# Effects of Questionnaire Design on Respondent Experience and Data Quality

Master Thesis Proposal

Ana Levordashka

VU University Amsterdam

Supervisor: Dr. Daniël Lakens, Eindhoven University of Technology

## Introduction

Psychometric assessment based on self-report is widely used in academic and non-academic settings. Usually in the form of a questionnaire, it can be conducted in various ways, including interview, pen and paper, and, more recently, online. The quality of the data acquired through self-report is highly contingent upon the respondents' capacity and willingness to provide reliable answers. When respondents are interested and focused, they will be more likely to think carefully about the questions and state their true opinions. When respondents are unmotivated, bored, or fatigued, on the other hand, they will be more likely to give meaningless or even random responses, which, in turn, add noise to the data (Oppenheimer et al., 2009), leading to measurement errors and poor data quality (Krosnick, 1999; Krosnick et al., 2002). Furthermore, fatigue and boredom can increase study abandonment (dropout) and thus lead to sampling biases (Cook, Heath, & Thompson, 2000), which pose serious problems for studies that rely on representative sampling (e.g., public opinion surveys) or random assignment (e.g., experiments), especially if attrition is non-random.

Interestingly, respondent engagement and behavior are not entirely outside of researchers' control. Problematic respondent behaviors are associated with properties of the assessment, ranging from content-related factors, such as length and complexity, to content-independent factors, such as layout, formatting, and visual design of questionnaires. Empirical evidence for the effects of various factors is reviewed in the upcoming sections.

Research on factors underlying respondent engagement and behavior typically results in suggestions on how to intervene on problematic behavior, for example through reducing study length or optimizing visual layout. Making content changes, such as shortening established psychometric measures, can compromise study quality, whereas changes to design and layout carry little risk. Furthermore, design-related solutions are cost-effective, which makes them a desirable form of intervention (Couper, 2000).

Despite the advantages of understanding the effect that design variables have on questionnaire respondents and researchers' explicit recommendations of doing so (Couper, 2000), experimental research on the topic has been scarce. The present investigation combines psychological research and principles of user interface design to investigate the role of questionnaire layout in respondent behavior and data quality. Although problems with respondent engagement apply to all forms of psychometric assessment, the present investigation focuses entirely on online questionnaires, both because of the increasing interest in the method and because of its enormous potential (Kiesler & Sproull, 1986; Reips, 2000; Gosling, Vazire, Srivastava, & John, 2004; Buchanan, Johnson, & Goldberg, 2005).

## Literature review

### Problematic respondent behavior

Research has identified a number of problematic behaviors in self-reported assessment, which can roughly be classified as instances of disengaged responding (often grouped under the term "satisficing;" Krosnick, 1991) or total survey abandonment ("dropout"). Giving a meaningful answer to a question, involves cognitive steps of comprehending the question, retrieving relevant information, and formulating an answer according to the scale requirements (Tourangeau & Rasinski, 1988). Instead of completing these steps and thinking carefully about a question, respondents can also use heuristics and immediate cues to conceive an acceptable response ("mild satisficing;" Krosnick, 1991) or they can give an entirely random responses without reading the question ("severe satisficing;" Krosnick, 1991). Severe satisficing can become apparent in visually non-random response patterns, such as uniform straight lines if respondents answer questions by always selecting one and the same location on the response scale (straightlining), or other non-random patterns (Herzog & Bachman, 1981). Satisficing can also be detected through instructional manipulation checks, where participants are asked to leave a question blank (Oppenheimer, Meyvis, & Davidenko, 2009). Other indexes of disengagement include taking less time to consider a question (speeding), opting for easy responses such as “not applicable” or “no opinion” (nonresponse; Krosnick et al., 2002), or entirely skipping optional questions (Heerwegh & Loosveldt, 2008). One study revealed that disengaged or fatigued participants can go as far as providing false information in order to avoid answering a set of optional questions (Survey Sampling International [SSI], 2010). The survey included a section on holidays and recreation, which could be skipped if respondents had not been on a holiday during the past year. For some respondents, this block of questions was presented at the beginning of a 30-minute study, and for others – towards the end. Although the positioning of the question should have no effect on respondents’ holiday history, the percentage of affirmative answers was significantly lower (47% vs. 64%) when the question was encountered later in the study.

### Factors underlying problematic behavior

Problematic behavior is associated with the state of the respondent (e.g., fatigue, boredom, low engagement) and properties of the assessment (e.g., questionnaire length and complexity). Evidence comes mainly from laboratory experiments and secondary data analyses of survey panel reports.

Study length has negative impact on data quality. Studies published in peer-reviewed journals have shown that towards the end of a lengthy questionnaire (after about 20 minutes) participants begin to spend less time answering a set of questions (speeding), give shorter answers to open-ended questions, answer in uniform patterns (e.g., straightlining), and opt for nonresponses such as “I don’t know” and “not applicable” (Deutskens, De Ruyter, Wetzels,

& Oosterveld, 2004; Galesic & Bosnjak, 2009; Herzog & Bachman, 1981). Similar effects of survey length were found in several reports using secondary survey data (SSI, 2010; Sleep & Puleston, 2008; MarketTools, 2010).

Design and layout variables can also influence respondent behavior. When too many items (e.g., 20-40 as opposed to 1-10) are presented on the same page, respondents are more likely to skip items (Toepoel, Das, & Van Soest, 2009) or entirely abandon the study (Sleep & Puleston, 2008; LightspeedAhead, 2010). In a market research study, including pictures of products along with product descriptions was shown to reduce the amount of nonresponse (Deutskens et al., 2004). Similarly, exemplifying question content and response scales (e.g. transforming response categories into a slider bar) led to better questionnaire evaluations (Downes-Le Guin, Baker, Mechling, & Ruylea, 2012). Excessively visual and cluttered layout, however, increased dropout rates from around 6 percent to 42 percent (Downes-Le Guin et al., 2012). Violating basic design aesthetic principles of color and shape (e.g. presenting question text in red boxes over violet background) also led to problematic respondent behavior, such as skipping questions, negative emotional tone of stated opinions, and decreased amount of time spent per question (Mahon-Haft & Dillman, 2010).

Question format is another important factor. Survey questions can be presented in different formats – typically as single items (e.g. Appendix 1b) or in a grid (Appendix 1a). In the separate-items format a single question or item is followed by choice options in horizontal or in vertical alignment. In the grid format, several questions or items are typically listed in the left-most column of a table, and each subsequent column represents a choice option. Several studies have shown that question grids are linked to problematic behavior (Galesic & Bosnjak, 2009), lower data quality (SSI, 2010), and lower respondent satisfaction (Grandmont, Goetzinger, Graff, & Dorbecker, 2010). According to data from over 25,000 survey respondents reported in a Lightspeed Research newsletter from 2010, 15 % of study participants reported dropout due to “dislike for and reluctance to complete grids.”

