Survey Research Methods (2010)
Vol.4, No.1, pp. 43-59
ISSN 1864-3361
http://www.surveymethods.org

Does Visual Appeal Matter? Effects of Web Survey Aesthetics on Survey Quality

Taj Alexander Mahon-Haft and Don A. Dillman
Washington State University

Due to recent scholarly attention to visual design, much is known about the effects that specific design features have on web surveys, but little attention has been paid to the effects of overall screen design. Applying Norman's (2004) emotional design concepts to web survey design, we propose and test hypotheses related to potential detrimental impacts of poor screen aesthetics. To do so, we compare data collected from two versions of the same Student Experience Survey, an experimental design representing poor aesthetics and a control design representing good emotional design. By triggering negative visceral responses and, thus, emotional reactions, the experimental design is predicted to reduce data quality along four avenues: respondent cooperation, substantive response patterns, respondent commitment, and subjective survey experience. Sixteen of 30 total comparisons show reduced data quality on the experimental forms, and all avenues to quality data show some impact. These findings suggest that screen design aesthetics and emotional reactions can impact data quality, independent of survey content. However, the inconsistent results according to some indicators suggest the need for further research on screen design effects to hone our understanding.

Keywords: aesthetics, web survey, screen design, emotional design, visceral reactions

Introduction

Since the late 1990s, increasing concerns regarding coverage and response rates have greatly reduced survey researchers' reliance on telephone surveys (Singer, 2006). Meanwhile, increasing internet accessibility, cheaper per-unit costs, and quicker response and data collection have contributed to the growth of web-based surveys (Couper and Miller, 2008). This shift, and the greater range of visual design opportunities offered by web surveys (Couper, 2008) has led survey researchers to spend considerable effort examining the impacts of visual features of web surveys (for example, Christian and Dillman, 2004; Toepoel, 2008; Tourangeau et al., 2004; 2007). The main focus of this research has been on specific structural considerations, such as the numbering and spatial layout of response options (Tourangeau, et al., 2007; 2004) and text box size (Christian and Dillman, 2004).

Only rarely, have studies examined the impacts of overall screen design and layout, leading to a paucity of knowledge of how screen aesthetics influence survey response rates and other aspects of data quality. The few previous experiments have examined the effect of various screen designs on response rates. For example, screens with bright colors, graphics and complicated display arrangements have been compared to simpler, plainer styles. In those tests, the elaborate screen designs failed to improve response rates and, in some cases, appear to have decreased them (Dillman, Tortora, Conradt and Bowker, 1998; Coates, 2004). More recent research has focused on the development of colorful and engaging ways of displaying questions in an effort to decrease survey terminations among respondents (Sleep and Pulleston, 2008). All of this research has lacked a guiding theoretical perspective for developing screen aesthetics or better understanding how style elements relate to one another and impact data quality.

The lack of previous research attention to screen aesthetics can be attributed in part to difficulties associated with defining and measuring aesthetics (Lindgaard, 2007). Still, task-unrelated aesthetic qualities have been found to impact users' experiences on web sites in a variety of ways outside of the survey realm (see, for example, Hassenzahl et al., 2001; Lindgaard and Dudek, 2003). It has been argued that these impacts are the result of the role of aesthetics in the formation of our emotional reactions to any product, which involve three levels of cognitive responses (visceral, behavioral, and reflective) that strongly impact our perceptions, mood, behavior, and cognitive functioning (Norman, 2004). Basic sensory stimuli determine the initial responses at the visceral level, which guide subsequent cognitive responses, and can influence judgment of subsequent perceptions and behavior (Norman, 2004). Thus, in visually-based web surveys, it is the aesthetic qualities (visual appeal) that determine visceral responses, either innate repulsion or attraction, that guide emotional reactions and can therefore can be to influence the rest of the survey experience and potentially impact data quality.

Previously, Norman's (2004) emotional design concepts were applied to the redesign of a government survey, when it was suggested that negative emotional reactions to an

Contact information: Don A. Dillman, Departments of Sociology, Rural Socioloy, and Social and Economic Sciences Research Center, Washington State University, e-mail: dillman@wsu.edu

aesthetically-displeasing initial design could contribute to reduced data quality (Dillman, Gertseva, and Mahon-Haft, 2005). Our purpose in this paper is to extend those initial suppositions by reporting results from a theoretically-guided test of the effects of visual appeal on web survey response behavior. To test the potential effect of aesthetic appeal on web screen design, we designed an experimental web survey with visual traits expected to elicit negative emotional reactions. Data from that screen design was compared to data from a more conventional, visually-appealing web survey design, lacking these aesthetically-displeasing attributes. Here, response rates, early terminations, item omissions, quality of open- and close-ended answers, likelihood of satisficing, time burden, and the subjective experiences of respondents are compared between the two designs. Our purpose is to understand the potential impact of screen designs for influencing respondent behavior, gain insight about possible explanations for such effects, and extend knowledge of optimal screen design traits.

