CentER

Discussion Paper

No. 2006–30

DESIGN OF WEB QUESTIONNAIRES: THE EFFECT OF LAYOUT IN RATING SCALES

By Vera Toepoel, Marcel Das, Arthur van Soest

April 2006

ISSN 0924-7815

Tilburg University

DESIGN OF WEB QUESTIONNAIRES:

THE EFFECT OF LAYOUT IN RATING SCALES

Vera Toepoel,[*] Marcel Das[*] and Arthur van Soest[**1]

Abstract

This article shows that respondents gain meaning from visual cues in a web survey as well as from verbal cues (words). We manipulated the layout of a five point rating scale using verbal, graphical, numerical, and symbolic language. This paper extends the existing literature in four directions: (1) all languages (verbal, graphical, numeric, and symbolic) are individually manipulated on the same rating scale, (2) a heterogeneous sample is used, (3) in which way personal characteristics and a respondent's need to think and evaluate account for variance in survey responding is analyzed, and (4) a web survey is used. Our experiments show differences due to verbal and graphical language but no effects of numeric or symbolic language are found. Respondents with a high need for cognition and a high need to evaluate are affected more by layout than respondents with a low need to think or evaluate. Furthermore, men, the elderly, and the highly educated are the most sensible for layout effects.

JEL codes: C42, C81, C93

Keywords: web survey, questionnaire lay out, context effects, need for cognition, need to evaluate

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

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

[1] The authors like to thank Don A. Dillman for his helpful comments provided on an earlier version of this article.

Ordinal scale questions are probably the most widely used measurement instrument used in web surveys. These questions are presented in various ways: categories can be presented in a single column or in multiple columns, in rows, with labels for all categories or the endpoint categories only, with radio buttons or an answer box, etc. Differences in layout yield detectable differences on responses to survey questions (Christian and Dillman, 2004; Tourangeau et al., 2004; Christian, 2003; Dillman and Christian, 2002). Christian, Dillman and Smyth (2005) suggest that writing effective questions for web surveys may depend as much or more on the presentation of the answer categories as the question wording itself.

While a theory of web questionnaire design may draw from the principles for the visual layout and design of paper questionnaires, it will also have new features and require independent testing and evaluation (Dillman et al., 1998). The cursor, the mouse, and the landscape orientation of monitors add dimensions that are different than those presented by the hand-eye coordination aspects involved in completing paper questionnaires. Despite the enormous use of web questionnaires, the knowledge of what people read and comprehend, and why, remains in its infancy (Redline et al., 2003). The understanding of the quality of respondent answers depends upon it.

In order to contribute to the development of a theory of questionnaire design for web surveys, this paper examines how verbal and visual languages influence answers to web surveys. Verbal, graphical, numeric, and symbolic languages are individually manipulated on a rating scale. The results in this paper are based on a representative sample of the Dutch population. We report results focusing on respondents with different characteristics, which has received little attention.

Background

The primary components that contribute to overall measurement error in a survey are the respondent, the data collection mode, the questionnaire, and if present, an interviewer. These components are interrelated, and interact during the measurement process (Biemer and Lyberg, 2003). A different mode of data collection can result in associated errors. While web surveys are conducted since the last decade, little is known about effects in questionnaires using the computer.

Survey researchers recognize the potential for alternative wordings of a question or of answer categories to affect the answers respondents provide. For example, the choosing of response categories can have a significant effect on respondent answers (see Schwarz et al., 1985; Schwarz and Hippler, 1987; Strack and Martin, 1987; Krosnick and Alwin, 1987, Rockwood et al., 1997, Toepoel et al., 2006). But not only verbal information can influence respondents, non-verbal information accounts for variances in survey responses as well. Context effects usually refer to an effect in which questions or response categories are read or presented (Schaeffer, 1992). Papers on these effects draw on social information-processing models of how people answer questions. Interpreting the question, retrieving information, generating an opinion or a representation of the relevant behavior, and reporting it are the main psychological components of a process that starts with respondents' exposure to a survey question and ends with their report (Sudman et al., 1996; Strack and Martin, 1987). In the next subsections these concepts will be discussed in more detail.

Interpreting the question

The first step in the question-answering process, interpreting the question, is to understand what is meant by the question. There must be a shared meaning between the researcher and the respondent with respect to each of the words in the question as well as the question as a whole. To comprehend the question, the respondent considers the question and attempts to understand what information is requested. In doing so, the respondent is lead by cues in the questionnaire.

Retrieving information

Given the respondent's understanding of the question, the respondent then goes to retrieve whatever information is necessary to respond to the question. Information needed to formulate a response is retrieved from memory. Some questions do not require the retrieval of factual data, but information may still be retrieved from memory in the form of feelings, viewpoints, positions on issues and so on (Biemer and Lyberg, 2003). The amount in which the respondent searches information for answering the question may differ because of the respondent's cognitive activity in answering the survey. Cacioppo and Petty (1982) developed a scale to measure the need for cognition. Need for cognition (NFC) represents the tendency for individuals to engage in and enjoy thinking. People with a high need for cognition (HNC) undergo different processes in formatting an answer than people with a low need for cognition (LNC). People with a HNC tend to seek more information and think more carefully about it than people with a LNC. People with a LNC are more easily influenced by peripherical cues.

Generating an opinion

In the third step of the question-answering process, the respondent is generating an opinion on the subject. This stage includes the process of reflecting on the issues raised by the questions in order to arrive at a report, attitude, belief, or opinion. Jarvis and Petty (1996) developed a measure to assess individual differences in the propensity to engage in evaluation, the Need to Evaluate Scale (NES). Although attitudes are a fundamental concept in psychology, little research exists on how the process of reflecting on issues can be used to predict meaningful mental and behavioral processes. Bizer et al. (2004) found that respondents high in need to evaluate (HNE) reported their answers more quickly than those low in the need to evaluate (LNE). Petty and Jarvis (1996) suggest that people with a LNC and LNE are expected to be more susceptible to various low effort biases than people with a HNC and HNE, such as being influenced by cues in a survey suggesting one response over another. On the other hand, Tormala and Petty (2001) found that HNE individuals formed attitudes in a spontaneous, on-line fashion, whereas LNE individuals formed them in a less spontaneous, more memory-based fashion. From this perspective, people with a HNE could be more susceptible to verbal and non-verbal cues in a survey. Evaluation by no means requires effortful thought. The relation between NES and NFC was tested by Jarvis and Petty and was found to be moderate and positive ($r=.35$, $p<.001$).