**Solutions to problematic respondent behavior**

The issue of problematic respondent behavior is typically dealt with by translating research findings into prescriptions against bad practices (e.g., shortening survey length, avoiding grid questions; ) and, as a final step, identifying and screening out flawed responses (e.g., Downes-Le Guin, 2006; Oppenheimer et al., 2009).

Changes to the design of studies in order to influence response quality can be content-oriented (e.g., phrasing of questions, choosing appropriate study topics, shortening of existing scales), as well as independent of content (e.g., visual design, question type, items per screen). Content changes can compromise study quality (e.g. scale reliability) and are thus not always possible or desirable. Layout changes are less problematic, easier to implement, and cost-effective. Research has shown that attempts to create engaging questionnaires can improve respondent experience and data quality (VisionCritical, 2008; Drolet, Butler, & Davis, 2009). Nevertheless, the prescriptions based on prior research findings are not always put into practice (Sleep & Puleston, 2008). For example, a brief visual analysis of online psychology studies active in March 2013 revealed that 9 out of 14 designs included grid questions (see Table 1).

Data validation and cleaning is a necessary step in psychometric research, especially for studies conducted online or in other non-laboratory settings (Downes-Le Guin, 2006; Oppenheimer et al., 2009). Despite the clear benefits of these procedures, removing cases can have important disadvantages, such as loss of statistical power due to sample size and sampling bias in case attrition is not entirely random. It is therefore desirable to minimize the amount of responses that need to be screened-out (Downes-Le Guin, 2006).

Several authors conclude that effort should be made to improve response quality by enhancing respondent experience and engagement, especially through cost-effective methods (Couper, 2000; Sleep & Puleston, 2008). This thesis offers an analysis of previously unexplored visual design elements and their effect on online questionnaires.

## Psychological research on processing fluency

Psychology provides a solid theoretical background for understanding the impact of incidental variables on cognitive and motivational processes. Under conditions of complexity or uncertainty, people might automatically incorporate affective, cognitive, metacognitive, and even bodily experiences into their judgments (Tversky & Kahnemann, 1974; Schwarz & Clore, 1983; Strack, Martin, & Stepper, 1988). In an experiment by Schwarz and colleagues (1991), when asked to list only a few (as opposed to many) examples of being assertive, people rated themselves as more assertive – an effect likely due to the fact that recalling few examples is easier than recalling many and people interpreted this metacognitive experience of ease as evidence for their assertiveness (Schwarz, 1991). Further research has shown that experiences of ease and difficulty are not limited to information recall, but can also arise from the processing of information itself (processing fluency). Processing fluency can be manipulated in various ways, including exposure, priming, and visual clarity (for a review, see Alter & Oppenheimer, 2009). High fluency has been shown to create positive affect (Winkielman & Cacioppo, 2001) and influence a wide range of judgments, such as enjoyment (Reber, Schwarz, & Winkielman, 2004), competence (Oppenheimer, 2006), and effort (Song & Schwarz, 2008).

**Visual fluency**. Particularly relevant for the current research is the effect that visual design and text formatting have on processing fluency. Typeface (readable vs. non-readable), font size, color, and background contrast have all been used to manipulate visual fluency and have been shown to influence subsequent judgments. For example, when presented in an easy-to-read font, words were rated as more familiar (Reber & Zupanek, 2002), instructions as less effortful (Song & Schwarz, 2008), and choice options inspired greater confidence (Novemsky, Dhar, Schwarz, & Simonson, 2007). Similarly, statements presented in colors that ensured high contrast from background were rated as more likely to be true (Reber & Schwarz, 1999). Regarding font size, studies have shown people thought they were more likely to remember words, which were presented in a larger font size (Rhodes & Castel, 2008). Another study found that reading text presented in small font size (8pt and 10pt) lead to increased strain in participants' neck muscles and a more tense posture (Elouri, Akladios, Peres, & Amos, 2010).

**Motor fluency.** Motor actions can also be experienced as fluent or dysfluent (e.g., writing with a non-dominant hand; Petrova, 2006). It has been argued that experiences of fluency affect cognition in similar way, regardless of the origin of fluency (Alter & Oppenheimer, 2009). Facilitating respondents' motoric actions should, therefore, lead to more positive experiences during the questionnaire. In online questionnaires, responses are indicated by clicking on different locations of a response scale. The most commonly used type of scale is multiple choice scale with radio buttons, where respondents click a small circle, corresponding to a certain choice option. If motoric dysfuency is indeed experienced as unpleasant, radio buttons may not be the optimal way of indicating responses. The target area of radio buttons is relatively small and, according to Fitt's law, smaller targets require more effort to hit. Text buttons, where the choice label is surrounded by a visually designated clickable area, are an alternative to radio buttons that might ensure higher motor fluency.

**Fluency in existing online questionnaires.** Given the role of visual and motor fluency on enjoyment and motivation, questionnaires should be designed in a way that ensures optimal levels of fluency. Some good practices of visual fluency have already been established. A review of 14 randomly selected active online studies suggested that choosing a pleasant, readable typeface (e.g., Arial) is rather common. All of the 14 reviewed studies used plain sans serif font (see Table 1). Font size, on the other hand, appears to be more problematic - 75% of the reviewed studies used font size that is smaller than 12pt, which is considered a standard reading size. The practice of using small font likely originates in pen-and-paper surveys, where reducing the amount of printed materials could substantially reduce study costs. Authors have argued that using small font in computer-based studies is an outdated practice and have recommended using font size larger than 10pt (Fanning, 2005; Brace, 2008). However, the effect of font size on respondent experience, to my knowledge, has not been explicitly addressed in research.

**Processing dysfluency and cognitive performance.** While a substantial body of research has focused on demonstrating the positive effects of processing fluency, for the past several years, investigations by Alter and colleagues have revealed a different aspect of fluency. Namely, that the experience of fluency can lead to diminished cognitive effort and increased use of heuristics, whereas dysfluency - to more deep, systematic processing (e.g., Alter, Oppenheimer, Epley, & Eyre, 2007). For example, when people are asked “If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?” (Cognitive Reflection Test; Frederick, 2005), they are less likely to give the intuitive and incorrect response “100,” when the question is presented in a difficult to read font (Alter et al., 2007). The experience of dysfluency signals the presence of difficulty – a need to invest effort, rather than rely on intuition. At a first glance, their compelling findings might seem to contradict the hypotheses regarding fluency and respondent engagement outlined thus far, but a more careful consideration reveals that this is not necessarily the case. Dysfluency research typically looks at situations where a first, intuitive response is incorrect and should be overwritten by a more controlled judgment (e.g., the Cognitive Reflection Test). In psychometric assessment, however, people are specifically asked to indicate their intuitive response. Although responding does require careful processing, experiences of difficulty are not likely to facilitate it, given that respondents are prompted to think about something they know well. Furthermore, dysfluent presentation might exacerbate respondent fatigue and boredom, because of effort and decreased enjoyment, and thus diminish data quality.

