

Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 509–528

## Design of Web Questionnaires: The Effect of Layout in Rating Scales

Vera Toepoel<sup>1</sup>, Marcel Das<sup>1</sup>, and Arthur van Soest<sup>2</sup>

This article shows that respondents gain meaning from verbal cues (words) as well as nonverbal cues (layout; numbers) in a web survey. We manipulated the layout of a five-point rating scale in two experiments. In the first experiment, we compared answers for different presentations of the responses: in one column with separate rows for each answer (“linear”), in three columns and two rows (“nonlinear”) in various orders, and after adding numerical labels to each response option. Our results show significant differences between a linear and non-linear layout of response options. In the second experiment we looked at effects of verbal, graphical, and numerical language. We compared two linear vertical layouts with reverse orderings (from positive to negative and from negative to positive), a horizontal layout, and layouts with various numerical labels (1 to 5, 5 to 1, and 2 to –2). We found effects of verbal and graphical language. The effect of numerical language was only apparent when the numbers 2 to –2 were added to the verbal labels. We also examined whether the effects of design vary with personal characteristics. Elderly respondents appeared to be more sensitive to verbal, graphical, and numerical language.

*Key words:* Web survey; questionnaire design; measurement error; context effects; scalar questions.

### 1. Introduction

Ordinal scale questions are probably the most widely used measurement instrument in web surveys. These questions are presented in various ways: answer categories can be presented in (one or more) columns, with labels for all categories or for the endpoint categories only, with radio buttons or an answer box, etc. It is well-known that differences in layout can lead to substantial differences in responses (Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Schwarz and Hippler 1987; Tourangeau, Couper, and Conrad 2004, 2007). Christian, Dillman, and Smyth (2005) suggest that writing effective questions for web surveys may depend at least as much on the presentation of the answer categories (“visual language”) as on the question wording itself.

Researchers have developed a theoretical framework that draws on linguistics and Gestalt psychology to explain how visual language influences the question-answering process (Jenkins and Dillman 1997), and a growing body of empirical research now

<sup>1</sup> Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. Emails: V.Toepoel@uvt.nl and das@uvt.nl

<sup>2</sup> Tilburg University and Netspar. Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. Email: avas@uvt.nl

**Acknowledgments:** The authors would like to thank Mick Couper, Don Dillman, and three anonymous reviewers for helpful comments.

provides a foundation for visual design theory. Four languages for communication are distinguished: verbal, graphical, numerical, and symbolic (Dillman 2007). Despite the growing empirical evidence, the theory of visual design is virtually without any reference to respondent characteristics (Stern, Dillman, and Smyth 2007). In line with this observation, studies have not examined how questionnaire format effects may vary with demographic characteristics of the respondents.

The purpose of this article is to determine how visual design elements in a rating scale influence survey answers. We manipulated verbal, graphical, and numerical languages individually. In addition, we explore how the effects of visual design vary with respondent characteristics.

## **2. Background**

In constructing ordinal scales for self-administered questionnaires, the visual layout of the scale is an important source of information for respondents when selecting an answer (Christian 2003; Christian, Parsons, and Dillman 2009). Tourangeau et al. (2004, 2007) argue that respondents use several visual heuristics to interpret a question. Each heuristic assigns a meaning to a visual cue. For example, respondents will see the middle option in a set of response options as the most typical. In addition, they will expect that the response options are presented in some kind of logical order. Another interpretive heuristic states that with a vertically oriented list, the top option will be seen as the most desirable. Also, visually similar options will be seen as closer conceptually. In addition to these visual heuristics, grouping principles from Gestalt Psychology can be used to understand visual design effects. For example, the Law of Pragnanz states that elements displaying simplicity, regularity, and symmetry are easier to perceive and to remember (Dillman 2007). Presenting the response scale with a layout that is inconsistent with these heuristics and principles results in different responses (see, for example, Christian and Dillman 2004; Smith 1995; Smyth et al. 2006; Tourangeau et al. 2004, 2007).

Verbal and nonverbal cues can independently and jointly influence the survey answers. For example, Redline et al. (2003) provide evidence 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. Dillman and Christian (2002) find that simultaneously manipulating several aspects of the visual language changes respondent behavior significantly.

### *2.1. Verbal Language*

We define verbal language as the verbal orientation of a scale; for example, by changing the verbal orientation of a scale (decremental versus incremental), responses may be altered because of different verbal labels at the first through last response options. Primacy effects lead to options at the beginning of a response list being selected more often, while recency effects lead to options near the end of a response list being chosen more often (Krosnick and Alwin 1987). Satisficing occurs when respondents are more likely to choose items earlier in a list because they settle for the first response option they consider satisfactory, rather than going through all of them (see Krosnick and Alwin 1987;

Krosnick, Narayan, and Smith 1996; and Tourangeau, Rips, and Rasinski 2000, for a detailed description of satisficing).

Research on orientation effects in rating scales is inconclusive. In some studies on visual presentation respondents altered their responses when the orientation of a scale changed, while in other studies responses remained unaffected (Weng and Cheng 2000).

## 2.2. Graphical Language

Dillman (2007) defines graphical language as: elemental visual features such as size, shape, location, spatial arrangement, color, brightness, contrast, and figure/ground composition that convey meaning in questionnaires. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales do not elicit the same responses. However, the direction of the difference varied across items. Their results are thus inconclusive and future research is warranted.

A non-linear layout (where options are presented in multiple rows and columns) can also result in different responses compared to a linear<sup>3</sup> layout, because the graphical language conveying the scale is interrupted (Christian 2003; Christian and Dillman 2004).