Formatting a report

Following the opinion-stage, the next stage of the response process is referred to as the response formatting process. Answers to survey questions have to be reported in a format that is provided by the survey researcher. This format contains verbal and nonverbal cues that influence respondent behavior. Nonverbal cues include numeric,

symbolic and graphical languages that convey meaning in addition to the verbal language (Dillman and Christian, 2002). A conceptual framework for explaining how visual languages may influence respondent behavior has been provided by Jenkins and Dillman (1997). Verbal and nonverbal cues can independently and jointly influence answers to questions. Redline et al. (2003) confirm in their study on item non-response in self-administered paper questionnaires that the visual and verbal complexity of information on a questionnaire affects what respondents read, the order in which they read it, and ultimately, their comprehension of the information. Dillman and Christian (2002) found that manipulating several aspects of the visual languages simultaneously significantly changed respondent behavior in a paper questionnaire. In 2004 Christian and Dillman individually manipulated graphical and symbolic languages, and found significantly different answers in their manipulations.

Graphical language There are a number of studies conducted on the influence of visual layout on self-administered questionnaires. Christian and Dillman (2004), in their study of graphical and symbolic languages, show that the visual design of questions on self-administered questionnaires could significantly impact respondent behavior. Friedman and Friedman (1994) demonstrated that equivalent horizontal and vertical rating scales in paper questionnaires do not elicit the same responses. Christian (2003) compares a vertical linear layout of scalar questions to nonlinear layouts. The linear version produced significantly different responses from the nonlinear versions. The triple- nonlinear versions produce greater use of the response option in the middle of the top line, just to the right of the first option regardless of the labels given to the category. The addition of numbers on the nonlinear vertical versions did not seem to significantly change how respondents answered any of the

questions. Friedman and Leefer (1981) find that in scalar questions respondents seem to respond to the labels rather than to the position of an answer category relative to the endpoints. They suggest further research on rating scale design should be conducted, in order to determine the relative importance of context effects due to verbal and visual cues, using a sample other than students.

Numeric language Schwarz et al. (1985) show that respondents gain information about the researcher's expectations using the numeric labels as frames of reference. Further, Schwarz et al. (1991) find that changing the numeric values attached to scales resulted in different respondents' answers. Respondents may use numerical language as additional meaning to the verbal labels of the scale.

Symbolic language Symbolic language uses symbols that have cultural meaning to convey information to respondents. Schwarz et al. (1991) find that respondents hesitate to assign a negative score in a face-to-face interviews to themselves. An eleven point scale with numbers 0-10 resulted in lower scores than a –5 to 5 format. Negative signs on the scale influence respondents' interpretation of the endpoint labels.

Effect of personal characteristics

The extent to which personal characteristics, such as education, age and gender affect respondents' performance is relatively unknown. Couper (2000) argues that design may interact with the type of web survey conducted and the population at which the survey is targeted. McFarland (1981) did not find evidence that personal characteristics might interact with the ordering of questions. The effects of question

order were consistent for both sexes and across education levels. Toepoel et al. (2006), in their study on response category effects in a web survey, find men to be more affected by cues than women. Also, younger people were more distracted by cues than older people, although people of 65 year and older show the highest deviation in reports between a high and low response scale. Krosnick and Alwin (1987) find respondents with less education and more limited vocabularies to be influenced more by different answer categories.

Literature suggests that additional research on the visual design of web questionnaires is needed to develop more general principles of how the visual layout of questions influences answers to web surveys (Dillman et al., 2005; Christian and Dillman, 2004; Dillman and Christian, 2002; Jenkins and Dillman, 1997; Schwarz et al.,1991; Friedman and Leefer, 1981). Such work is essential for effective survey construction and offers the possibility for methodological improvements of survey research.

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 numeric 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. Because a researcher has so many possibilities for presenting a 5-point scalar question, this type of question is used to manipulate verbal, graphical, numeric, and symbolic cues. These languages were individually manipulated as suggested by Redline et al. (2003).

Two experiments using eight different formats were used in the CentERpanel, an online household panel consisting of more than 2,000 households administered by CentERdata. This panel is representative for the Dutch population (see Appendix B for more details about the CentERpanel). Because not all people own a computer or have access to Internet, CentERdata provides a set-top box (and, if necessary, a television) for people who do not have a computer to make it possible to complete the questionnaires online. Two questions were used measuring the quality of education (1) and life (2) in the Netherlands. These questions were based on an experiment conducted by Christian (2003), who measured the quality of education and the quality of student life at Washington State University. The study was conducted in week 37 (September) and week 41 (October) 2005. The response percentage was 78.3%[2] (2787 were selected, 2182 people responded) for the first experiment and 78.8% (2830 were selected, 2229 people responded) for the second. The first group of each experiment answered a rating scale with answer categories excellent, very good, good, fair, and poor in a linear vertical format from positive to negative. In the first experiment 3 different manipulations were used, in the second experiment 4 (see Appendix A).

The first experiment is a replication of an experiment done by Christian (2003), to find out if similar results occur using a representative sample. We compared a linear vertical format (Appendix A: 1a) to two non-linear formats: a triple banked format with options running horizontally (Appendix A: 1b) and a triple banked format with options running vertically (Appendix A: 1c). To test whether numbers would help reading the triple vertical format, a fourth group answered the questions in a triple vertical format with numbers (Appendix A: 1d).

[2] Response Rate 1 defined in the Standard Definitions of AAPOR (www.aapor.org). Note that this definition is not primarily designed for an online panel. We compared the personal characteristics of respondents and non-respondents, and concluded that non-response in this study was non-selective.

In the second experiment, the panel was randomly divided in five groups. The first group answered the rating scale in a linear vertical format from positive to negative (Appendix A: 2a). This group served as a reference group. All other groups have individually different 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). The fifth group was offered a symbolic manipulation. The numbers 5 to 1 were added in the education question, while in the life question the numbers varied from 2 to minus 2 (Appendix A: 2e).