### Visual attention and focus on important content

Respondent behavior can be affected by basic attentional processes. The cognitive process of answering a question involves comprehending the meaning of the question and retrieving relevant information (Tourangeau et al., 1988), during which the respondent should remain focused on the question itself. Distractions can interrupt the process and interfere with the respondent's ability to provide a meaningful answer. Internet offers a highly information rich environment where distractions are constantly present and where people are more likely to "scan" and "meddle around" rather than remain concentrated on one thing and follow instructions (Krug, 2006). One of the major concerns of web-designers, for example, is to make the most out of internet users' browsing behavior and they widely acknowledge the benefit of making website content intuitive and self-explanatory.

In an influential book on design, Steven Krug (2006) describes some fundamental principles of making content self-explanatory, which are likely to influence questionnaire respondents, much like they influence website users. One of the major principles he discusses is creating visual hierarchies by emphasizing important content. Important content, according to Krug, should be presented in larger size, bolded style, distinctive color, or a combination of these means. Visual emphasis can help direct attention towards important content, as well as prevent diverting attention away from it. In the case of questionnaires, this would be the question or item that people are currently responding to.

**Emphasis in existing online studies.** The two most common ways of presenting rating scales in questionnaires are the separate-items format (single item followed by response options; Appendix 1b) and the gird format (table with several items listed in the left-most column, and response options listed in columns; Appendix 1a). Authors have argued that the  grid format predisposes participants toward satisficing (e.g., Galesic & Bosnjak, 2009), an effect that many others have observed (SSI, 2010; Sleep & Puleston, 2008). The exact reasons for this negative effect have not been discussed in the literature, but one possibility is that the format violates basic principles of presenting important content. According to Krug's (2008) principles, if the question content is not centered or emphasized and is more likely to escape the respondent's attention, which, in turn, makes it more likely for the respondent to skip the necessary step of comprehending and carefully considering the question (Krosnick, 1991).

Whereas the separate-items format appears less problematic, analyzing the visual design of existing questionnaires revealed that there is room for improvement. Only in 4 out of 14 online studies question content was emphasized through means of design (e.g., font size and style, background color). In the remaining 11 questionnaires, question content was not visually distinct from the rest of the text (instructions, response options) and thus not spontaneously attracting attention. Adding subtle emphasis, such as background color, will highlight the question and, according to Krug (2008), help fixate the people's attention.

Another issue has to do with the spacing of items. According to the principles of Gestalt psychology, objects that are closer together are perceived as related and intuitively grouped together. It is therefore important to ensure proper spacing between separate questions, in order to make each question stand out and to emphasize its content. Although well known to psychologists and designers alike, these principles are not always utilized in study design – only 3 of the reviewed online studies used spacing to visually separate different questions.

## Present investigation

The study proposed here aims at systematically investigating the effects three factors - visual emphasis on important content, button type, and text size - on respondent engagement and behavior. Visual emphasis refers to the ways in which important content is highlighted and related elements (e.g., a question and its corresponding scale) are grouped together in order to guide the respondents' attention and facilitate the cognitive processes involved in questionnaire completion (Tourangeau & Rasinski, 1988). Responses in computer based-questionnaires are indicated through pressing buttons that correspond to different choice options and can be of different type (e.g. radio buttons, where respondents click a single, small circle or text buttons, where respondents click a designated area surrounding the option text). Button type can influence the ease with which respondents indicate their choices and hence their experiences during the questionnaire, due to perceptions of motoric fluency. Lastly, questionnaires can be presented in various font sizes. Small font can make the text difficult to read and in turn lead to negative affect and unfavorable evaluations (for a review see Alter & Oppenheimer, 2009).

Study 1a will be designed to test the effects of questionnaire visual emphasis and button type on respondent experience and data quality. A questionnaire consisting of several commonly used psychometric scales will be presented in different formatting styles. Visual emphasis and button type will be manipulated in a 2 (emphasis: high, low) by 2 (buttons-type: radio buttons, text buttons) design. Self-reported measures of participant enjoyment and engagement will be combined with objective measures of data quality, including scale reliability, correlations between opposing statements, instructional manipulation checks, and instances of inattentive responding (severe satisficing). Since careful consideration of question content is an essential step in providing meaningful responses (Tourangeau & Rasinski, 1988), it is hypothesized that questionnaire design, which does not direct respondents' visual attention towards the question content, will result in lower data quality. Regarding button type, it is hypothesized that radio buttons, which are small and harder to press will diminish respondent enjoyment and thus have a negative impact on data quality. More specifically, high visual emphasis (as compared to low emphasis) and text buttons (as compared to radio buttons), should lead to 1) lower dropout rates and lower number of inattentive respondents; 2) greater satisfaction with the questionnaire and self-reported engagement; 3) better data quality, reflected in (3a) higher scale reliability, (3b) negative correlations between opposing statements, (3c) higher correlations between related scales.

In Study 1b will be use the same questionnaire content and dependent measures as Study 1a. The questionnaire will be presented in two different font sizes: regular (13pt) and small (8pt). It is hypothesized that using regular (as opposed to small) font size will lead to 4) lower dropout rates and lower number of inattentive respondents; 5) greater satisfaction with the questionnaire and self-reported engagement; 6) better data quality, reflected in (6a) higher scale reliability, (6b) negative correlations between opposing statements, (6c) higher correlations between related scales.

## Methods

### Participants

Participants will be recruited from the Amazon Mturk participant pool, through the CrowdFlower platform (crowdflower.com) and will complete the questionnaire online. A total of 300 U.S. citizens (60 per condition) will participate.

## Materials and Procedure

Six versions of the same questionnaire, containing five commonly used psychometric scales will be presented in different visual formatting. In Study 1a, visual emphasis on question content and button type will be systematically varied and fully crossed (Condition 1: high emphasis, text buttons; Condition 2: high emphasis, radio buttons; Condition 3: low emphasis, text buttons; Condition 4: low emphasis, radio buttons). Each participant will be randomly assigned to completing one of the four visual formatting versions. In Study 1b, font size will be manipulated. Condition 4 from Study 1a will be compared to an additional condition (Condition 5), which is visually identical to Condition 4 but presented in smaller font size (8pt vs. 13pth).

The questionnaire will include a total of 114 questions and is expected to take approximately 30 minutes to complete. Prior research has indicated that fatigue effects become increasingly pronounced after about 20 minutes. The proposed questionnaire duration will, therefore allow for examining how the manipulations influence respondent behavior at different levels of fatigue. All versions of the questionnaire will use separate-items question format (10 questions per page) with horizontal multiple choice response scales and web-safe sans serif font (Arial).