## 2.3. Numerical Language

This refers to the numbers associated with response options. Schwarz et al. (1985) showed that respondents seek information about the researcher's expectations using the numerical labels on a scale as frames of reference. Schwarz et al. (1991) found that changing the numerical values attached to scales resulted in different answers. In particular, respondents hesitated to assign a negative score to themselves (see also Tourangeau et al. 2000, p. 248, and Tourangeau et al. 2007).

Most of the research on layout in scalar questions is based on paper surveys. One could argue that web questionnaires may prompt stronger response effects than paper questionnaires. According to De Leeuw (2005), the web is the most dynamic of the available modes of administration, and the possibility of multitasking may lead to more superficial cognitive processing and more satisficing in answering survey questions. Christian and Dillman (2004) indeed found larger visual effects on the web than in a mail survey.

Visual design theory is virtually without any reference to respondent characteristics (Dillman 2007). As a result, the empirical tests have not analyzed how the effects of questionnaire format vary with respondent characteristics. Couper (2000) argues that design may interact with the type of web survey conducted and the survey's target population. Of the few available studies on personal characteristics, some suggest that effects do occur (mainly caused by working memory capacity), while others find no variation in response effects due to personal characteristics. Tourangeau et al. (2007) observe no consistent variation in the impact of a response scale layout due to gender, age, or education group. Stern et al. (2007) also show that the layout of survey questions affects

<sup>3</sup> In line with the literature, in this article a linear presentation refers to a presentation of response options in a single row or a single column.

different demographic groups in similar ways. In addition, McFarland (1981) finds no evidence that gender and educational level interact with the ordering of questions. Krosnick and Alwin (1987), on the other hand, find that respondents with less education and more limited vocabularies are influenced more than others by different answer categories. Knäuper, Schwarz, and Park (2004) and Borgers, Hox, and Sikkel (2004) also find differences due to working memory capacity. Fuchs (2005) finds that the effects of which response categories are used, of the order in which responses are presented, and of labeling response categories with numerical values decrease with age when children and adolescents are compared, supporting the hypothesis that response effects decrease with increase in the level of cognitive sophistication. The groups that are compared are very distinct groups, however, and results cannot be compared with an entire (heterogeneous) population.

The existing literature suggests that additional research on the visual design of web questionnaires is needed to develop more general principles for how the visual layout influences answers (Christian and Dillman 2004; Dillman and Christian 2002; Dillman, Gertseva, and Mahon-Haft 2005; Friedman and Leefer 1981; Jenkins and Dillman 1997; Schwarz et al. 1991). Several authors (see Deutskens et al. 2004; Dillman, Caldwell, and Gansemer 2000; Friedman and Leefer 1981; Hofmans et al. 2007; Stern et al. 2007) conclude that future research on visual design should study the effects of presentation on different questionnaires and populations. Such work is essential of effective survey construction and offers the possibility for methodological improvements of survey research. The experiments described in the following section examine whether visual languages influence respondents' answers in a Dutch online panel and if any effects are tied to personal characteristics.

### **3. Design and Implementation**

Studies on scalar questions have focused on the number of scale points, the use of verbal labels, the use of a midpoint, the use of numerical labels, the use of a "don't know" filter, and the graphical layout of scales. See Christian (2003); Krosnick and Fabrigar (1997), and Schwarz (1996) for a discussion of these factors in relation to response scales of ordinal questions. We use five point scalar questions, as this provides many possibilities of manipulating visual cues.

We carried out two experiments using eight different formats in the CentERpanel, a web-based household panel consisting of more than 2,000 households. This panel is administered by CentERdata (Tilburg University, the Netherlands). The panel is designed to be representative of the Dutch-speaking population aged 16 and older in the Netherlands. Households that do not have access to the Internet when recruited are provided with a so-called Net.Box, which can be used to establish a connection via a telephone line and a television set. If the household does not have a television, CentERdata provides that as well. New panel members are recruited in three stages. In the first stage, a random sample (landline numbers) of persons is interviewed by telephone. The interview ends with the question whether the person would like to participate in survey research projects. If so, the person and his or her household are included in a database of potential panel members. If a household drops out of the panel, a new household is selected from the

database of potential panel members. This is done on the basis of demographic characteristics, such that the panel will remain representative (for more details, see <http://www.centerdata.nl/en/CentERpanel>).

It is difficult to give response rates for the three different recruitment stages since this is an ongoing process. Hoogendoorn and Daalmans (2008) looked at the recruitment process for the CentERpanel from 2001 to 2003 and reported a response rate of 60% for the first interview, 38% for the willingness to become a panel member and 50% for actual participation, resulting in an overall response rate of 12%. Our study was conducted in Week 37 (September) and Week 41 (October) of 2005. The percentage of panel members responding to our specific survey was 78.3% (2,787 panel members were selected, 2,182 responded) for the first experiment and 78.8% (2,830 panel members were selected, 2,229 responded) for the second experiment. There was no partial nonresponse, probably due to the fact that the questionnaire was quite short. In some households there is more than one respondent. In Experiment 1, the 2,182 respondents belonged to 1,535 households. In Experiment 2, the 2,229 respondents represented 1,537 households. The overlap in respondents between the two experiments was about 70%.

The amount of time that the respondents in our experiment were already participating in the CentERpanel varied from a few months to seventeen years. We used this information to test for interactions between survey experience and the effects of visual language in the questions under consideration, but we did not find any significant interactions.