Theoretical Background

Previous methodological consideration of web survey screen design, though limited, has generally concluded that certain traits are undesirable, including screen clutter (Weller, 2004), complex design (Bowker and Dillman, 2001), irregular organization and discordant colors (Brady and Phillips, 2003), and distracting (but functionally unnecessary) icons (Coates, 2004). It has been suggested that such characteristics may diminish data quality by making the survey experience less pleasant (Coates, 2004), producing lower completion rates (Bowker and Dillman, 2001), and reducing respondent effort and focus (Brady and Phillips, 2003).

More broadly, aesthetic qualities have been shown in other mediums to have a large impact on users' experiences with various interfaces. Certain fundamental aesthetic properties, such as color, layout, and simplicity have been found to reliably predict overall visual appeal ratings across users (Lindgaard et al., 2006). On web sites, ratings of visual appeal have been shown to have a strong influence on subsequent perceptions of web site enjoyment (van der Heijden, 2003; Hassenzahl et al., 2001), user satisfaction (Lindgaard and Dudek, 2003), and usability (Jennings, 2000). Similarly, visual appeal has demonstrated similar influence on users' broader perceptions of their experiences using ATM screens (Tractinsky et al., 2000) and MP3 player skins (Mahlke, 2006). Aesthetics have have such influence over user perceptions that initial aesthetic appeal has been found to guide overall perceptions of product experience and usability, even when performance actually suffered (Damasio 2000; Lindgaard and Dudek, 2003). Aesthetic appeal has even been shown to influence behavior, with task-unrelated visual traits enhancing usage of info systems (Hassenzahl et al., 2001) and improving work quality for call center agent and hotel receptionists (Draper, 1999).

In using any product, the importance of aesthetics is a result of the manner in which fundamental visual traits determine our visceral responses, which involve innate attractions and repulsions to physiologically-recognized sensory input (LeDoux, 1996; Damasio, 2000; Norman, 2004). These visceral responses are the first of three levels of cognitive responses, which also include behavioral and reflective responses, that collectively respond to the use of any product to form an emotional (or affective) reaction (Norman, 2004). He explains that this affective system is constantly passing judgment about any information encountered, literally making us feel good or bad, relaxed or tense.

These emotional reactions do more than just alter users' affective states (or moods), more broadly impacting experiences with any product in a number of ways because they occur quicker than intellectual responses (LeDoux, 1996; Goleman, 1996; Norman, 2004). This is particularly true for emotional reactions stemming from visceral responses, which occur most quickly, having been found robust at as little as 50 microseconds (Lindgaard and Dudek, 2003). As a result, these emotional reactions often result in a "confirmation bias" (Mynatt et al., 1977), in which users selectively interpret their overall experiences with a product in accordance with their emotional reactions stemming from initial subconscious impressions (Norman, 2004; Lindgaard, 2007). Norman (2004) further explains that our emotional reactions are also strongly impactful upon human behavior because the neurochemicals released by these affective reactions modify our decision making and behavior, as the emotional reactions are evolutionary mechanisms meant 'to motivate appropriate action' (Niedenthal et al., 1999). Because our cognitive response systems attach emotion to cognition (LeDoux, 1994), emotional reactions to aesthetics have even been associated with differences in cognitive functioning, with positive reactions elevating task-related performance, and vice versa (Norman, 2004). Thus, the impact of emotional reactions is so widespread that Norman concludes "[t]he emotional side of design may be more critical to a product's success than its practical elements" (Norman, 2004:26).

In web surveys, where the medium is entirely visual, users' visceral responses will come from reactions to the fundamental visual stimuli, or the aesthetic properties. Those visceral reactions will guide the overall emotional reaction, so we expect web surveys with aesthetically-displeasing design traits to produce negative emotional reactions in respondents. As shown in Figure 1, it is the production of those negative emotional reactions that we expect to act as the causal mechanism connecting aesthetically-displeasing screen design with reduced data quality. By projecting the reduced motivation, worsened mood, negative perceptions, and reduced cognitive functioning associated with negative emotional reactions to the web survey setting, we expect aesthetically-displeasing screens to detrimentally affect data quality via four avenues to potential data quality impact: response rates, response behavior, respondent commitment, and subjective survey experience (also see Figure 1). The specific indicators by which we expect these impacts to be evident are described in more detail below and are summarized in Table 1.