**Visual emphasis manipulation.** Visual emphasis will be defined through a combination of three design elements: background color, difference in font size between question and response scale, and additional spacing to visually separate the questions. In the high-emphasis conditions, questions will have a unique background color, distinguishing them from background and response scales. Question text will be 13pt and the text of response scales will be 8pt. There will be additional spacing before each question, resulting in larger space between two separate questions than between a question and its corresponding scale. In the low-emphasis conditions, all background color will be the same (white), all text size will be set to 13pt, and there will be equal spacing between all elements. See Appendix 1 for an illustration of high emphasis.

**Button type.** In the radio-button conditions, participants will indicate their responses by clicking on a radio button, which is approximately 20x20 pixels in size and is located above the corresponding choice label. In the text-button condition, the choice text will be surrounded by a subtly colored clickable area (approximately 120x40 pixels) to form a button.

**Font size manipulation.** Study 1b will include an additional condition, identical to Condition 4 of Study 1a (low emphasis, radio buttons) with the exception of font size. All text in Condition 5 will be set to 8pt.

**Questionnaire length and content.** The following scales will be included: (a) Emotionality, conscientiousness, and agreeableness subscales from HEXACO-60 (30 items; Ashton & Lee, 2009); (b) The Patient Health Questionnaire - PHQ9 (10 items; Kroenke, Spitzer, & Williams, 2001); (c) Self-report Depression Scale – SDS (20 items; Zung 1956); (d) Need for cognition – NFC (18 items; Cacioppo, Petty, & Kao, 1984); (e) Positive and Negative Affect Scale – PANAS (20 items; Watson, Clark, & Tellegen, 1988). The presentation order of the two depression scales (PHQ9 and SDS) will be randomized. In half of the participants SDS would appear at the beginning of the questionnaire PHQ-9 – towards the end. For the other half of the participants, this order will be reversed. There will be an instructional manipulation check (Oppenheimer et al., 2009), assessment of respondents' experience and motivation (12 items), and basic demographic variables. Participants will be asked to estimate how long they think it took them to complete the questionnaire. The actual questionnaire duration will be recorded along with the duration of a two single pages (one at the beginning and one towards the end of the study).

**Dependent variables.** The dependent variables will include self-reported experience and motivation, an instructional manipulation check, indexes of data quality, and dropout rates. Respondent experience will be assessed with 12 items, partially adopted from the Intrinsic Motivation Inventory (IMI; Ryan, 1982). The questions will cover perceptions of invested effort (e.g. "I thought deeply about how to answer the questions"), enjoyment (e.g., "This questionnaire was more enjoyable than most"), and visual appeal (e.g., "I liked the design of this questionnaire").

Respondents who abandoned the study prematurely will be classified "dropouts." Respondents who did not dropout but failed to complete the instructional manipulation check or showed instances of "straightlining" (more than 10 consecutive answers appearing a uniform line)  will be classified as "inattentive." Dropout and inattentiveness rates will be used as dependent variables.

Data quality will be inferred from (a) scale reliability, (b) correlations between opposing statements, such as reversed and non-reversed items in a single scale, (c) instances of straightlining, (d) completion of instructional manipulation checks. Furthermore, the correlation between two scales measuring the same concept (PHQ9 and SDS) will be compared across condition.

The questionnaire duration (time it took respondents to complete the entire questionnaire), along with respondents' estimation of how long the questionnaire lasted (in minutes; open-ended question) will also be recorded as dependent variables.

## Planned Analyses and Expected Results

### Study 1a

The different variants of the questionnaireformats (1HETB=high emphasis, text buttons; 2HERB=high emphasis, radio buttons; 3LETB =low emphasis, text buttons; 4LERB=low emphasis, radio buttons) will be combined to reflect the two main factors, emphasis and button type. The conditions will be: emphasis (high, 1HETB + 2HERB; low, 3LETB + 4LERB), button-type (text, 1HETB + 3LETB; radio, 2HERB + 4LERB).

**Dropout rates and instructional manipulation checks.** As a first step, I will identify respondents who did not complete the questionnaire (dropouts) or completed it inattentively, that is, without reading the questions, and thus failed to complete an instructional manipulation check (Oppenheimer, 2008) or gave a 10 or more consecutive uniform responses (straightlinging). Two Chi-square tests will be conducted to test whether instances of dropout and inattentiveness differ significantly between the two experimental conditions. In line with the general hypotheses of the study, dropout and inattentiveness should be higher in the low-emphasis (as compared to the high-emphasis) condition and in the radio-buttons (compared to the text-buttons) condition. These predictions are not expected to hold if there are too few cases of dropout and inattentiveness in the sample.

**Respondent experience and questionnaire evaluation.** The psychometric properties of the Respondent Experience scale will be assessed. The questionnaire consists of 11 items (including five reverse items), designed to capture three different aspects of respondents' experience: (a) their level of focus and concentration; (b) enjoyment of the questionnaire; and (c) perceived visual pleasantness of the questionnaire. An exploratory factor analysis will be conducted. The results of this analysis (Scree plot interpretation and factor loadings) will determine whether the scale will be treated as a unified measure or broken down into its separate subscales.

The effect of condition on respondent experience and questionnaire evaluation will be tested in a general linear model or a multivariate analyses of variance (ANOVA), depending on whether one (unified) or separate evaluation scales are used. The two experimental conditions (emphasis and button type) and their interaction, will be used as independent variables. Two main effects of emphasis and button type are expected, such that the high-emphasis and the text-buttons conditions are, on average, associated with higher engagement, enjoyment, and ratings of visual pleasantness. Alternatively, the effect of emphasis and button type might emerge only in combination, in which case I expect an interaction, indicating that the combination of high emphasis and text buttons leads to higher engagement and better evaluations.

**Affect**. To complement the direct self-reported evaluation with a less explicit measure of enjoyment, I will compare respondents' positive and negative affect (assessed through PANAS), across experimental conditions. Higher scores on positive affect and lower scores on negative affect are expected for people in the high-emphasis (compared to the low-emphasis) and the text-buttons (compared to the radio-buttons) conditions.

**Scale reliability.** The reliability (Cronbach's alpha) of each scale included in the questionnaire, will be calculated for each experimental condition. The separate conditions will be treated as independent samples. I will then compare the alpha values using the Fisher-Bonett test (Kim & Feldt, 2008).

As an alternative index of scale reliability, I will compute the correlation between two reversed and two non-reversed items of a single scale (the correlation is expected to be negative) and compare these correlations between conditions using Fisher's Z-test.

For both indexes, reliability is expected to be higher for the high-emphasis (compared to the low-emphasis) and the text-buttons (compared to the radio-buttons) conditions.