The experiments used two questions: one on the quality of education and one on the quality of life in the Netherlands. These questions were taken from an experiment conducted by Christian (2003), who measured the quality of education and the quality of student life at Washington State University. A first group of respondents in each experiment answered a rating scale with the answer categories of excellent, very good, good, fair, and poor in a linear vertical format from positive to negative. In the first experiment we varied the graphical presentation using three non-linear manipulations. In the second experiment we manipulated graphical, numerical, and verbal languages in a linear format (see Appendix A for screenshots).

The first experiment is a replication of an experiment performed by Christian (2003) on graphical language interrupting a scale due to banking (presenting response options in columns). We compared a linear vertical format (Appendix A: 1a) to two nonlinear formats: a triple banked format with options presented horizontally in 2 rows and 3 columns (Appendix A: 1b) and a triple banked format with options running vertically in 3 columns and 2 rows (Appendix A: 1c). To test whether numbers would help respondents read the triple vertical format, a fourth group answered the questions in a triple vertical format with numbers (Appendix A: 1d). The banking of response options tests how respondents react when the graphical language denoting the scale is interrupted. Banking response options is more common in pen and paper surveys than in web surveys (to save space), but particularly in mixed mode surveys, where web and paper questionnaires are presented in the same way to give the same (visual) stimulus, banked options are implemented in web questionnaires as well. This makes it important to develop a better understanding of banked response options in web surveys.

In the second experiment, the first group again answered on a rating scale in a linear vertical format from positive to negative (Appendix A: 2a). All other groups have different

linear manipulations in relation to this format. The second group answered on the same scale, but from negative to positive (poor to excellent, Appendix A: 2b). For the third group the graphics were changed: a linear horizontal format was used (Appendix A: 2c). In the fourth group we added numbers 1 to 5 (Appendix A: 2d). For the fifth group the numbers 5 to 1 were added in the education question (first question), while in the life question (second question) the numbers varied from plus 2 to minus 2 (Appendix A: 2e). Since the objective was to determine which respondents are more sensitive to nonverbal cues, we compared scores for different gender, age, and education groups.

#### **4. Results**

In this section we discuss the results of the two experiments. We first consider the effects for the complete sample and then consider subsamples with specific demographic characteristics. In analyzing personal characteristics we also looked at the influence of personality factors such as the Need for Cognition (NFC, Cacioppo and Petty 1982) and Need to Evaluate (NES, Jarvis and Petty 1996), but since their effect on responses was small and insignificant we decided not to include these analyses in the article. We also used information on survey experience (e.g., the number of weeks on the panel) to test for an interaction between survey experience and the effects of visual language in the two questions that we fielded, but we again found no significant interactions.

##### *4.1. Experiment 1: Graphical Manipulations of Layout*

Following Christian (2003); Christian and Dillman (2004), and Tourangeau et al. (2004), we hypothesized that a nonlinear layout would result in different responses compared to the linear layout because the graphical language denoting the scale is interrupted. Christian (2003) shows that some respondents read the top line only and select particularly the response option right next to the first one. We therefore expected that, in the nonlinear format, respondents would more often choose options in the first line. We also expected that response times would be different across formats because of visual heuristics and Gestalt Psychology, with the reference level (linear format) showing the shortest completion time because this layout is easier to perceive and remember.

Table 1 displays response distributions for the two questions. In addition, Chi-square and *t*-tests statistics are presented to test for differences in the distribution of individual responses across formats and in mean responses. These tests are the same as those in previous research (Christian 2003; Christian and Dillman 2004; Dillman and Christian 2002; Stern et al. 2007). Lower mean scores indicate more positive ratings (1 = “excellent”, . . . , 5 = “poor”).

The results in Table 1 show that graphical language influences the answers to both questions. The overall Chi-square test and differences of means test reject the null hypothesis of no differences between the four versions ( $\chi^2 = 33.86, p < .001; F = 6.71, p < .01$  in the education question and  $\chi^2 = 43.96, p < .001; F = 8.96, p < .01$  in the life question). Separate tests show that the linear version produces significantly different responses and mean scores compared to each of the triple versions, as hypothesized. We found no evidence that respondents are more likely to select an option from the first line in nonlinear formats, however. Some frequencies seem to confirm the conjecture that

Table 1. Experiment 1. Frequencies (in %), mean scores, correlations and mean differences in linear and nonlinear formats

	Nonlinear – Triple			
	1a. Linear	1b. Horizontal	1c. Vertical	1d. Vertical with numbers
<i>Overall, how would you rate the quality of education in the Netherlands?</i>				
1 Excellent	1.5	0.9	0.6	1.5
2 Very good	17.8	12.9	10.8	14.7
3 Good	51.3	44.0	52.1	48.9
4 Fair	25.1	36.2	31.9	28.3
5 Poor	4.4	6.0	4.6	6.6
N	550	552	545	530
Mean	3.13	3.34	3.29	3.24
<i>Overall, how would you rate the quality of life in the Netherlands?</i>				
1 Excellent	2.9	2.0	1.5	4.4
2 Very good	32.3	21.4	24.1	26.4
3 Good	49.9	51.6	56.3	47.3
4 Fair	13.9	23.4	17.0	20.7
5 Poor	0.9	1.7	1.1	1.2
N	545	543	536	518
Mean	2.78	3.01	2.92	2.88
			Chi Square tests	Diff. of means
			$\chi^2$	T
<i>Overall, how would you rate the quality of education in the Netherlands?</i>				
1a versus 1b			20.69**	– 4.20**
1a versus 1c			16.12**	– 3.44**
1a versus 1d			5.43	– 2.14*
1c versus 1b			7.66	– 0.93
1c versus 1d			9.30*	1.12
Overall across all 4 formats			33.86**	F = 6.71**
<i>Overall, how would you rate the quality of life in the Netherlands?</i>				
1a versus 1b			27.32**	– 5.12**
1a versus 1c			12.84**	– 3.26**
1a versus 1d			12.19*	– 2.07*
1c versus 1b			8.49	– 2.02*
1c versus 1d			14.43*	0.95
Overall across all 4 formats			43.96**	F = 8.96**

\* =  $p < .05$ , \*\* =  $p < .01$ .