While this study is designed to isolate the data quality effects resulting from aesthetically-displeasing screen design

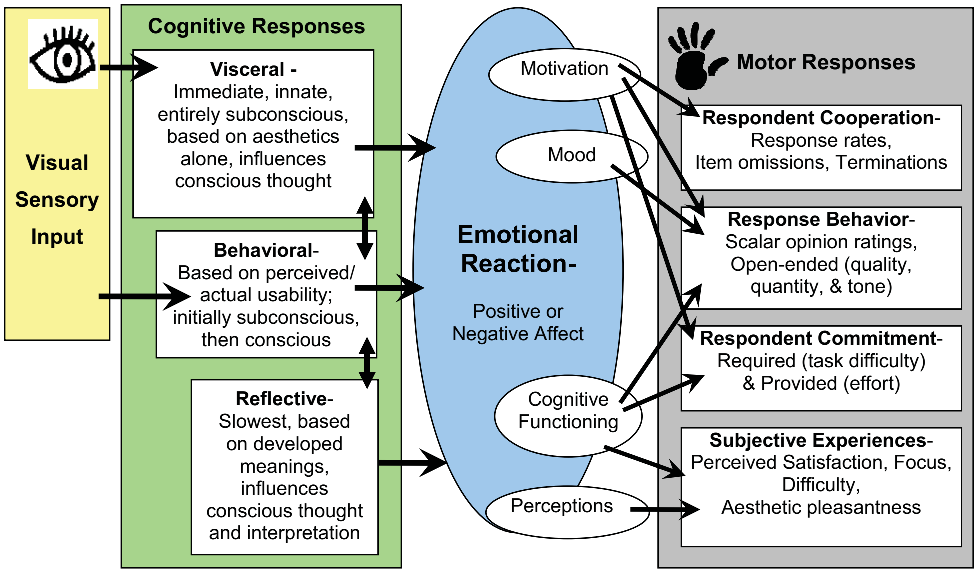

Visual Sensory Input
Cognitive Responses
Visceral - Immediate, innate, entirely subconscious, based on aesthetics alone, influences conscious thought
Behavioral- Based on perceived/ actual usability; initially subconscious, then conscious
Reflective- Slowest, based on developed meanings, influences conscious thought and interpretation
Emotional Reaction-
Positive or Negative Affect
Motivation
Mood
Cognitive Functioning
Perceptions
Motor Responses
Respondent Cooperation- Response rates, Item omissions, Terminations
Response Behavior- Scalar opinion ratings, Open-ended (quality, quantity, & tone)
Respondent Commitment- Required (task difficulty) & Provided (effort)
Subjective Experiences- Perceived Satisfaction, Focus, Difficulty, Aesthetic pleasantness

Figure 1. Explanation of link between visual input from web surveys, cognitive responses to that input, the resulting cognitive reactions, and their potential effects on human behavior during survey completion.

and the resulting negative visceral responses, it should be noted that the impacts that we project below would therefore be expected when a web survey elicits a negative emotional reaction for any reason. In fact, because negative visceral responses are subconscious and are but one level of cognitive response contributing to the emotional reaction (LeDoux, 1996; Norman, 2004), many respondents may not even consciously connect their reactions to the survey with its aesthetic properties. Instead, the production of negative emotional reactions is expected to negatively impact respondents' mood, general perceptions of the survey, motivation, and cognitive functioning, so it is the projection of those impacts to the web survey setting from which we derive our predictions. Thus, since the visceral responses affected by aesthetics precede and guide rational consideration (Goleman, 1996; LeDoux, 1996; Norman, 2004), are physiologically-based (LeDoux, 1996; Damasio, 2000), cannot be controlled (Zajonc, 1980) and have repeatedly been found to be robust beyond personal differences (Bornstein, 1992), even when respondents do not consciously associate their reactions to the form with aesthetics, those reactions can still be interpreted as being the result of their innate visceral responses. Therefore, by comparing the data from the aesthetically-displeasing experimental screen design with that from an aesthetically-pleasing control design, any observed differences in data quality can be interpreted as stemming ultimately from aesthetic differences.