The objective was to learn which respondents are more sensitive to verbal and to non-verbal cues. Therefore, scores of different gender, age and education groups were compared. Because research indicates that the need for cognition and the need to evaluate construct account for independent variation in survey responding, we take this into the analysis[3]. NFC is measured through a scale consisting of 34 items. The NES scale is measured through 16 items. The mean score on the scale defines the distinction between a high and a low NFC and NES group.

Results

Linear versus Non-Linear Layout

In the first question, respondents rated the quality of education in the Netherlands. See table 1 for response distributions and some tests that were carried out. Results from the chi square tests indicate differences in individual responses across formats and results from the t-tests indicate differences in the mean number of responses. We use

[3] Wording of the items is available upon request.

these tests to stay in line with previous research. Lower mean scores indicate more positive ratings (1 = the first category and 5 = the last category).

[table 1]

The overall chi-square test indicates significant differences in the responses across all four versions (χ^2 =33.86, p=.00). Individual tests show that the linear version has significantly different responses and mean scores than the triple versions. Respondents rated the education in the Netherlands significantly more favorably on the linear version than on all the nonlinear versions, indicating a primacy effect. The second response option “very good” is selected more often in the linear format; while the fourth option “fair” is selected less than in the triple formats. Comparing the triple horizontal and triple vertical format, respondents select the response option “very good” more often in the triple horizontal format as opposed to the triple vertical format (respectively 12.9% versus 10.8%), while the response option “good” is more often chosen in the triple vertical format (44.0% in the triple horizontal format and 52.1% in the triple vertical format), supporting the hypothesis that respondents more easily select the answer right next to the first option on the first line. In the triple horizontal version, the option “fair” is selected more than in other versions. Stimulated to read horizontally, the first option on the second line is chosen more often. We therefore did not find evidence that respondents fail to read the second line in a non-linear format. Adding numbers to the vertical format did significantly change how respondents answered the question, but the difference in means between the triple vertical format and the triple vertical format with numbers is not significant (χ^2 =9.30, p=.05, t=1.12, p=.26). Table 2 shows similar results in the second question

about life in the Netherlands. Our results are in line with the experiment conducted by Christian (2003) and other research (Christian and Dillman, 2004; Dillman and Christian, 2002).

[table 2]

Verbal and Visual Manipulations of Layout

Verbal Table 3 and 4 show statistically different answer distributions and mean scores in a negative-positive format in relation to a positive-negative format, indicating that respondents are affected by verbal language.

[table 3]

[table 4]

Chi square tests indicate significant differences in the responses across the two versions (χ^2 =14.76, p=.01 in the education question, and χ^2 =103.79, p=.00 in 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 in pos/neg format and 3.28 in neg/pos format in the education question and respectively 2.60 and 2.88 in the life question). Different responses result in selecting the second response alternative more, supporting a primacy effect. The response option “very good” is selected by 24% when it is presented as second alternative, and by 10.7% when it is presented as fourth alternative. The option “fair” is chosen by 31.1% when it is presented as second alternative and by 16.5% as fourth alternative in the education question. Despite the label, the second option is selected more often. The same results are found in the life question.

Graphical Chi square tests indicate significant differences in the responses across the vertical and horizontal versions (χ^2 =10.43, p=.04 in the education question, and χ^2 =71.92, p=.00 in the life question), but the mean scores do not statistically differ (t=-1.82, p=. 07 in the education question and t=-1.80, p=. 07 in the life question). Differences result in selecting the fourth option "fair" in the horizontal format more. Thus, in the horizontal format a shift to the left is not detected. Respondents may be more willing to read all options in the horizontal format (assuming respondents first read horizontally before they read vertically). Therefore, a primacy effect is more likely to emerge in a vertical format. Lower mean scores in the vertical format support this hypothesis, but these differences do not reach statistical significance.

Numeric No evidence was found that the adding of numbers 1 to 5 causes different responses. Chi square tests indicate no significant differences in the responses across the linear version and the linear versions with numbers 1 to 5 (χ^2 =.58, p=.97 in the education question, and χ^2 =13.29, p=.10 in the life question). No differences of mean scores were found (t =.55, p=.58 in the education question, and t =1.08, p=.28 in the life question). The numbers 1 to 5 are probably seen as answer category numbers, so respondents do not see these numbers as an additional meaning to the verbal labels.

Symbolic Comparing the numbers 1 to 5 and 5 to 1 in the first question, we do not find significant differences. The mean score in the 5 to 1 version is lower than in the 1 to 5 version (respectively 2.88 and 2.94), indicating that respondents select a

positive answer more easily when a higher number is added. The different mean scores do not statistically differ, however (t=1.07, p=.29). The mean score in the 2 to –2 format (2.54) is lower than the mean score in the 1 to 5 format (2.64), although this difference does not reach significance (t=1.85, p=.07) either. The chi square test indicates no significant differences in the responses across the two versions (χ^2 =7.03, p=.14). Therefore, we did not find statistical evidence for respondents to be more eager to assign positive scores.

Finding information and generating an opinion: Need For Cognition and Need to Evaluate

In this section we discuss whether there are significant differences between formats for respondent with a high or a low NFC/NES in the two experiments. The strength of the differences is presented using eta as measure of association. The results for the education question are discussed in the text, while tables 5 to 8 present the results for the life question in more detail.

In our first experiment, we did not find significant differences in the education question for respondents with a high NFC/NES or a low NFC/NES. Differences for the whole population are not found for homogeneous subsets for levels of cognition and evaluation. In the life question, only respondents with a low NFC show significant differences in answers between all four formats.

In the education question in the second experiment, we find little evidence for an effect of need for cognition and need to evaluate as well. In the verbal manipulation, respondents with a high need for cognition report different answer scores in the positive to negative format in relation to the negative positive format (eta=.188). Apparently they try to find information on the spot, influenced by verbal

cues in the questionnaire. Respondents with a different need to evaluate do not report different answers in the verbal manipulation. The overall test, across all formats, shows different answer scores for respondents with a high need for cognition (eta=.177) and respondents with a high need to evaluate (eta=.089). In contradiction to our expectations, respondents with a high rather than a low score on the NFC and NES are affected by layout.

