meaning to the scale. They may be more uncertain which answer to select and therefore base their answers more often on cues in a questionnaire (see also Tourangeau et al., 2007, who make a similar argument for the greater impact of nonverbal cues for ambiguous questions). Survey researchers should pay attention to these differences between trained and fresh respondents, and additional research is needed to determine whether these conclusions hold in different settings.

6. References

Baltagi, Badi H. 2001. *Econometric Analysis of Panel Data*. Wiley: Chichester.

Bartels, Larry M. 1999. "Panel Effects in the American National Election Studies." *Political Analysis*, 8, 1-20.

Brannen, Julia. 1993. "The effects of research on participants: Findings from a study of mothers and employment." *Sociological Review*, 41, 328-346.

Chan, Jason C. and Kathleen B. McDermott. 2007. "The Testing Effect in Recognition Memory: A Dual Process Account." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 431-437.

Choquette, Keith A. and Michie N. Hesselbrock. 1987. "Effects of Retesting with the Beck and Zung depression scales in alcoholics." *Alcohol and Alcoholism,* 22, 277-283.

Christian, Leah, M. 2003. "The Influence of Visual Layout on Scalar Questions in Web Surveys. Unpublished Master's Thesis." Retrieved 03-01-2007 on http://survey.sesrc.wsu.edu/dillman/papers.htm.

Christian, Leah M. and Don A. Dillman. 2004. "The influence of graphical and symbolic language manipulations to self-administered questions." *Public Opinion Quarterly,* 68, 57-80.

Coen, Terrence, Jacqueline Lorch, and Linda Piekarski. 2005. "The Effects of Survey Frequency on Panelists' Responses." ESOMAR retrieved July 27 2007 on www.websm.org.

Coombs, Lologene C. 1973. "Problems of Contamination in Panel Surveys: A Brief Report on an Independent Sample; Taiwan 1970." *Studies in Family Planning,* 4, 257-261.

Couper, Mick P. 2000. "Web Surveys. A review of issues and approaches." *Public Opinion Quarterly*, 64, 464-494.

Couper, Mick P. , Michael W. Traugott, and Mark J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly,* 65, 230-253.

Das, Marcel, Vera Toepoel, and Arthur van Soest. 2007. "Can I Use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys." CentER Discussion Paper 2007-56, CentER: Tilburg.

Dennis, Michael. 2001. "Are Internet Panels Creating Professional Respondents?" *Marketing Research*, 13, 34-39.

Dillman, Don. A. 2007. *Mail and Internet Surveys. The Tailored Design Method.* Wiley: Hoboken NJ.

Dillman, Don A. and Leah Christian. 2002. "The Influence of Words, Symbols, Numbers, and Graphics on Answers to Self-Administered Questionnaires: Results from 18 Experimental Comparisons." Retrieved 01-03-2007 on http://survey.sesrc.wsu.edu/dillman/papers.htm.

Friedman, Linda W., and Hershey H. Friedman. 1994. "A comparison of vertical and horizontal rating scales." *The Mid-Atlantic Journal of Business,* 30, 107-202.

Fuchs, Marek. 2005. "Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options." *Journal of Official Statistics*, 21, 701-725.

Golob, Thomas F. 1990. "The Dynamics of Household Travel Time Expenditures and Car Ownership Decisions." *Transportation Research*, 24A, 443-463.

Jenkins, Cleo R. and Don A. Dillman. 1997. " Towards a theory of Self-administered Questionnaire Design." P. 165-196, in *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. Wiley series in probability and statistics: New York.

Knauper, Barbel, Norbert Schwarz, and Denise Park. 2004. "Frequency Reports Across Age Groups." *Journal of Official Statistics*, 20, 91-96.

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.

Lozar Manfreda, Katja, Zenel Batagelj, and Vasja Vehovar. 2002. "Design of Web survey questionnaires: Three basic experiments." *Journal of Computer-Mediated Communication* 7, 3 (http://jcmc.indiana.edu/vol7/issue3/vehovar.html).

Mathiowetz, Nancy A. and Tamra J. Lair. 1994. "Getting better? Changes or Errors in the Measurement of Functional Limitations." *Journal of Economic & Social Measurement*, 20, 237-262.

Meurs, Henk, Leo van Wissen, and Jacqueline Visser. 1989. "Measurement Biases in Panel Data." *Transportation*, 16, 175-194.

Norman, Kent L., and Zachary Friedman, Kirk Norman, and Rod Stevenson. 2001. "Navigational Issues in the Design of On-Line Self-Administered Questionnaires." *Behavior & Information Technology*, 20, 37-45.

Peytchev, Andy, Mick P. Couper, Sean Esteban McCabe, and Scott D. Crawford. 2006. "Web Survey design. Paging Versus scrolling." *Public Opinion Quarterly,* 70, 596-607.