Expected Effects of Viscerally-Displeasing Screens

The multitude of possible effects of viscerally-deficient screen design can be discussed as part of the four avenues to potential data quality impact summarized in Table 1. Below, we describe the theoretical expectations of each avenue and hypothesize their expected effects on a range of data quality indicators. Based on our results, future tests could then hone our understanding of particular effects.

Respondent Cooperation

The emotional response system is connected to the motivational center of the brain, and is triggered by affective stimuli (Cuthbert et al., 2000). According to social exchange theory, respondent cooperation is less likely with increased burden and reduced motivation (Dillman, 2007). Similarly, leverage saliency theory proposes that participation is dependent upon the saliency to each respondent of motivating and non-motivating characteristics, and that, for some potential respondents, the perception of increased burden is salient enough to lead to non-response (Groves et al., 2000). Studies in other fields have shown that aesthetically-pleasing visual layouts are more likely to be perceived as easy to use (Kurosu and Kashimura, 1995; Tractinsky, 2000). Thus, the aesthetically-deficient design is expected to negatively affect response rates, potentially increasing by a variety of non-response measures.

Thus, one expected impact of the experimental panel is an overall reduction in completion rates. Additionally, early terminations and item omissions have been found to be more likely among those reporting greater burden and reduced motivation (Galesic, 2006). Thus, we further predict that the experimental design will produce a higher rate of early terminations and an increased likelihood that respondents who complete the questionnaire will skip items.

Substantive Response Behavior

Since emotional reactions guide behavior, the substantive responses offered by respondents are also expected to

Table 1: Avenues of Potential Data Quality Impact of Screen Design and Associated Indicators

Avenues to Potential Impact	Indicators Compared Between Designs
Response Rates	Completion Rate Early Termination Rates Likelihood of Omitting Survey Items
Substantive Response Behavior	Likelihood of Reporting Most Favorable Scalar Opinions Likelihood of Providing Open-Ended Responses Depth of Open-Ended Responses Tone of Open-Ended Responses
Respondent Commitment	Commitment Provided= Rate of Satisficing • Mean Selections of Top/Bottom Categories Commitment Required= Time Burden • Mean Time per Response Screen (Overall, by Question Type)
Subjective Survey Experience	Likelihood of Desirable Survey Experience Ratings • Satisfaction, Difficulty, Focus, Visual Appeal

vary between screen designs. The experimental design is expected to affect measurement by altering two aspects of response behavior: (1) scalar opinion ratings, and (2) open-ended response quality.

With immediate visceral responses framing further experiences with a product, the impact of emotional reactions to the screen design on respondents' mood (affective state) is expected to be reflected in subjects' responses to opinion questions. Both screen designs include four identical scalar opinion questions, in which respondents rated their satisfaction with their education, classes, and advising, as well as the desirability of Pullman as place to live. Response options are presented on a seven-point scale, with 'very satisfying/desirable' at the top, 'neither satisfying nor dissatisfying' at the midpoint, and 'very dissatisfying/undesirable' at the bottom of the list. Previously, screen designs rated as more pleasing by respondents have been associated with higher opinion ratings (Coates, 2004). Therefore, the experimental design is expected to reduce the likelihood that respondents will respond in the most favorable response categories (very satisfied/desirable and satisfied/desirable).

The lack of motivation resulting from negative emotional reactions (Cuthbert, 2000) is predicted to have an even greater effect on open-ended responses, which inherently require more effort. To compare the data quality of open-ended questions, two such questions were included. First, a topical, list-style question (Q11) asked respondents, 'What businesses would you like to see in the Pullman/Moscow area that are not currently here?' Three single-row response boxes are then offered for responses, aligned one above another. Then, the final question is an optional open-ended comments page (Q31), instructing respondents, 'If you have any additional comments about this survey, please share them here.' Comments can then be inputted into a paragraph-sized text box.

Open-ended questions provide better quality data when more respondents answer with greater depth, so less thorough responses equate to increased measurement error. More interested respondents have been found to spend more time answering questions (Galesic, 2006), and those motivated to provide longer responses by graphic manipulations provide more thorough answers to open-ended questions (Christian and Dillman, 2004). Hence, the viscerally-deficient design is expected to reduce the quality of open-ended responses by reducing both the volume of respondents that provide open-ended answers and reducing the depth of responses that are offered.

To test for differences in the detail of open-ended responses, we compare the likelihood of respondents providing open-ended responses for both styles of open-ended question. On the list-style question (Q11), we predict that reduced motivation associate with the experimental design will reduce the likelihood that respondents will provide a response in each of the three text boxes. Similarly, we predict a reduced likelihood that experimental panel respondents will provide comments on the open-ended comments page (Q31).