**Presentation order effects.** Two of the scales – SDS and PHQ-9 – were randomly presented either at the beginning or towards the end of the questionnaire. To estimate general fatigue effects, I will compute Cronbach's alphas and correlations between reversed items for each presentation order (beginning vs. end). Then I will compare the values of these statistical tests, using the procedures described in the previous section. I expect reliability to be higher for scales presented at the beginning, rather than at the end of the questionnaire, due to fatigue effects.

**Correlations between related scales.** The questionnaire includes two depression scales, one presented at the beginning and one at the end of the questionnaire. The presentation order (which scale appears first and which second) was randomized and counterbalanced. I will compute the correlation between these two scales for each condition, treating the conditions as separate samples. I will then compare the correlations using Fisher's Z-test. Correlations should be higher in the high-emphasis (compared to the low-emphasis) and the text-buttons (compared to the radio-buttons) conditions.

**Timing and time perception.** Research has shown that as respondents become fatigued (or bored), they begin to spend less time on answering questions. At the same time, perceiving a task as shorter than it actually is, shows engagement, whereas the feeling that time drags is related to boredom.

The questionnaire duration (time it took respondents to complete the entire questionnaire) was recorded, along with respondents' estimation of how long the questionnaire lasted (in minutes). Additionally, the time it took to complete a single page (10 questions) was recorded for a page that was presented either at the beginning or near the end of the study. Total questionnaire duration, single-page duration, and total-duration estimate will be compared across conditions (multivariate ANOVA).

No directional hypotheses are made regarding the actual questionnaire duration. On one hand, shorter times can be interpreted as lack of engagement. On the other, they can reflect differences in perceptual and motoric fluency of the experimental conditions, which make the questions easier (and faster) to complete.

The effect of presentation order (beginning vs. end) on page duration will be estimated by comparing the duration of the page when it was presented at the beginning to its duration when presented towards the end (t-test). Shorter duration when the page is presented towards the end will be interpreted as evidence for fatigue effects and diminished engagement.

To see whether emphasis and button type affected completion time, and whether their effects depend on presentation order, I will conduct a general linear model, in which timing will be regressed on emphasis, button type, presentation, all two-way and a three-way interaction. A main effect of presentation order, with shorter duration when page is presented towards the end, will be interpreted as evidence for fatigue effects and diminished engagement. Similar to the reasoning regarding total questionnaire length, no directional hypotheses are made regarding the directions of possible main effects of emphasis and button type. In line with previous predictions, significant two-way interactions of emphasis and button type with presentation order, are expected to indicate reduced fatigue effects in the high-emphasis and text-buttons conditions (i.e., the effect of presentation order will be diminished in these conditions). A significant three-way interaction might emerge if the high emphasis and text buttons have an effect only in combination.

Time perception refers to participant’s estimate of how long it took them to complete the questionnaire (in minutes), assessed by a single, open-ended question. Severe outliers (Z-score>3) and nonsensical responses (e.g., zero or 100) will be excluded from the analysis. Duration estimates will be regressed on emphasis, button-type, and their interaction. I expect significant main effects of the two conditions, such that high emphasis and text buttons are associated with lower time estimates.

**Exploratory analyses.** If there are no significant effects of condition, I will further explore the data by running two types of contrasts to compare the condition with high emphasis and text buttons to: (a) the rest of the conditions and (b) to the low-emphasis radio-buttons condition.

## Study 1b

Study 1b addresses the potential influence of text size on questionnaire data quality and respondent engagement. Two experimental conditions will be compared: small (8pt) vs. regular (13pt) font size.

The analyses will follow the logic outlined in Study 1a and the general predictions are that regular (as compared to small) font size will be associated with less dropout and inattentiveness (given that there are enough cases), higher scores on the Respondent Experience scale (engagement, enjoyment, visual pleasantness), more positive and less negative affect (PANAS), higher scale reliabilities, higher negative correlations between reverse items, higher correlations between the two depression scales (SDS and PHQ9), less pronounced fatigue effects (i.e., differences in duration when a page is presented at the beginning vs. towards the end of a study), and higher completion times but lower completion time estimates (self-reported).

Dropout and inattentiveness will be compared across conditions using two Chi-square tests. Respondent engagement, affect, total questionnaire duration, and duration estimates, will be compared using a series of t-tests. Scale reliabilities (alphas) and correlations will be compared using Fisher-Bonett tests and Fischer's Z-tests, respectively.

## References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review*, *13*, 219–35. doi:10.1177/1088868309341564

Alter, A., Oppenheimer, D., Epley, N., & Eyre, R. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569-576. doi: 10.1037/0096-3445.136.4.569

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*, 340-345.

Brockmole, J. R., Carlson, L. A., & Irwin, D. E. (2002). Inhibition of attended processing during saccadic eye movements. *Perception & Psychophysics, 64*, 867-881.

Brace, I. (2008). Questionnaire design: How to plan, structure, and write survey material for effective market research (2nd ed.). India: Replica Press.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment, 21*, 115–127.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement*, *60*, 821–836. doi:10.1177/00131640021970934

Couper, M. (2000). Web surveys: a review of issues and approaches. *Public opinion quarterly*, *64*, 464–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11171027

Deutskens, E., De Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*, 21–36. doi:10.1023/B:MARK.0000021968.86465.00

Downes-Le Guin, T. (2006). Great results from ambiguous sources: Cleaning internet panel data. *Panel Research*.

Downes-Le Guin, T., Baker, R., Mechling, J., & Ruylea, E. (2012). Myths and realities of respondent engagement in online surveys *International Journal of Market Research, 54*, 1-21. doi:10.2501/IJMR-54-5-000-000

Drolet, J., Butler, A., & Davis, S. (2009). The Survey Burden Factor: How important is the respondent's perception of survey length? *Quirks, November, 2009*, 44-51.

Elouri, Y., Akladios, M., Peres, S. C., & Amos, a. (2010). Effects of Font Size on Muscle Activity. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*, 1105–1109. doi:10.1177/154193121005401503

Fanning, E. (2005). Formatting a Paper-based Survey Questionnaire: Best Practices. *Practical Assessment Research & Evaluation, 10*, 1-14.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42. doi: 10.1257/089533005775196732

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349–360. doi:10.1093/poq/nfp031

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93-104.

Grandmont, J., Graff, B., Goetzinger, L., & Dorbecker, K. (2010). Grappling with grids: How does question format affect data quality and respondent engagement? Paper presented at the annual meeting of the American Association for Public Opinion Research, Chicago, IL, May. Retrieved February 2013 from http://www.amstat.org/

Heerwegh, D., & Loosveldt, G. (2008). Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality. *Public Opinion Quarterly*, *72*(5), 836–846. doi:10.1093/poq/nfn045

Herzog, A. R., & Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly*, *45*, 549–559.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*, 537–67. doi:10.1146/annurev.psych.50.1.537

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, M., Kopp, R. J., Mitchell, R. C., et al. (2002). The Impact of “No Opinion” Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice? *Public Opinion Quarterly*, *66*, 371–403.