Note: A high mean score indicates a negative judgment.

respondents tend to select the answer right next to the first option on the first line. For example, the response option “good” was chosen significantly more often in the triple vertical format than in the triple horizontal format (52.1% versus 44.0% in the education question and 56.3% versus 51.6% in the life question). However, the option “fair” was chosen more often when presented in the first column (36.2% versus 31.9% in the education question and 23.4% versus 17.0% in the life question), indicating that some respondents read in columns rather than rows. The sizes of these effects are similar to the findings of Christian (2003). The effect of visual language decreased when numbers were



added to the vertical format. There may be a hierarchy of features that respondents pay attention to, with numerical labels dominating purely visual cues, as suggested by Tourangeau et al. (2007).

We found no difference in response times between formats for either question, and therefore no evidence to support the conjecture that it takes longer to process a question if the graphical language does not accord with visual heuristics and principles.

#### *4.2. Experiment 2: Verbal, Graphical, and Numerical Manipulations of Layout*

Tables 2 and 3 show the results for our second experiment. One significant contrast between the two questions is that a joint Chi-square test and differences of means test for all nonverbal manipulations (excluding verbal manipulation 2b) did not show differences in the education question ( $\chi^2 = 15.97, p = .19; F = 1.98, p = .12$ ) while they did in the life question ( $\chi^2 = 115.16, p < .001; F = 32.01, p < .001$ ). We suspect that this difference is due to the use of different numbers in format 2e. In the education question the numbers 5 to 1 were added to this format, while in the life questions the numbers 2 to -2 were used. We also looked at the duration of response times to find out if some formats take longer to process, but we found no significant differences between formats.

##### *4.2.1. Verbal Language*

Dillman (2007) distinguishes four languages for communicating visually; one of them is verbal language. By changing the verbal orientation of a scale, visual heuristics like “left and top means first” and “up means good” (Tourangeau et al. 2004, 2007) are violated. In addition, the theory of satisficing (Krosnick and Alwin 1987; Krosnick, Narayan, and Smith 1996; and Tourangeau et al. 2000) states that respondents are more likely to choose items earlier in a list because they find the first position that they can reasonably agree with to be a satisfactory answer, instead of processing each response option separately. Therefore, we hypothesized that respondents would select more positive responses in the reference format (positive to negative) compared to the reversed verbal manipulation (negative to positive).

Our two questions show statistically different answer distributions and mean scores between a decremental and an incremental scale (2a versus 2b), indicating that respondents are affected by verbal language ( $\chi^2 = 52.23, p < .001$  for the education question, and  $\chi^2 = 39.92, p < .001$  for the life question). The mean score in the positive to negative scale is lower than the mean of the negative to positive scale in both questions (mean = 2.91 for the decremental scale and 3.28 for the incremental scale in the education question; 2.60 and 2.88, respectively, in the life question), providing evidence for a primacy effect. Our results thus provide empirical support, in a different country and culture than the literature, for the theory of satisficing and primacy effects. Moreover, in the Netherlands it is much more usual than in English-speaking countries to use an incremental scale than a decremental scale, e.g., when using Likert scales such as “fully disagree,” . . . “fully agree” (Hofmans et al. 2007). Therefore, our results suggest that the effect of satisficing leading to a primacy effect is larger than the effect of violating visual heuristics.



Table 2. Experiment 2: education question. Frequencies (in %), mean scores, correlations and mean differences in the verbal, graphical, and numerical manipulations

	2a. Reference: Linear vertical positive to negative	2b. Verbal: Linear vertical negative to positive	2c. Graphical: Linear horizontal	2d. Numerical: Linear vertical with numbers 1 to 5	2e. Numerical: Linear vertical with numbers 5 to 1
<i>Overall, how would you rate the quality of education in the Netherlands?</i>					
1 Excellent	2.7	1.5	0.5	3.1	2.5
2 Very good	24.0	10.7	23.4	22.8	25.4
3 Good	54.8	51.3	52.8	53.8	55.1
4 Fair	16.5	31.1	21.9	17.9	15.2
5 Poor	2.0	5.4	1.4	2.4	1.8
N	442	460	415	457	448
Mean	2.91	3.28	3.00	2.94	2.88
				Chi Square tests $\chi^2$	Diff. of means T
<i>Overall, how would you rate the quality of education in the Netherlands?</i>					
Verbal: 2a versus 2b				52.23**	- 7.17**
Graphical: 2a versus 2c				10.43*	- 1.82
Numerical: 2a versus 2d				.68	- .52
Numerical: 2a versus 2e				.58	.55
Overall across all nonverbal manipulations (excluding 2b)				15.97	F = 1.98
Overall across all 5 formats				95.21**	F = 20.42**

\* =  $p < .05$ , \*\* =  $p < .01$ .

Note: A high mean score indicates a negative judgment.