In the life question we find more significant differences for respondents with different levels of NFC and NES (see table 5). The overall test across all formats shows different answer scores for all NFC/NES groups. Again, respondents with a high score on the NFC and NES are more affected by layout. In the verbal manipulation, respondents with a high need for cognition are more sensible to verbal cues than respondents with a low need for cognition (eta=.388 vs. eta=.310). We also find differences for need to evaluate, although the strength of the relationship is somewhat similar (eta=.340 for high NES and .340 for low NES). The life question also reports differences for the graphical manipulation. High NFC (eta=.335) and high NES (eta=.315) are more affected by graphical manipulations than low NFC (eta=.246) and low NES (eta=.315). Our results are in line with Tormala and Petty (2001). High NFC/NES individuals seem to form attitudes in a spontaneous, on-line fashion, whereas low NFC/NES individuals form them in a less spontaneous, more memory-based fashion. Therefore, high NFC/NES individuals are more sensitive for cues that suggest one response option over another.

[table 5]

Effects of gender, age and education

In Toepoel et al. (2006) it is found that men are more sensitive for context effects than women. In our first experiment, significant differences across all 4 formats for men in both questions are found, while women do not report statistically different answers across all formats. Men select the second response option more often in the linear format, and the third option less in the triple horizontal version.[4]

In an overall test across all formats for our second experiment, we find significant differences for men and women in the education question: eta=.128 for men and eta = .123 for women. Analyzing the different formats, we did not find evidence that gender affected answers on the verbal, numeric, and symbolic manipulation. However, women answer differently in the graphical manipulations (vertical versus horizontal). As one can see in table 6, this holds for the life question as well. In this question we find large differences between men and women (eta=.275 versus eta=.322), but this time women report higher differences between formats. Our results indicate that women are more influenced by verbal and graphical manipulations in a more personal question than men. Further research on verbal and non-verbal manipulations with different question types can make this effect more clear.

[table 6]

With regard to the effect of age, the effect of layout seems to decrease with age until the age of 55. As of then, the effect increases. In the first experiment, respondents older than 55 years select the fourth response option 'fair' in the linear

[4] This option is presented at the right of the screen in the triple horizontal format (see Appendix A 1b).

format less. In the triple horizontal format, they select ‘fair’ (presented right under the first option) more often, and they select ‘good’ (presented at the utmost right) less. In the second experiment, respondents of 65 years and older select the second response alternative more often in the negative to positive format. More details can be seen in table 7. Our results hint at an increase in context effects due to decreases in cognitive functioning.

[table 7]

Looking at the respondent’s education level, the largest differences between formats are found for respondents with a university degree (see table 8). Because previous research is mostly based on a student population, this research shows that context effects found on a student population may not apply to the population as a whole. We did not find that respondents with lower secondary education are the least affected, but the effect of layout is relatively small for this education group.

[table 8]

Ordinal regression with the linear vertical format as reference level shows significant interaction effects of format with gender, age, and education in both experiments.

Discussion and Conclusions

This article shows that respondents gain meaning from non-verbal cues in a web survey as well as from verbal cues. We manipulated the layout of a five point scalar

question in two experiments using two questions. In the first experiment, a linear layout with three non-linear layouts was compared. In the second experiment we manipulated verbal, graphical, numeric, and symbolic language individually, to learn how these verbal and non-verbal cues influence respondents' answers in rating scales. This paper extends previous research as verbal, graphical, numeric, and symbolic languages are individually manipulated on the same rating scale, a representative sample is used, it is analyzed in which way personal characteristics and a respondent's need to think and evaluate account for variance in survey responding, and because the experiment is based on a web survey.

In the linear versus non-linear versions we find differences across all versions. Triple horizontal and triple vertical format show significant different means than the linear format. In a triple visualization, respondents are more eager to select the second answer on the top line. Our results support a primacy effect in answering scalar questions. Options that require less movement of the mouse might be more easily chosen than answers requiring more hand/eye movements. The addition of numbers to stimulate respondents to read vertically did not influence mean scores. Our results are in line with the experiment conducted by Christian (2003).

In experiment 2, again different correlations and mean scores are found between the five different manipulations. The verbal manipulation shows significant different means than the other manipulations. This indicates that a negative tone of the first option deviates reports in a negative manner. Despite the label, respondents select the second option more often. Different results are also found comparing the non-verbal manipulations with each other in the question about life in the Netherlands: large statistical differences are found caused by the graphical manipulation. Changing the answer categories to a horizontal format changes answer scores. Respondents may

be more willing to read all options in the horizontal format (because respondents may first read horizontally before they read vertically). The addition of numbers 1 to 5 to the vertical format did not influence respondent answers. Comparing the numbers 1 to 5 versus 5 to 1 and 1 to 5 versus 2 to -2, no significant differences due to symbolical language occur. Thus, no evidence was found of respondents being less eager to assign negative scores in a five point rating scale.

But which format shows the least deviation to the overall scores across al formats? Looking at the mean scores in the different formats, one format has almost exactly the same mean score for the education question: the symbolic manipulation with numbers 5 to 1 (where 5 is the most positive). Adding numbers, with the highest number for the most positive score, seems to validate the scale in this question type. But, in the life question the graphical manipulation has the closest mean to the overall score. Because all other formats have a vertical format, the conclusion that the horizontal format has the closest mean to the overall score is remarkable. While we already have seen that this format is also the least sensible for primacy effects, it could be that presenting a 5 point scale horizontally makes sure that respondents read the answer categories more accurate, therefore decreasing the influence of layout. Further research in web surveys on a horizontal layout of scalar questions in different contexts can make this effect more clear.

The effect of format is not the same for respondents with different personal characteristics. Respondents with a high need for cognition (NFC) and a high need to evaluate (NES) are more sensible for verbal and visual cues. Apparently they think and evaluate in an ongoing online process, influenced by cues in a questionnaire. This is in line with results of Tormala and Petty (2001). Men, the elderly, and the highly educated are more sensitive for layout effects. Deriving conclusions on a student-

based sample might show more differences between different formats than a heterogeneous sample of the population. Future research should be conducted comparing student based and representative samples to find out if studies using students show more significant results.

This paper shows that the visual presentation of answer categories must be taken into consideration in order to reduce measurement error. This goes especially for researchers who want to compare results across surveys. Similarly worded questions may be presented to respondents in visually dissimilar ways. Do different results then come from a different time of measurement or from a different visualization? This is a challenge for further research.

References

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. Wiley series in Survey Methodology, New Jersey.