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 01-03-2007 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." *Sociological Methods and Research*, 26, 118-140.

Schwarz, Norbert, Barbel Knauper, Hans-J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly*, 55, 570-582.

Schwarz, Norbert and Seymour Sudman. 1996. *Answering Questions*. Jossey-Bass Publishers: San Francisco.

Schwarz, Norbert and Hans-J. Hippler 1987. "What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives." Pp. 163-178, in *Social Information Processing and Survey Methodology*, edited by H.-J. Hippler, N. Schwarz, and S. Sudman. New York: Springer-Verlag.

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

Sharpe, J. Patrick and David G. Gilbert. 1998. "Effects of Repeated Administration of the Beck Depression Inventory and Other Measures of Negative Mood States."*Personal Individual Differences*, 24, 457-463.

Stern, Michael J., Don A. Dillman, Jolene D. Smyth. 2007. "Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey." *Survey Research Methods*, 1, 121-138.

Strack, Fritz and Leonard L. Martin. 1987. "Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys." Pp. 123-148, in *Social Information Processing and Survey Methodology,* edited by H.-J. Hippler, N. Schwarz, and S. Sudman. New York: Springer-Verlag.

Strack, Fritz, Norbert Schwarz, and Michaela Wanke. 1991. "Semantic and Pragmatic Aspects of Context Effects in Social and Psychological Research." *Social Cognition*, 9, 111-125.

Strahan, Robert and Kathleen C. Gerbrasi. 1972. "Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale." *Journal of Clinical Psychology*, 28, 191-193.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith. 2007. "Attitudes Over Time: The Psychology of Panel Conditioning." Pp. 1-13, in *Methodology in Longitudinal Surveys*, edited by P. Lynn. Wiley: Chichester.

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

Toepoel, Vera, Marcel Das, and Arthur van Soest. 2006. "Design of Web Questionnaires: the Effect of Layout in Rating Scales." CentER Discussion Paper 2006-30, CentER: Tilburg.

Toepoel, Vera, Marcel Das, and Arthur van Soest. 2005. "Design of Web Questionnaires: a Test for Number of Items per Screen." CentER Discussion Paper 2005-114, CentER: Tilburg.

Toepoel, Vera, Corrie Vis, Marcel Das, and Arthur van Soest. 2006. "Design of Web Questionnaires: an Information-Processing Perspective for the Effect of Response Categories." CentER Discussion Paper 2006-19, CentER: Tilburg.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2007. "Color, Labels, and Interpretive heuristics for response scales." *Public Opinion Quarterly,* 71, 91-112.

Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2004. "Spacing, position, and order. Interpretive heuristics for visual features of survey questions." *Public Opinion Quarterly,* 68, 368-393.

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

Trivellato, Ugo. 1999. “Issues in the Design and Analysis of Panel Studies: A Cursory Review.” *Quality & Quantity*, 33, 339-352.

Van der Zouwen, Johannes and Theo van Tilburg. 2001. “Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses.” *Sociological Methods & Research*, 30, 35-56.

Wang, Kevin, David Cantor, and Adam Safir. 2000. “Panel conditioning in a Random Digit Dial Survey.” *Proceedings of the Section on Survey Research Methods*, 822-827.

Williams, William H. 1970. “The Systematic Bias Effects of Incomplete Responses in Rotation Samples.” *Public Opinion Quarterly*, 33, 593-602.

Williams, William H. and Colin L. Mallows. 1970. “Systematic Biases in Panel Surveys Due to Differential Nonresponse.” *Journal of the American Statistical Association*, 65, 1338-1349.

Winter, Joachim K. 2002a. “Bracketing Effects in Categorized Survey Questions and the Measurement of Economic Quantities.” Discussion Paper No. 02-35, Sonderforschungsbereich 504, University of Mannheim. Retrieved 01-03-2007 on http://www.sfb504.uni-mannheim.de/publications/dp02-35.pdf.

——— 2002b. “Design Effects in Survey-Based Measures of Household Consumption.” Discussion Paper No. 02-34, Sonderforschungsbereich 504, University of Mannheim. Retrieved 01-03-2007 on http://www.sfb504.uni-Mannheim.de/publications/dp02-34.pdf.

Appendix A: Questions and Answer Categories in Experiment 2: Response Category Effects.

Response scales	Format A	Format B	Format C
How many hours do you typically watch 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	
How many birthday parties do you typically attend per year?			
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	
How many times did you go to the hairdresser last year?			
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	
How many days did you leave your home (have a holiday) last year?			
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 one to five in Format A match answer category one in Format B. Answer category six in Format A matches answer categories two to six in Format B.

Appendix B: Frequencies (in %), Number of Observations, and Mean Scores in Experiment 3: Layout Effects. Fresh panel between parentheses.

Question:
Overall, how would you rate the quality of education in the Netherlands?