Table 3. Experiment 2: life question. Frequencies (in %), mean scores, correlations and mean differences in the verbal, graphical, and numerical manipulations

	2a. Reference: Linear vertical positive to negative	2b. Verbal: Linear vertical negative to positive	2c. Graphical: Linear horizontal	2d. Numerical: Linear vertical with numbers 1 to 5	2e. Numerical: Linear vertical with numbers 2 to -2
<i>Overall, how would you rate the quality of life in the Netherlands?</i>					
1 Excellent	5.7	3.7	2.7	4.2	8.1
2 Very good	35.7	25.6	37.4	40.4	40.1
3 Good	52.3	51.1	49.0	43.3	41.3
4 Fair	5.7	18.5	10.1	11.3	9.4
5 Poor	0.7	1.1	0.7	0.9	0.9
N	440	454	414	453	446
Mean	2.60	2.88	2.69	2.64	2.54
				Chi Square tests $\chi^2$	Diff. of means T
<i>Overall, how would you rate the quality of life in the Netherlands?</i>					
Verbal: 2a versus 2b				39.92**	-5.50**
Graphical: 2a versus 2c				71.92**	-8.20**
Numerical: 2a versus 2d				14.37**	-.85
Numerical: 2a versus 2e				13.29**	1.07
Overall across all nonverbal manipulations (excluding 2b)				115.16**	F = 32.01**
Overall across all 5 formats				136.05**	F = 28.84**

\* =  $p < .05$ , \*\* =  $p < .01$ .

Note: A high mean score indicates a negative judgment.

#### 4.2.2. Graphical Language

By changing the graphical orientation of the scale from vertical to horizontal, the graphical language is altered. Friedman and Friedman (1994) demonstrate that equivalent horizontal and vertical rating scales do not elicit the same responses. However, the direction of the difference they found was not consistent. We hypothesized that responses would be shifted to the left in a horizontal format (more positive ratings), due to the necessity of more hand/eye movement to select the last options in a horizontal format.

Chi-square tests indicate significant differences in the responses across the vertical and horizontal versions ( $\chi^2 = 10.43$ ,  $p = .04$  in the education question, and  $\chi^2 = 71.92$ ,  $p < .001$  in the life question); the mean score differs statistically in the life question ( $t = -8.20$ ,  $p < .001$ ), but only marginally so in the education question ( $t = -1.82$ ,  $p = .07$ ). Differences resulted from selecting the fourth option “fair” more often in the horizontal format. Thus, in the horizontal format a shift to the left is not detected.

#### 4.2.3. Numerical Language

Based on the literature (Fuchs 2005; Schwarz et al. 1985; Schwarz et al. 1991), we hypothesized that respondents would choose response options with low numbers less often when numbers are added to the verbal labels compared to the reference level without numbers. In addition, we hypothesized that response options with negative numbers would be chosen less often.

Little evidence was found that adding the numbers 1 to 5 produced different response distributions. Chi-square tests indicated significant differences in the responses across the linear version (2a) and the linear version with numbers 1 to 5 (2d) for the life question only ( $\chi^2 = .68$ ,  $p = .95$  in the education question, and  $\chi^2 = 14.37$ ,  $p < .01$  in the life question). No differences in mean scores were found ( $t = -0.52$ ,  $p = .60$  in the education question, and  $t = -0.85$ ,  $p = .39$  in the life question). When adding the numbers in the reverse order (format 2e) in the education question, we did not find significant differences compared to the reference format (2a) either.

The Chi-square test indicated significant differences in the response distribution when numbers 2 to  $-2$  were added compared to the reference format in the life question ( $\chi^2 = 13.29$ ,  $p = .01$ ; format 2a versus 2e). This confirms that negative numbers are interpreted as implying more extreme judgments than no numbers on verbal labels (scale label effect, see Tourangeau et al. 2000, p. 248; Schwarz et al. 1991; Tourangeau et al. 2007).

Since we experimented with all visual languages in the same experiment with the same questions, we were also able to test the manipulations against each other instead of the reference level only. When we tested the verbal manipulation (reversal of the scale; format 2b) against the graphical manipulation (horizontal layout; format 2c), we found smaller differences compared to the verbal manipulation against the reference level (format 2a versus 2b). Chi-squares are smaller ( $\chi^2 = 40.19$ ,  $p < .001$  in the education question and  $\chi^2 = 12.56$ ,  $p = 0.01$  in the life question). In addition, the differences in mean scores get smaller if we compare the verbal manipulation (2b) with the graphical manipulation (2c) instead of the reference level (2a). The differences in mean scores are .37 in the education question and .28 in the life question when we compare the verbal manipulation with the

reference level (2a versus 2b) and .28 and .19, respectively, when we compare the verbal manipulation with the graphical manipulation (2b versus 2c, horizontal layout). This effect is also confirmed by the smaller value of the test statistic ( $t = -5.43$ ,  $p < .001$  in the education question and  $t = -2.59$ ,  $p = .01$  in the life question). This suggests that the effect of the verbal orientation of the scale is smaller when using a horizontal layout. Toepoel and Dillman (2008) also suggest that the effect of visual language is smaller when using a horizontal layout than when using a vertical layout. We did not have a fourth experimental condition (reversal of the scale in a horizontal format) to thoroughly test the relation between verbal and graphical orientation of a scale. Future research should be conducted to elucidate the relation between response effects and a horizontal layout.

#### 4.3. *Effects for Different Demographic Subgroups*

Based on previous research on personal characteristics (Krosnick, Narayan, and Smith 1996; Stern et al. 2007), we defined the following demographic subgroups: men and women; two educational categories (with and without a college degree), and two age categories (65 years and older, and under the age of 65). We performed ordered logistic regressions on the two questions in both experiments. Results are presented in Table 4. Following Borgers et al. (2004); Fuchs (2005); Knäuper et al. (2004), and Krosnick and Alwin (1987), we hypothesized that the effects of design are larger for older respondents and respondents without a college degree. We hypothesized no effect of gender.