Bizer, George Y., Jon A. Krosnick, Allyson L. Holbrook, S. Christian Wheeler, Derek D. Rucker, and Richard E. Petty. 2004. "The Impact of Lifeity on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate". *Journal of Lifeity* 72:996-1028.

Cacioppo, John T., and Richard E. Petty. 1982. "The Need for Cognition". Journal of Lifeity and Social Psychology 42:116-131.

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

Christian, Leah, M. 2003. The Influence of Visual Layout on Scalar Questions in Web Surveys. Unpublished Master's Thesis. Retrieved 10-25-2005 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.

Christian, Leah M., Don A. Dillman, and Jolene D. Smyth. 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. Retrieved 20-02-2006 on http://survey.sesrc.wsu.edu/dillman/papers.htm

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 10-25-2005 on http://survey.sesrc.wsu.edu/dillman/papers.htm

Dillman, Don A., Arina Gertseva, and Taj Mahon-Haft. 2005. "Achieving

Usability in Establishment Surveys Through the Application of Visual Design Principles". *Journal of Official Statistics* 21:183-214.

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

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.

Friedman, Hershey H., and Joanna R. Leefer. 1981. "Label Versus Position in Rating Scales". Journal of the Academy of Marketing Science 9:88-92.

Jarvis, W Blair G., and Richard E. Petty. 1996. "The Need to Evaluate". *Journal of Lifeity and Social Psychology* 70:172-194.

Jenkins, Cleo R., and Don A. Dillman. 1997. " Towards a theory of Self-administered Questionnaire Design". In: Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathrijn Dippo, Norbert Schwarz, and Dennis Trewin (Eds). *Survey Measurement and Process Quality (pp165-196).* Wiley series in probability and statistics, New York.

Krosnick, Jon A., and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement". *Public Opinion Quarterly* 51: 201-219.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997."Designing Rating Scales for Effective Measurement in Surveys". In: Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathrijn Dippo, Norbert Schwarz, and Dennis Trewin (Eds). *Survey Measurement and Process Quality (pp141-164).* Wiley series in probability and statistics, New York.

McFarland, Sam G. 1981. "Effects of Question Order on Survey Responses". *Public Opinion Quarterly* 45:208-215.

Petty, Richard E., and W. Blair G. Jarvis. 1996. "An Individual Differences Perspective on Assessing Cognitive Processes". In: Norbert Schwar, and Seymour Sudman (Eds). *Answering Questions (pp221-257).* Jossey-Bass Publishers, San Francisco.

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

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

Schaeffer, Nora C. 1992. "Context Effects in Social and Psychological Research". *The Public Opinon Quarterly* 57:280-283.

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

Schwarz, Norbert and Hans-J. Hippler. 1987. "What response scales may tell your respondents: informative functions of response alternatives ". In: Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman (Eds), *Social Information Processing and Survey Methodology (pp 163-178).* Springer-Verlag, New York.

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

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. *The Public Opinon Quarterly* 55:570-582.

Strack, Fritz and Leonard L. Martin. 1987. "Thinking, Judging, and communicating: a process account of context effects in attitude surveys". In: Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman (Eds), *Social Information Processing and Survey Methodology* (pp 163-178). Springer-Verlag New York.

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

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.* Retrieved on January 9, 2006 on http://greywww.kub.nl:2080/greyfiles/center/ctr_py_2006.html

Tormala, Zakary L. and Richard E Petty. 2001. "On-Line Versus Memory-Based Processing: The Role of "Need to Evaluate" in Person Perception". *Pers Soc Psychol Bull* Vol. 27 No.12: 1599-1612.

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.

Table 1. Experiment 1. Education Question: Frequencies (in %), mean scores, correlations and mean differences in the different formats

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

	Linear	Nonlinear - Triple		
		Horizontal	**Vertical**	**Vertical with Numbers**
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

	Chi Square Tests χ^2	**Diff. Of means t**
Linear versus Triple horizontal	20.69**	-4.20**
Linear versus Triple vertical	16.12**	-3.44**
Linear versus Triple vertical with numbers	5.43	-2.14*
Triple vertical versus Triple horizontal	7.66	-0.93
Triple vertical versus Vertical with numbers	9.30*	1.12
Overall-across all 4 formats	33.86**	F= 6.71**

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

Note: A high mean score indicates a negative judgment.

Table 2 Experiment 1. Life Question: Frequencies (in %), mean scores, correlations and differences of means in the different formats

Overall, how would you rate the quality of life in the Netherlands?

	Linear	Nonlinear - Triple		
		Horizontal	**Vertical**	**Vertical with Numbers**
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 χ^2	**Diff. Of means t**
Linear versus Triple horizontal	27.32**	-5.12**
Linear versus Triple vertical	12.84**	-3.26**
Linear versus Triple vertical with numbers	12.19*	-2.07*
Triple vertical versus Triple horizontal	8.49	-2.02*
Triple vertical versus Vertical with numbers	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.

Table 3 Experiment 2. Education Question: Frequencies (in %), mean scores, correlations and differences of means in the different formats

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

	Reference: Linear Vertical Positive to Negative	**Verbal:** Linear Vertical Negative to Positive	**Graphic:** Linear Horizontal	**Numeric:** Linear Vertical With Numbers 1 to 5	**Symbolic:** Linear Vertical With Numbers 5 to 1
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 χ^2	**Diff. Of means** **t**
Verbal: Positive to Negative versus Negative to Positive	14.76**	-7.17**
Graphic: Linear Vertical versus Linear Horizontal	10.43*	-1.82
Numeric: Linear Vertical versus Linear Vertical With Numbers 1 to 5	.58	.55
Symbolic: 1 to 5 versus 5 to 1	2.51	1.07
Overall across all non-verbal manipulations (without linear neg to pos)	15.97	F=1.98
Overall across 5 all formats	47.68**	F= 8.74**

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

Note: A high mean score indicates a negative judgment.

Table 4 Experiment 2. Life Question: Frequencies (in %), mean scores, correlations and differences of means in the different formats

Overall, how would you rate the quality of life in the Netherlands?