%	1 Reference: Linear Vertical Positive to Negative	2 Verbal: Linear Vertical Negative to Positive	3 Graphical: Linear Horizontal	4 Numerical: Linear Vertical With Numbers 1 to 5	5 Numerical: Linear Vertical With Numbers 5 to 1	6 Numerical: Linear Vertical With Numbers 2 to -2
Excellent	.0 (.2)	1.6 (.0)	.0 (.5)	.8 (.0)	.6 (.2)	3.9 (.9)
Very Good	14.2 (11.1)	5.2 (4.5)	19.9 (8.6)	15.4 (9.3)	9.0 (7.6)	19.9 (13.3)
Good	42.1 (46.5)	49.4 (40.5)	40.7 (43.8)	46.5 (53.4)	42.2 (44.3)	45.8 (50.2)
Fair	36.8 (37.0)	33.8 (48.2)	34.9 (41.0)	33.8 (31.9)	38.1 (39.8)	25.1 (29.2)
Poor	6.9 (5.3)	10.1 (6.8)	4.5 (6.1)	3.5 (5.5)	10.1 (8.1)	5.3 (6.5)
N	162 (453)	181 (460)	159 (460)	172 (474)	138 (466)	162 (483)
Mean	3.36 (3.36)	3.46 (3.57)	3.24 (3.44)	3.24 (3.34)	3.48 (3.48)	3.08 (3.27)

Note: Scores for all versions are transformed back to the reference layout. Thus, a high mean score indicates a negative judgment.

Table 1. Overview of Frequencies (%) from Different Response Formats for the Trained and Fresh Panel.

	Low response scale		High response scale		Open-ended	
	Trained panel	Fresh panel	Trained panel	Fresh panel	Trained panel	Fresh panel
	more than X*	more than X*	more than X*	more than X*	more than X*	more than X*
Hours watching TV	20	18	51	49	46	44
Birthday parties	24	28	40	41	42	44
Visiting a hairdresser	14	17	28	33	25	21
Days on holiday	35	41	44	45	45	43

*X=two and a half for hours watching TV, nine for visiting a hairdresser, and 17 for birthday parties and days on holiday.

Table 2. Overview of Correlations between Answer Score and Response Format.

	High response scale versus low response scale		Low response scale versus open-ended		High response scale versus open-ended	
	Trained panel	Fresh panel	Trained panel	Fresh panel	Trained panel	Fresh panel
	eta	eta	eta	eta	eta	eta
Hours watching TV	.329 (p<.0001)	.325 (p<.0001)	.267 (p<.0001)	.243 (p<.0001)	.062 (p=.137)	.067 (p<.0001)
Birthday parties	.168 (p<.0001)	.137 (p<.0001)	.505 (p<.0001)	.482 (p<.0001)	.352 (p<.0001)	.358 (p<.0001)
Visiting a hairdresser	.182 (p<.0001)	.180 (p<.0001)	.136 (p=.002)	.044 (p=.001)	.045 (p=.225)	.133 (p<.0001)
Days on holiday	.089 (p=.056)	.050 (p=.102)	.097 (p=.036)	.019 (p=.548)	.008 (p=.830)	.031 (p=.348)

Note: A higher correlation coefficient (eta) between the answer score and the scale that was used indicates greater differences between response scales.

Table 3. Chi Square Tests and Differences of Means in the Different Manipulations.

	Trained panel		Fresh panel	
	Chi Square Tests χ^2	Diff. Of means t	Chi Square Tests χ^2	Diff. Of means t
Verbal: 1 versus 2	13.901 (p=.016)	1.311 (p=.253)	23.430 (p<.0001)	14.834 (p<.0001)
Graphical: 1 versus 3	2.557 (p=.634)	1.829 (p=.177)	3.492 (p=.625)	1.594 (p=.207)
Numerical: 1 versus 4	4.477 (p=.483)	1.757 (p=.186)	5.743 (p=.332)	.310 (p=.578)
Numerical: 4 versus 5	9.082 (p=.059)	7.081 (p=.008)	13.424 (p=.020)	9.509 (p=.002)
Numerical: 5 versus 6	16.337 (p=.006)	17.361 (p<.0001)	30.988 (p<.0001)	27.091 (p<.0001)
Overall across all non-verbal manipulations (except 2)	37.727 (p=.010)	F=5.399 (p<.0001)	67.840 (p<.0001)	F=8.871 (p<.0001)
Overall across all 6 formats	55.618 (p<.0001)	F=5.944 (p<.0001)	102.906 (p<.0001)	F=11.943 (p<.0001)

Note:
1 Reference: Linear Vertical Positive to Negative
2 Verbal: Linear Vertical Negative to Positive
3 Graphical: Linear Horizontal
4 Numerical: Linear Vertical with Numbers 1 to 5
5 Numerical: Linear Vertical with Numbers 5 to 1
6 Numerical: Linear Vertical with Numbers 2 to -2