In Experiment 1 we found no significant interaction effects between format and demographic subgroups. The effects of linear and nonlinear layouts are the same for men and women, respondents with and without a degree, and respondents over and under the age of 65. In Experiment 2, we found again no significant differences with regard to layout effects between men and women. In addition, no effect of education was found. We did find significant differences between respondents over and under the age of 65 with regard to changes in a linear layout.

Respondents aged 65 and older were more likely to select lower ratings when the scale was reversed (difference between 2a and 2b in both questions; verbal language). They showed a larger primacy effect when the visual heuristic “up means good” was violated compared to younger respondents. Both questions in Experiment 2 showed significant differences (see Table 4,  $B = .861$ ,  $p = .01$  in the education question and  $B = 1.397$ ,  $p < .0001$  in the life question). Note that a high answer indicates a negative judgment (1 = excellent, . . . , 5 = poor).

Older respondents were also more influenced by a change in the graphical orientation of the scale compared to their younger counterparts. When the graphical orientation was changed from vertical to horizontal (2a versus 2c), older respondents showed a larger recency effect than younger respondents. Ordinal regression results in Table 4 show a significant effect in the life question ( $B = 1.103$ ,  $p = .001$ ). The education question shows a significant interaction effect at the 10% level ( $B = .580$ ,  $p = .08$ ).

The adding of numbers 1 to 5 to the verbal labels (2d) did not lead to significant differences between older and younger respondents. However, the adding of numbers 5 to 1 (format 2e in the education question) did. Older respondents chose the first options (with numbers 5 and 4) less often. This effect can be seen in Table 4. Where the main format

Table 4. Ordinal logistic regression estimates (B) explaining the respondent's rating by dummies for format, gender (1 = female), age (1 = 65 and older), education (1 = with college degree), and interactions between explanatory variables

	Experiment 1		Experiment 2	
	Education question	Life question	Education question	Life question
format b	.519**	.554**	.982**	.691**
format c	.390*	.333	.226	.609**
format d	.430*	.289	.040	-.011
format e			-.160	-.336
gender	.308	.244	.215	.011
age	.585**	.081	.078	-.449
education	.280	-.468**	.074	-.472*
gender*format b	-.018	-.124	-.232	-.299
gender*format c	-.211	-.271	-.037	.163
gender*format d	-.230	.077	-.089	-.091
gender*format e			-.022	.189
age*format b	.215	.254	.861**	1.397**
age*format c	.218	.076	.580	1.103**
age*format d	.430	-.031	.281	.598
age*format e			.701*	.713*
edu*format b	.006	.174	-.232	-.232
edu*format c	.337	.424	-.230	.318
edu*format d	-.316	-.155	.124	.067
edu*format e			-.119	-.122

\* $p < .05$ , \*\* $p < .01$ .

Note: Experiment 1: format b = nonlinear-horizontal, format c = nonlinear-vertical, format d = nonlinear-vertical with numbers.

Experiment 2: format b = linear vertical negative to positive, format c = linear horizontal, format d = linear vertical with numbers 1 to 5, format e = linear vertical with numbers 5 to 1 (education question) or numbers 2 to -2 (life question).

effect (2e, adding numbers 5 to 1) in the education question was negative and insignificant ( $B = -.160, p = .47$ ), the interaction effect of format (2e) and a dummy for age showed a positive effect ( $B = .701, p = .03$ ; compared to younger respondents, older respondents more easily selected one of the last options when numbers 5 to 1 were added). No significant differences between young and old respondents were found when the numbers 2 to  $-2$  (2e in the life question) were added to the verbal labels.

In all, we found evidence that reduction in cognitive functioning as part of the aging process may lead to larger response effects due to verbal, graphical, and numerical language. Respondents aged 65 and older showed larger primacy effects in the verbal manipulation and larger recency effects in the graphical and numerical manipulation, compared to younger respondents.

## 5. Discussion and Conclusions

This article shows that respondents draw meaning from nonverbal as well as verbal cues in a web survey. We manipulated the layout of a five-point scalar question in two experiments using two questions. In the first experiment, a linear layout was compared with three nonlinear layouts (graphical manipulation). In the second experiment we manipulated verbal, graphical, and numerical languages individually, to learn how these verbal and nonverbal cues influence answers in rating scales. This article advances previous research by manipulating linear and nonlinear formats as well as verbal, graphical, and numerical languages on the same rating scale. Moreover, it examines how the effects of layout manipulations vary with personal characteristics.

Comparing linear and nonlinear formats, we found differences across all versions. Triple horizontal and triple vertical formats show significantly different means compared to the linear format. The effect of visual language decreases if numbers are added to the vertical format. This seems to point toward a hierarchy of features that respondents pay attention to, with numerical labels taking precedence over purely visual cues, as suggested by Tourangeau et al. (2007). Future research can elucidate this effect. Our results may be specific to the particular distributions of the variables being studied. Similar experiments with other survey questions would be useful to test the robustness of our findings. We leave this for future research.

In Experiment 2, differences in responses caused by variation in visual language were also found. The verbal manipulation (“excellent”-“poor” versus “poor”-“excellent”) shows significantly different responses compared to the other manipulations. This indicates satisficing and also that a negative tone of the first option changes reports in a negative manner (an anchoring effect, as suggested by Schwarz 1996). Respondents select the second option more often. Presenting the answer categories in a horizontal format resulted in different response distributions than presenting them in a vertical format. One interpretation is that respondents may be more willing to read all options in the horizontal format (because they first read horizontally and then vertically), but the lack of a difference in the time respondents spent on answering the questions does not support this conjecture. Adding numerical labels had little effect on the answers respondents provided. When we compare the reversal of the scale (verbal manipulation, format 2b) to a vertical layout (reference level, format 2a) and a horizontal layout (graphical manipulation,

format 2c), our results suggest that a horizontal format is less susceptible to layout effects than a vertical layout. This is also suggested by Toepoel and Dillman (2008). We do not have a fourth condition with a reversal of the scale in a horizontal layout, however. Further research in web surveys on a horizontal layout of scalar questions in different contexts (e.g., question types, scale points) is warranted.