	Reference: Linear Vertical Positive to Negative	**Verbal:** Linear Vertical Negative to Positive	**Graphic:** Linear Horizontal	**Numeric:** Linear Vertical With numbers 1 to 5	**Symbolic:** Linear Vertical With numbers 2 to -2
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 χ^2	**Diff. Of means** **t**
Verbal: Positive to Negative versus Negative to Positive	103.79**	-5.50**
Graphic: Linear Vertical versus Linear Horizontal	71.92**	-1.80
Numeric: Linear Vertical versus Linear Vertical With Numbers 1 to 5	13.29	1.08
Symbolic: 1 to 5 versus 2 to -2	7.03	1.85
Overall across all non-verbal manipulations (without linear neg to pos)	115.16**	F=32.01**
Overall across 5 all formats	220.57**	F= 52.27**

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

Note: A high mean score indicates a negative judgment.

Table 5. Overview of significance (chi square) and association (eta) between formats for Need for Cognition and Need to Evaluate in the life question

Exp. 1	**Linear versus Triple hori-zontal**	**Linear versus Triple vertical**	**Linear versus Triple vertical with numbers**	**Triple vertical versus Triple horizontal**	**Triple vertical versus Vertical with numbers**	**Overall-across all 4 formats**
NFC						
1 low	.162**	.022	.059	.145**	.039	.130*
2 high	.162	.165*	.093	.000	.062	.134
NES						
1 low	.179*	.096	.066	.096	.026	.132
2 high	.162*	.099	.097	.067	.061	.115
Exp. 2	**Verbal:**	**Graphic:**	**Numeric:**	**Symbolic:**	**Overall-across all 5 formats**	
NFC						
1 low	.310**	.246**	.041	.072	.252**	
2 high	.388**	.335**	.025	.076	.352**	
NES						
1 low	.344**	.253**	.019	.031	.299**	
2 high	.340**	.315**	.009	.096	.301**	

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

Note: A higher correlation coefficient (eta) indicates greater differences between formats.

Table 6. Overview of significance (chi square) and association (eta) between formats for gender in the life question

Exp. 1	**Linear versus Triple hori-zontal**	**Linear versus Triple vertical**	**Linear versus Triple vertical with numbers**	**Triple vertical versus Triple horizontal**	**Triple vertical versus Vertical with numbers**	**Overall-across all 4 formats**
men	.171**	.141**	.044*	.039	.088**	.137**
women	.137**	.053	.081	.088	.054	.099
Exp. 2	**Verbal:**	**Graphic:**	**Numeric:**	**Symbolic:**	**Overall-across all 5 formats**	
men	.308**	.262**	.490	.098	.275**	
women	.353**	.284**	.210	.018	.322**	

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

Note: A higher correlation coefficient (eta) indicates greater differences between formats.

Table 7. Overview of significance (chi square) and association (eta) between formats for age in the life question

Exp. 1	**Linear versus Triple hori-zontal**	**Linear versus Triple vertical**	**Linear versus Triple vertical with numbers**	**Triple vertical versus Triple horizontal**	**Triple vertical versus Vertical with numbers**	**Overall-across all 4 formats**
15-24	.273*	.172	.078	.114	.101	.208
25-34	.177	.191	.042	.007	.142	.164
35-44	.101	.237	.015	.023	.106	.104
45-54	.079	.057	.120*	.023	.062	.086
55-64	.169	.019	.074*	.187*	.091*	.150*
>64	.215*	.127**	.080*	.116	.041*	.174**
Exp. 2	**Verbal:**	**Graphic:**	**Numeric:**	**Symbolic:**	**Overall-across all 5 formats**	
15-24	.423	.251	.072	.015	.372**	
25-34	.436**	.241	.035	.023	.368**	
35-44	.395**	.216	.126	.153	.360**	
45-54	.167	.197**	.012	.035	.180*	
55-64	.390**	.220	.170	.180	.367**	
>64	*.225*	.444**	.111	.062	.319**	

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

Note: A higher correlation coefficient (eta) indicates greater differences between formats.

Table 8. Overview of significance (chi square) and association (eta) between formats for age in the life question

	Linear versus Triple hori-zontal	**Linear versus Triple vertical**	**Linear versus Triple vertical with numbers**	**Triple vertical versus Triple horizontal**	**Triple vertical versus Vertical with numbers**	**Overall-across all 4 formats**
primary	.125	.068	.122	.179	.056	.190
lower secondary	.180*	.075	.192*	.110	.121	.161
higher secondary	.098	.113	.037	.020	.068	.090
Inter-mediate vocational	.049*	.008	.014	.044	.023	.049
higher vocational	.132	.085	.021	.049	.102	.122
university	.337**	.379**	.174	.019	.199	.304**
	Verbal:	**Graphic:**	**Numeric:**	**Symbolic:**	**Overall-across all 5 formats**	
primary	.380*	.264*	.022	.178	.283	
lower secondary	.196**	.235**	.015	.061	.200**	
higher secondary	.217**	.265*	.102	.083	.285**	
Inter-mediate vocational	.242*	.248*	.017	.040	.239*	
higher vocational	.475**	.276**	.109	.074	.425**	
university	.495**	.410**	.047	.148	.427**	

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

Note: A higher correlation coefficient (eta) indicates greater differences between formats.

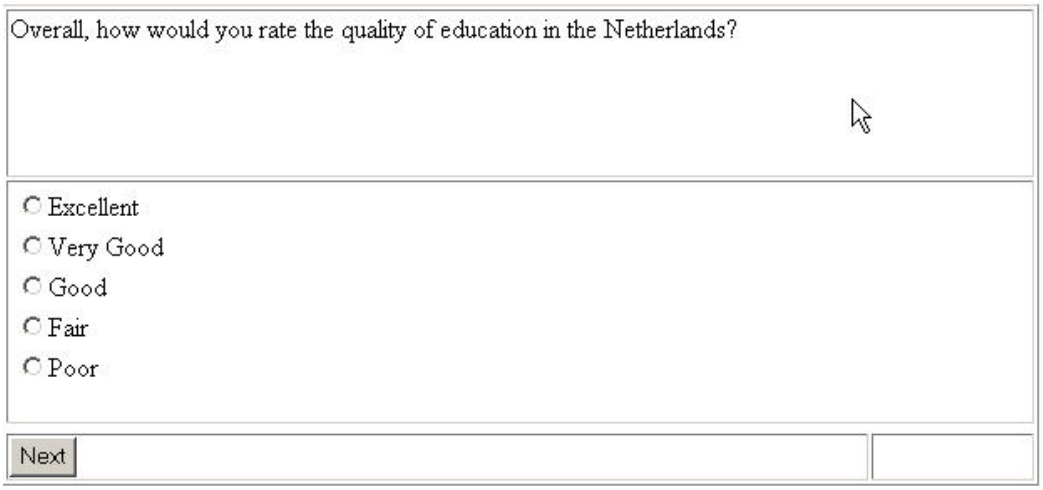
Appendix A: screenshots

Experiment 1
Four different layouts were used, using a linear and a non-linear format, in two questions, namely