In testing for differences in the depth of open-ended responses to the list-style question (Q11), we expect that the aesthetically-displeasing design will reduce the average number of text boxes used per respondent, as well as the number of businesses offered per respondent. Additionally, reduced response depth is expected to be reflected in paradata showing a reduced length of time spent responding to this question. On the comments page (Q31), we expect that reduced motivation on the experimental form will lessen the likelihood that respondents will provide comments with more than one theme. Additionally, of those comments provided, we expect a reduced likelihood that they will be substantively meaningful, and not simply related to the visual design of the form.

Additionally, we expect that emotional reactions to the experimental form will negatively impact the tone of open-ended responses on the comments page (Q31). Due to the production of more negative moods (affect), we predict an increased likelihood of comments with a negative tone, generally and those specifically criticizing the visual design, as well as a reduced likelihood that the comments provided will have a positive tone.

Respondent Commitment

Obtaining quality data also depends on respondent commitment. On one hand, high quality data requires respondents to provide full commitment when considering their responses, rather than simply try to 'get it over with'. Decreased effort leads to satisficing, or selecting the first justifiable response, the likelihood of which depends on respondent motivation and interest and task difficulty (Krosnick et al., 1996). Previously, respondents reporting less interest, focus, and effort have provided responses with higher inter-correlations, suggesting that their answers were not as thought out (Fricker et al., 2005). Therefore, reduced motivation associated with the experimental design is expected to increase the likelihood of satisficing, which involves picking the first justifiable response from near the top of the options (Krosnick and Alwin, 1987). Thus, we predict that the experimental screen design will result in a greater mean number of selections of 'top categories', or the options at the top of the list that include those selected by the upper-most quartile of the overall response distribution. To confirm that this is indeed the result of satisficing, we further predict that the experimental design will also produce a lower mean number of selections of 'bottom categories', or options at the bottom of the list that represent the lowest quartile of the overall response distribution.

On the other hand, surveys should seek to minimize the commitment required by respondents to effectively complete form by seek minimizing response burden (Krosnick et al., 1996). Norman (2004) proposes that poor emotional design makes using a product more difficult because being in a negative affective state reduces cognitive functioning, increasing the time needed to perform tasks and reducing the breadth of consideration that they are given. There is also evidence that variations in visual layout can lead to longer response times, presumably due to added burden (Couper, Traugott, and Lamias, 2001; Peytchev et al., 2006). Hence, the experimental screen design is expected to reduce cognitive functioning, thereby increasing respondent burden (commitment required) and resulting in lengthier response times.

Thus, we predict that the experimental form will lead to increased mean times on each question screen, overall. Additionally, we predict that the increased mean response time will be concentrated in the 'check all that apply' questions with lengthy lists of options that require more cognitive processing, and in the early stages of the survey, before respondents become accustomed to the poor visual design. Since subjects were randomly assigned to one of two versions with identical written content, statistically significant differences in response times can be attributed to screen design effects. Such an effect, in combination with an increase in satisficing, would suggest less thorough, accurate responses.

Subjective Survey Experience

With emotional reactions impacting users' perceptions and moods (Norman, 2004), we also expect that respondents' subjective interpretations of their survey experiences will be affected by screen design. The last four questions of both designs asked respondents to rate their experience completing the questionnaire as follows: (Q27) how satisfying was the survey, overall; (Q28) how difficult was it to complete; (Q29) how focused were they while completing it; and (Q30) how pleasing was the visual layout and appearance. Those ratings were reported on five-point scales, adapted to the content of each question, with '5' being 'extremely ________', and '1' being 'not at all ________'. We predict that negative perceptions associated with the experimental design will result in a reduced likelihood of respondents selecting the two most favorable categories when subjectively rating each of these four aspects survey experience.

Previously, respondents that report greater interest and lessened perceived effort have more broadly reported higher assessments of their overall survey experience after completion (Galesic, 2006). Thus, the viscerally-deficient design is predicted to result to be less likely to have respondents report that their overall experience completing the survey was a '4' or '5-very satisfying.'

In other studies, ATM screens rated as less visually satisfying design were associated with greater perceived difficulty of use (Kurosu and Kashimura, 1995; Tractinsky, 2000). Thus, we predict that the experimental design, though identical in task content, is less likely to be rated as a '1-not at all difficult' or a '2' by respondents reporting on its perceived difficulty.