Krug, S. (2006). *Don't make me think! A common sense approach to web usability* (2nd ed.). USA: New Riders

LightspeedAhead. (2010). *Respondent Engagement: How Much Does it Matter?* [Newsletter]. Retrieved February 2013 from http://www.lightspeedaheadnewsletter.com/?p=219

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods*, *4*, 43–59.

MarketTools. (2010). *When good respondents go bad: How unengaging surveys lower data quality* [White Paper]. Retrieved February 2013 from http://www.markettools.com/KnowledgeCenter

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference Fluency in Choice, *Journal of Marketing Research, 44*, 347–356.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, *20*, 139–156. doi:10.1002/acp.1178

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872. doi:10.1016/j.jesp.2009.03.009

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450-461.

Reber, R, & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, *8*, 338–42. doi:10.1006/ccog.1999.0386

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and social psychology review*, *8*, 364–82. doi:10.1207/s15327957pspr0804_3

Reber, R., & Zupanek, N. (2002). Effects of processing fluency on estimates of probability and frequency. In P. Sedlmeier & T. Betsch (Eds.), Frequency processing and cognition (pp. 175-188). Oxford, UK: Oxford University Press.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of experimental psychology. General*, *137*, 615–25. doi:10.1037/a0013684

Sanchez, E. (2013). Effects of questionnaire design on the quality of survey data. A*merican Association for Public Opinion Research, 56*, 206–217.

Schwarz, N., & Clore, G. L. (1983). Mood, misatribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513–523.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatke, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. Journal of *Personality and Social Psychology, 61*, 195–202.

Sleep, D., & Puleston, J. (2008). The survey killer. *Quirks*, (November), 54–58.

Song, H., & Schwarz, N. (2008). If it’s hard to read, it's hard to do: processing fluency affects effort prediction and motivation. *Psychological science*, *19*, 986–8. doi:10.1111/j.1467-9280.2008.02189.x

Strack, F., Martin, L. & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768-777.

Survey Sampling International. (2010). *Questionnaire length, fatigue effects and response quality revisited* [White Paper]. Retrieved February 2013 from http://www.research-voice.com/en/KnowledgeLink/learn-more.aspx

Thomas, L. E., Lleras, A. (2007). Moving eyes and moving thought: on the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review, 14*, 663-668.

Toepoel, V., Das, M., & Van Soest, A. (2009). Design of Web Questionnaires: The Effects of the Number of Items per Screen. *Field Methods*, *21*, 200–213. doi:10.1177/1525822X08330261

Tourangeau, R., Rasinski, K. A., Abelson, R., Bradburn, N., Campbell, S., & An-, R. D. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement, *103*, 299–314.

VisionCritical. (2008). *Maximizing respondent engagement through survey design* [White Paper]. Retrieved February 2013 from http://vcu.visioncritical.com/

Watson, D.,Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

Winkielman, P., & Cacioppo, J. T. (2001). Psychophysiology processing fluency and positive affect. *Journal of Personality and Social Psychology*.

**Table 1.**

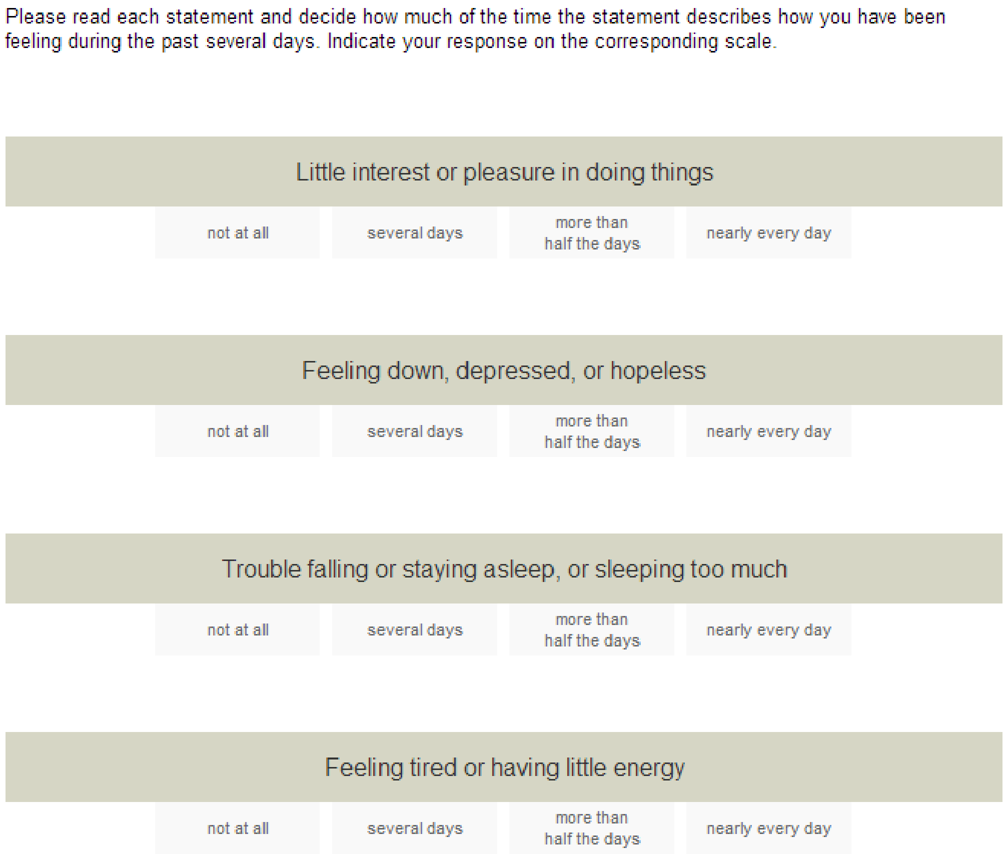
*Summary of visual design of reviewed online questionnaires*

| Institution | | Duration | Scale type | Typeface | pt | Visual emphasis | Question spacing |
|---|---|---|---|---|---|---|---|
| 1 | Harvard | 10 | Separate items | Verdana | 12 | Italicized font | Not applicable |
| 2 | University of Southampton | 10 | Separate items | Arial | 11 | None | Visually separated questions |
| 3 | Purdue University | 10 | Separate items | Arial | 10 | Background color | No visual separation |
| 4 | Columbia University | 25 | grid | Arial | 10 | None | No visual separation |
| 5 | Claremont University | 15 | grid | Verdana | 13 | None | No visual separation |
| 6 | Northwestern University | 20 | grid | Arial | 9 | None | No visual separation |
| 7 | Manhattan College | 40 | Separate items | Arial | 14 | Bold font | Visually separated questions |
| 8 | Mary Washington | 15 | grid | Arial | 9 | None | No visual separation |
| 9 | The Wright Institute | 30 | grid | Verdana | 10 | None | No visual separation |
| 10 | University College London | 15 | grid | Arial | 10 | None | No visual separation |
| 11 | Palo Alto University | 40 | grid | Arial | 12 | Bold font | No visual separation |
| 12 | Regent University | 15 | grid | Arial | 10 | None | No visual separation |
| 13 | Pepperdine University | 60 | Separate items | Arial | 11 | None | Visually separated questions |
| 14 | Curtin University | 30 | grid | Arial | 10 | None | No visual separation |

# Appendix 1

## Screenshots from visual emphasis experimental conditions

Condition 1. *High emphasis, text buttons*



Please read each statement and decide how much of the time the statement describes how you have been feeling during the past several days. Indicate your response on the corresponding scale.