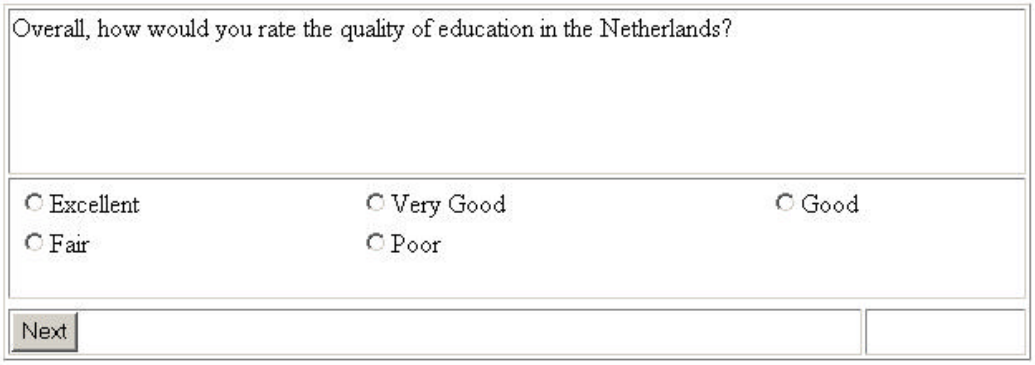
1. Overall, how would you rate the quality of education in the Netherlands?
2. How would you rate the quality of life in the Netherlands?

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

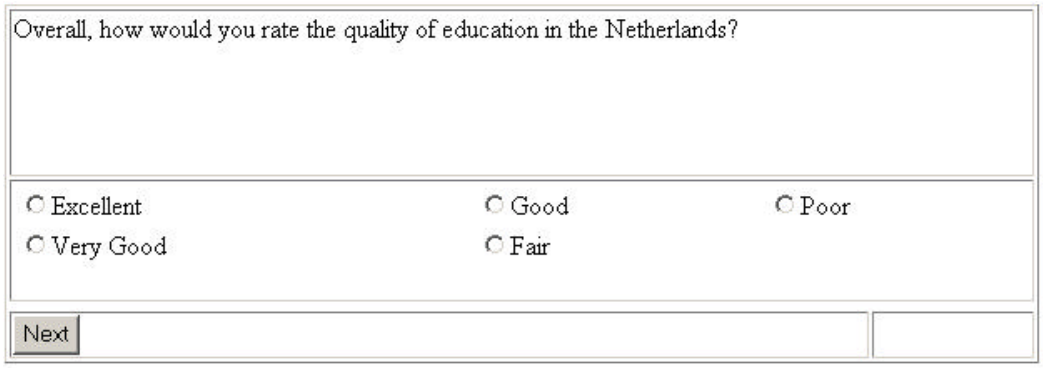
1a. Linear



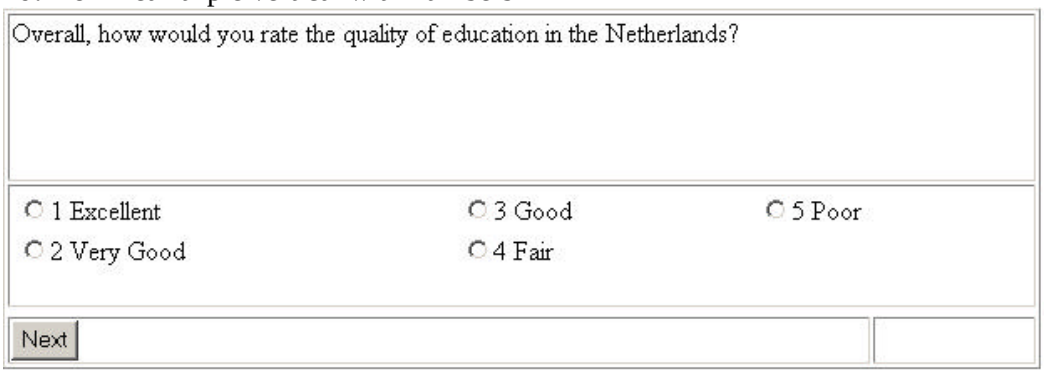
1b. Nonlinear - triple horizontal



1c. Nonlinear – triple vertical



1d. Nonlinear-triple vertical with numbers



Experiment 2

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

Format a: reference format (see 1a);

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

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

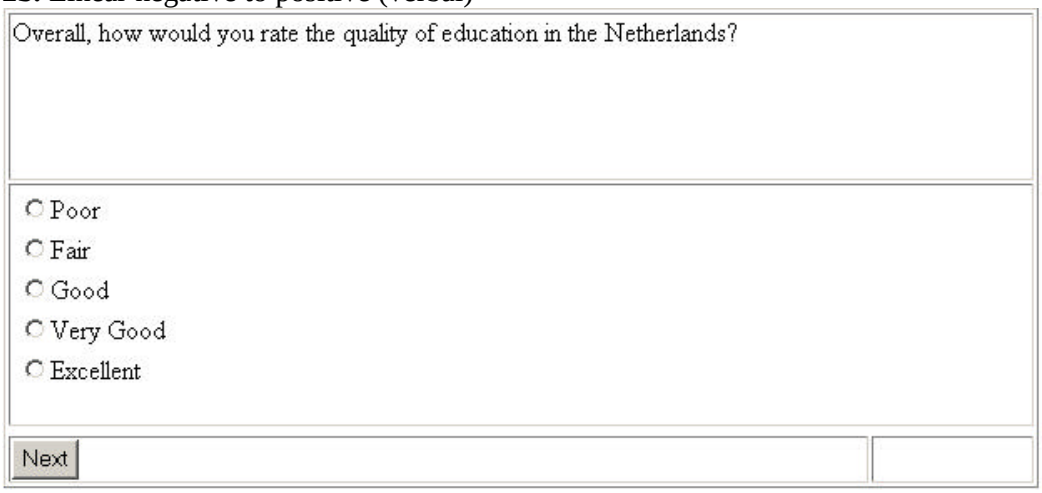
Format d: numeric manipulation: numbers 1 to 5 are added in this format;

Format e: symbolic manipulation: numbers 5 to 1 are added in this format (for the education question; numbers 2 to –2 for the life question).