Verbal, graphical, and numerical language appears to have a greater impact on response behavior for older respondents. We attribute this to the reduction in cognitive functioning as part of the aging process.

This article shows that, to reduce measurement error, the visual presentation of answer categories must be taken into account. This applies especially to researchers who want to compare results across surveys. Similarly-worded questions may be presented to respondents in visually dissimilar ways. Should different results then be attributed to a different time of measurement or to a different visualization? This is a challenge for further research. We recommend using a linear horizontal layout without numbers for a five-point fully labeled rating scale.

## Appendix A: Screenshots

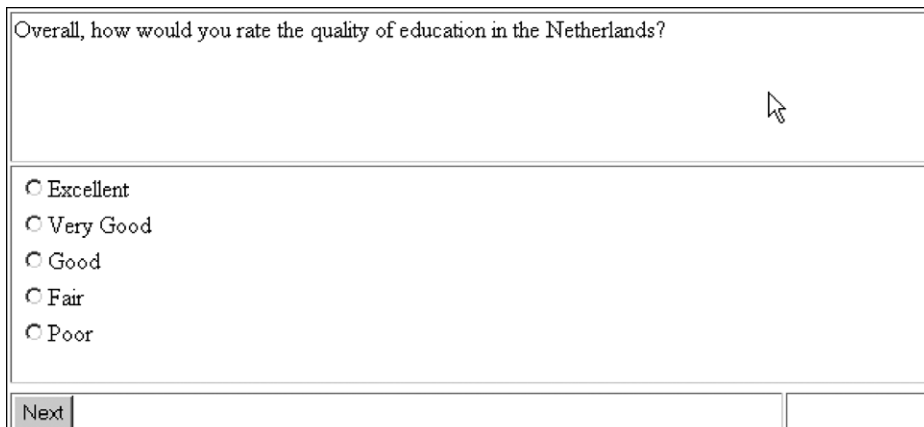
### Experiment 1

Using a linear and a nonlinear format, four different layouts were used to present two questions, namely:

1. Overall, how would you rate the quality of education in the Netherlands?
2. How would you rate the quality of life in the Netherlands?

The screenshots below show the different layout formats for the education question. The layout formats used in the life question are exactly the same.

- 1a. Linear



The screenshot shows a web form with a question: "Overall, how would you rate the quality of education in the Netherlands?". Below the question is a list of five radio button options: "Excellent", "Very Good", "Good", "Fair", and "Poor". At the bottom left of the form is a "Next" button.

- 1b. Nonlinear – triple horizontal



Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> Excellent	<input type="radio"/> Very Good
<input type="radio"/> Fair	<input type="radio"/> Good
<input type="radio"/> Poor	
Next	

## 1c. Nonlinear – triple vertical

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> Excellent	<input type="radio"/> Good
<input type="radio"/> Very Good	<input type="radio"/> Fair
	<input type="radio"/> Poor
Next	

## 1d. Nonlinear – triple vertical with numbers

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> 1 Excellent	<input type="radio"/> 3 Good
<input type="radio"/> 2 Very Good	<input type="radio"/> 4 Fair
	<input type="radio"/> 5 Poor
Next	

*Experiment 2*

Five different layouts were used to present the same two questions (as in experiment 1):

Format a: reference format (see 1a);

Format b: verbal manipulation: in this format, response scale is from negative to positive;

Format c: graphical manipulation: in this format, response scale is from vertical to horizontal;

Format d: numerical manipulation: numbers 1 to 5 are added in this format;

Format e: numerical manipulation: numbers 5 to 1 are added in the education question, while numbers 2 to –2 are added in the (life question).

The screenshots below show the different layout formats for the education question, the layout formats used in the life question are the same except for format e (see above).

2a. Linear positive to negative

See screenshot 1a.

2b. Linear negative to positive (verbal)

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> Poor <input type="radio"/> Fair <input type="radio"/> Good <input type="radio"/> Very Good <input type="radio"/> Excellent	
Next	

2c. Linear horizontal (graphical)

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> Excellent <input type="radio"/> Very Good <input type="radio"/> Good <input type="radio"/> Fair <input type="radio"/> Poor	
Next	

2d. Linear with numbers 1 to 5, 1 = positive (numerical)

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> 1 Excellent <input type="radio"/> 2 Very Good <input type="radio"/> 3 Good <input type="radio"/> 4 Fair <input type="radio"/> 5 Poor	
Next	

2e. Linear with numbers 1 to 5, 5 = positive in education question (numerical)

Overall, how would you rate the quality of education in the Netherlands?	
<input type="radio"/> 5 Excellent <input type="radio"/> 4 Very Good <input type="radio"/> 3 Good <input type="radio"/> 2 Fair <input type="radio"/> 1 Poor	
Next	

Note: Format 2e for the life question ranges from 2 (positive) to  $-2$  (negative).