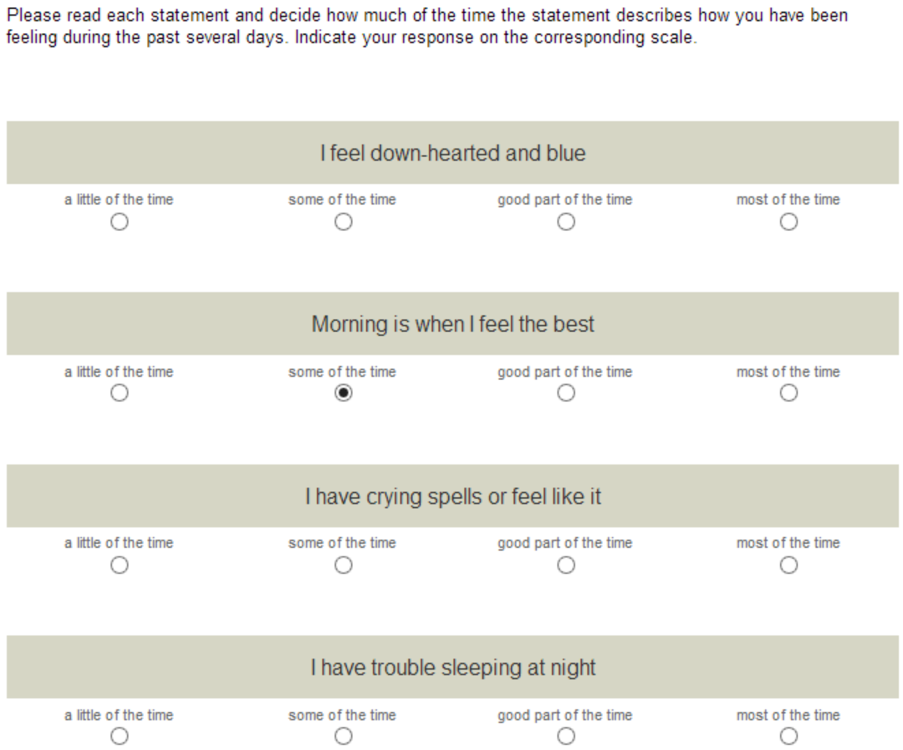
Little interest or pleasure in doing things

Feeling down, depressed, or hopeless

Trouble falling or staying asleep, or sleeping too much

Feeling tired or having little energy

Condition 2. *High emphasis, radio buttons*



Please read each statement and decide how much of the time the statement describes how you have been feeling during the past several days. Indicate your response on the corresponding scale.

I feel down-hearted and blue

| a little of the time | some of the time | good part of the time | most of the time |
|---|---|---|---|
| ○ | ○ | ○ | ○ |

Morning is when I feel the best

| a little of the time | some of the time | good part of the time | most of the time |
|---|---|---|---|
| ○ | ● | ○ | ○ |

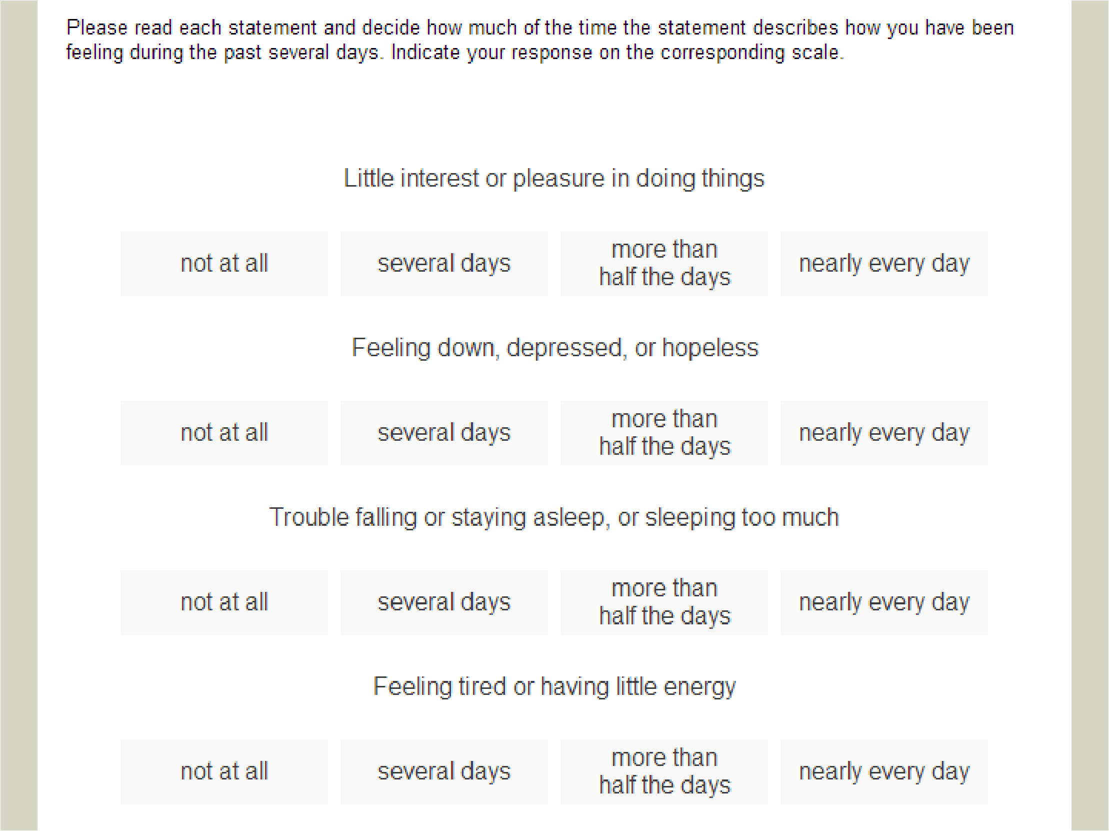
I have crying spells or feel like it

| a little of the time | some of the time | good part of the time | most of the time |
|---|---|---|---|
| ○ | ○ | ○ | ○ |

I have trouble sleeping at night

| a little of the time | some of the time | good part of the time | most of the time |
|---|---|---|---|
| ○ | ○ | ○ | ○ |

Condition 3. *Low emphasis, text buttons*



Please read each statement and decide how much of the time the statement describes how you have been feeling during the past several days. Indicate your response on the corresponding scale.