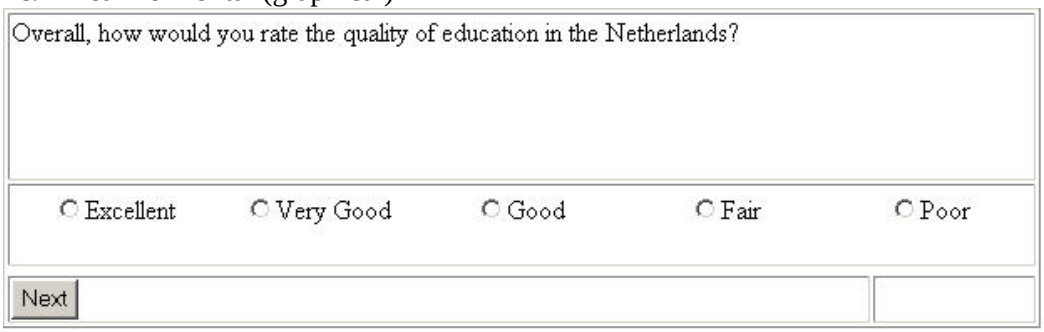
The screenshots below show the different layout formats for the education question, the layout formats used in the life question are the same except for the symbolic manipulation (see above).

2a. Linear positive to negative
See screen dump 1a.

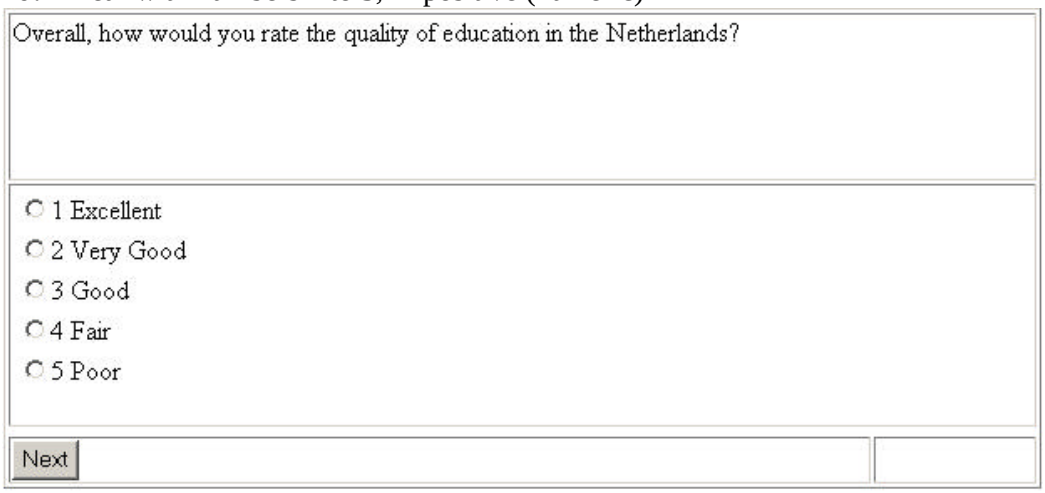
2b. Linear negative to positive (verbal)



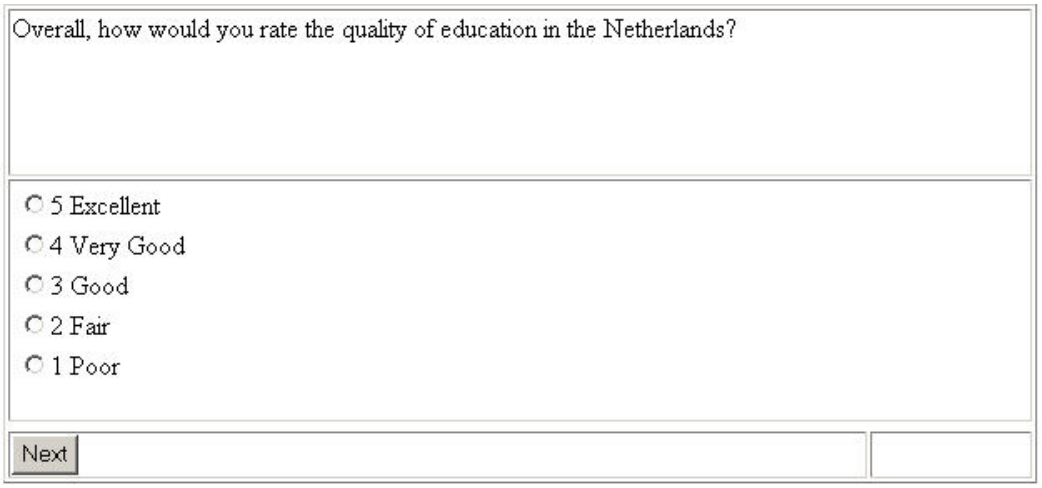
2c. Linear horizontal (graphical)



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



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



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

Appendix B

This Appendix presents the selection procedure of panel members.

The CentERpanel consists of over 2000 households in the Netherlands, the members of which fill in a questionnaire at their home computers every week. The CentERpanel is representative of the Dutch population.

The recruitment of new panel members consists of several stages. In the first stage, a random sample of candidates is interviewed by telephone. In the first telephone interview a number of questions are asked about the demographic characteristics of the household. The interview is concluded with the question whether the person would like to participate in survey research projects. If so, the household is 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 of the Dutch population). The selected household is asked whether the members of the household would like to become panel members, and if so, a number of additional questions are asked

Although the CentERpanel is an Internet-based panel, there is no need to have a personal computer with an Internet connection. Those households, who don't have access to Internet, are provided with a so-called set-top box, with which a connection can be established via a telephone line and a television set. If the household doesn't have a television, CentERdata provides one also.