## 6. References

- Borgers, N., Hox, J., and Sikkel, D. (2004). Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point. *Quality & Quantity*, 38, 17–33.
- Cacioppo, J.T. and Petty, R.E. (1982). The Need for Cognition. *Journal of Personality and Social Psychology*, 42, 116–131.
- Couper, M.P. (2000). Web Surveys. A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, 464–494.
- Christian, L.M. (2003). The Influence of Visual Layout on Scalar Questions in Web Surveys. Unpublished Master's Thesis. Retrieved on <http://survey.sesrc.wsu.edu/dillman/papers.htm>
- Christian, L.M. and Dillman, D.A. (2004). The Influence of Graphical and Symbolic Language Manipulations to Self-Administered Questions. *Public Opinion Quarterly*, 68, 57–80.
- Christian, L.M., Dillman, D.A., and Smyth, J.D. (2005). Instructing Web and Telephone Respondents to Report Date Answers in a Format Desired by the Surveyor. Technical Report #05-067. Social & Economic Sciences Research Center Pullman, Washington State University. Retrieved on <http://survey.sesrc.wsu.edu/dillman/papers.htm>
- Christian, L.M., Parsons, N.L., and Dillman, D.A. (2009). Designing Scalar Questions for Web Surveys. *Sociological Methods and Research*, 37, 393–425.
- De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, 233–255.
- Deutskens, E., De Ruyter, K., Wetzels, M., and Oosterveld, P. (2004). Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study. *Marketing Letters*, 15, 21–36.

- Dillman, D.A. (2007). *Mail and Internet Surveys. The Tailored Design Method*. Hoboken, NJ: Wiley.
- Dillman, D.A., Caldwell, S., and Gansemer, M. (2000). Visual Design Effects on Item Nonresponse to a Question About Work Satisfaction That Precedes the Q-12 Agree-Disagree Items. Paper supported by the Gallup Organization and Washington State University. Retrieved on <http://survey.sesrc.wsu.edu/dillman/papers.htm>
- Dillman, D.A. and Christian, L.M. (2002). The Influence of Words, Symbols, Numbers, and Graphics on Answers to Self-Administered Questionnaires: Results from 18 Experimental Comparisons. Retrieved on <http://survey.sesrc.wsu.edu/dillman/papers.htm>
- Dillman, D.A., Gertseva, A., and Mahon-Haft, T. (2005). Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles. *Journal of Official Statistics*, 21, 183–214.
- Friedman, L.W. and Friedman, H.H. (1994). A Comparison of Vertical and Horizontal Rating Scales. *The Mid-Atlantic Journal of Business*, 30, 107–202.
- Friedman, H.H. and Leefer, J.R. (1981). Label Versus Position in Rating Scales. *Journal of the Academy of Marketing Science*, 9, 88–92.
- Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N., and Cools, W. (2007). Bias and Changes in Perceived Intensity of Verbal Qualifiers Effected by Scale Orientation. *Survey Research Methods*, 1, 97–108.
- Hoogendoorn, A.W. and Daalmans, J. (2008). Nonresponse in the Recruitment of an Internet Panel Based on a Probability Sample. *Statistics Netherlands Discussion Paper*. 08007.
- Fuchs, M. (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.
- Jarvis, W.B.G. and Petty, R.E. (1996). The Need to Evaluate. *Journal of Personality and Social Psychology*, 70, 172–194.
- Jenkins, C.R. and Dillman, D.A. (1997). Towards a Theory of Self-Administered Questionnaire Design. In *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley Series in Probability and Statistics.
- Knäuper, B., Schwarz, N., and Park, D. (2004). Frequency Reports Across Age Groups. *Journal of Official Statistics*, 20, 91–96.
- Krosnick, J.A. and Alwin, D.F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51, 201–219.
- Krosnick, J.A. and Fabrigar, L.R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley Series in Probability and Statistics.
- Krosnick, J.A., Narayan, S., and Smith, W.R. (1996). Satisficing in Surveys: Initial Evidence. *New Directions for Program Evaluation*, 70, 29–44.
- McFarland, S.G. (1981). Effects of Question Order on Survey Responses. *Public Opinion Quarterly*, 45, 208–215.

- Redline, C.D., Dillman, D.A., Carley-Baxter, L., and Creecy, R. (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. Retrieved on <http://survey.sesrc.wsu.edu/dillman/papers.htm>
- Schwarz, N. (1996). *Cognition and Communication. Judgmental Biases, Research Methods, and the Logic of Conversation*. Hillsdale, New Jersey: Erlbaum.
- Schwarz, N. and Hippler, H.-J. (1987). What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives. In *Social Information Processing and Survey Methodology*, H.-J. Hippler, N. Schwarz, and S. Sudman (eds). New York: Springer-Verlag.
- Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49, 388–395.
- Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, 55, 570–582.
- Smith, T.W. (1995). Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1046–1051.
- Smyth, J., Dillman, D.A., Christian, L.M., and Stern, M.J. (2006). Effects of Using Visual Design Principles to Group Response Options in Web Surveys. *International Journal of Internet Science*, 1, 6–16.
- Stern, M.J., Dillman, D.A., and Smyth, J.D. (2007). Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey. *Survey Research Methods*, 1, 121–138.
- Toepoel, V., and Dillman, D.A. (2008). Words, Numbers and Visual Heuristics in Web Surveys: Is there a Hierarchy of Importance? *CentER Discussion Paper 2008-92*. CentER: Tilburg University.
- Tourangeau, R., Couper, M.P., and Conrad, F. (2004). Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68, 368–393.
- Tourangeau, R., Couper, M.P., and Conrad, F. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, 71, 91–112.
- Tourangeau, R., Rips, L.J., and Rasinski, R. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Weng, L. and Cheng, C.P. (2000). Effects of Response Order on Likert-type Scales. *Educational and Psychological Measurement*, 60, 908–924.

Received April 2007

Revised April 2009