Little interest or pleasure in doing things

| | | | |
|---|---|---|---|
| not at all | several days | more than half the days | nearly every day |

Feeling down, depressed, or hopeless

| | | | |
|---|---|---|---|
| not at all | several days | more than half the days | nearly every day |

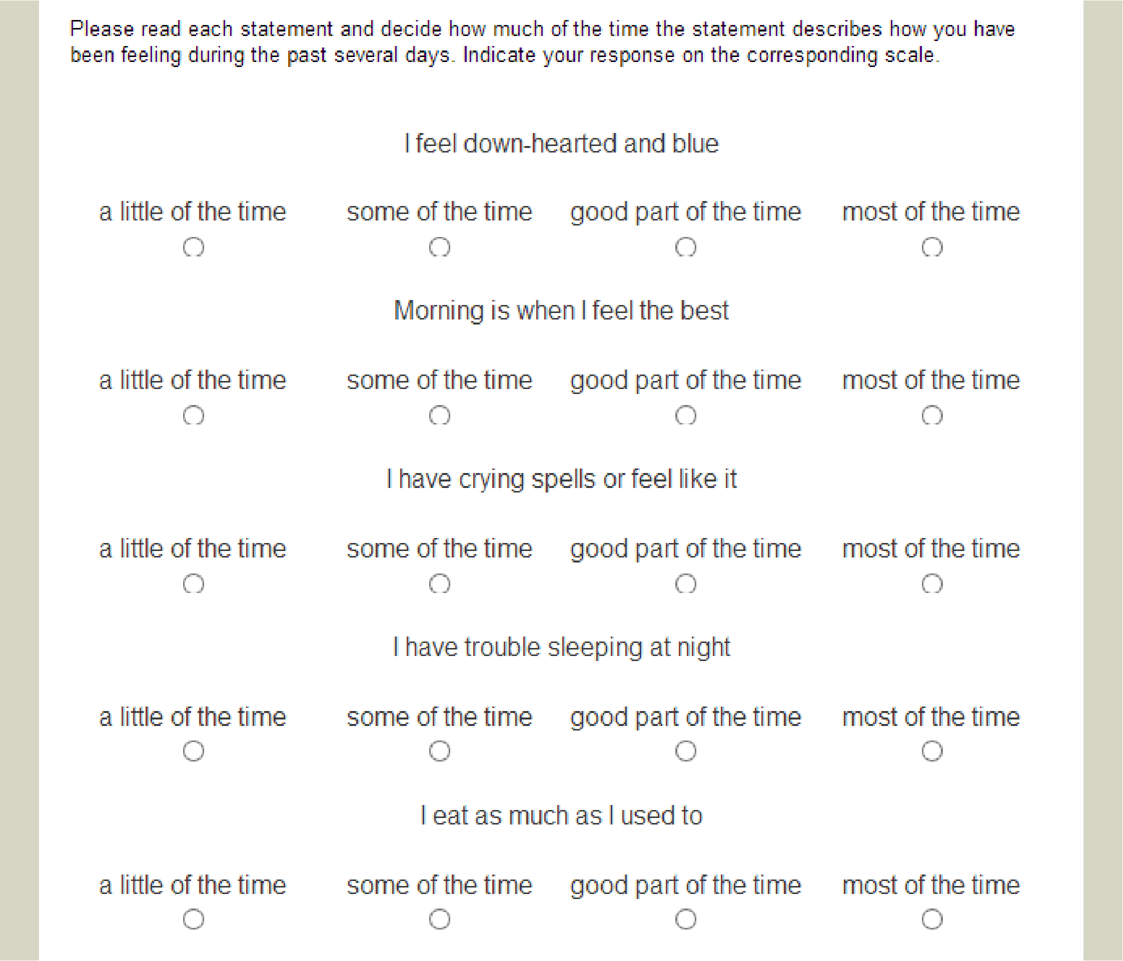
Trouble falling or staying asleep, or sleeping too much

| | | | |
|---|---|---|---|
| not at all | several days | more than half the days | nearly every day |

Feeling tired or having little energy

| | | | |
|---|---|---|---|
| not at all | several days | more than half the days | nearly every day |

Condtition 4. *Low emphasis, radio buttons*



Please read each statement and decide how much of the time the statement describes how you have been feeling during the past several days. Indicate your response on the corresponding scale.

I feel down-hearted and blue

| | | | |
|---|---|---|---|
| a little of the time | some of the time | good part of the time | most of the time |
| ○ | ○ | ○ | ○ |

Morning is when I feel the best

| | | | |
|---|---|---|---|
| a little of the time | some of the time | good part of the time | most of the time |
| ○ | ○ | ○ | ○ |

I have crying spells or feel like it

| | | | |
|---|---|---|---|
| a little of the time | some of the time | good part of the time | most of the time |
| ○ | ○ | ○ | ○ |

I have trouble sleeping at night

| | | | |
|---|---|---|---|
| a little of the time | some of the time | good part of the time | most of the time |
| ○ | ○ | ○ | ○ |

I eat as much as I used to

| | | | |
|---|---|---|---|
| a little of the time | some of the time | good part of the time | most of the time |
| ○ | ○ | ○ | ○ |

## Appendix 2

## Dependent variables

| | |
|---|---|
| Dropout and inattentiveness | |
| Dropout | |
| Instructional manipulation check | |
| Strainglining (more than 5 consecutive responses are identical) | |
| Data quality | |
| Scale reliability (alpha) | |
| Correlations between reversed items | |
| Correlations between SDS and PHQ | |
| Respondent experience and motivation | |
| Effort / Focus | I found it easy to concentrate on the questions<br>I thought deeply about how to answer the questions<br>I felt distracted while working on the questionnaire (reversed)<br>This questionnaire did not hold my attention (reversed)<br>It did not put much effort into completing the questionnaire (reversed) |
| Enjoyment | This questionnaire was more enjoyable than most<br>This questionnaire was fun to complete<br>I was so involved in the questionnaire that I lost track of time<br>This questionnaire was boring (reversed) |
| Visual appeal | I liked the design of this questionnaire.<br>The text was easy to read.<br>Visually, this questionnaire was unpleasant. (reversed) |
| General motivation | I typically enjoy completing surveys.<br>If all questionnaires look like this… (a) I will be more likely to take part in online questionnaires; (b) no change from my current |

| | |
|---|---|
| | participation; (c) I will be less likely to take part in online questionnaires; |
| Questionnaire duration | |
| Page duration | |
| Duration estimates | |
| Positive/Negative affect | |