Additionally, negative affective states have been associated with decreased concentration, in general (Norman, 2004), and decreased interest in completing surveys (Galesic, 2006). Therefore, it is expected that respondents will be less likely to report that their focus level was a '4' or '5-very focused'.

Just as web survey screens with viscerally-displeasing traits were previously rated as less aesthetically pleasing by users (Coates, 2004), we expect respondents to the experimental design to be less likely to rate the questionnaire as a '4' or '5-very pleasing' in terms of how 'visually pleasing' it is.

Methods

To test whether an aesthetically-displeasing web survey design reduces data quality by these indicators, we conducted an experimental comparison between two different screen designs for the same web survey. Since aesthetic differences are the focus, we intended for respondents to be capable of understanding and completing both designs, so the two designs have identical verbal content and differ only in visual appearance. Before comparing data from separate designs, we first had to develop theoretically-appropriate experimental and control screen designs, which are described below.

Development of the Aesthetically-Displeasing Screen

To effectively test the overall hypothesis that aesthetically-displeasing screen will influence response quality, the experimental design was designed not just to be unattractive, but to be displeasing at the visceral level where the emotional reactions are founded. Hence, it includes fundamental visual traits known to produce visceral repulsion, as described by Norman (2004). Since the design possibilities of web surveys allow for a wide range of visual choices and the predicted effects depend on the design producing negative emotional reactions, we aimed to ensure the potential impact by including an array of viscerally-displeasing traits throughout the experimental design. The experimental design thus represents the worst end of the aesthetic spectrum that might reasonably be encountered.

As a result, any data quality impacts observed indicate the impact of aesthetically-displeasing screens overall, not of any one trait. Thus, this initial experiment is meant to determine if the negative emotional reactions produced by aesthetically-displeasing screen design leads to reduced data quality in a number of ways. If such effects are observed, future studies could isolate the impacts of particular traits.

In applying the aesthetic traits that elicit such visceral responses to the strictly visual medium of web survey design, we discuss the screen designs in terms of whether or not they exhibit three broad, viscerally-pleasing characteristics: visual harmony, visual rhythm and visual comfort. Visual harmony is described as a pleasing agreement between various visual elements, particularly in terms of balance, color, and organization (Brady and Phillips, 2003). Rhythm involves repetition of visual elements and consistent use of graphical language. Visual comfort stems from stimuli that are smooth, simple, and symmetrical. The design of the experimental screen aimed to reduce all of these, while the control design aimed to increase them.

Elements of Screen Design Being Manipulated

The screen design variations that may impact emotional reactions can be understood within the framework of seven web survey design considerations found in Table 2. Below, we describe how each aspect can affect visceral reactions and their application in the experimental and control panels designs (also see Table 2). Screenshots of the log-in page, close-ended, and open-ended screen designs (Figures 2-4) provide visual examples of these manipulations.

Grouping of Visual Elements

Viscerally, we are repulsed by complex scenes (Norman, 2004), so web survey screens should be delineated into relatively few visual elements, to add comfort through simplicity, and the grouping should be consistent, to increase rhythm. Such use of visually delineated headers and spatial separation in web design has previously led to decreased burden and increased satisfaction for users (Chaparro et al., 2005). With more elements comes increased complexity, or "screen clutter", which has been tied to greater search times, more errors, and reduced respondent satisfaction (Tullis, 1984; Weller, 2004).

In the control version, all of the task-oriented information is grouped within a single visual element, via inclusion within the blue/green background area (Figure 2). This simple grouping adds visual comfort and its consistent application throughout the form adds rhythm. In the experimental design, the same information is grouped into multiple visual elements set against a variety of background colors, increasing complexity and creating visual inconsistency that reduces rhythm.

Shape of Visual Elements

Simple, basic geometric shapes are known to be viscerally-pleasing (Norman, 2004), explaining why web surveys using basic shapes and symmetry have been found to be rated as "good" and "more appealing" more often (Coates, 2004).

By grouping the task-oriented, header, and footer information into rectangular elements, the control panel thus creates visual comfort (Figure 2). The many visual elements in the experimental design, on the other hand, have boundaries that fit that section of text on each screen, creating complex, many-sided shapes that reduce comfort and vary by screen, reducing rhythm (Figure 2).

Color Scheme

Color has a particularly strong impact on visual appeal (Knutson, 1998) and is known to be a strong predictor of a website's overall appeal (Lindgaard, 1999; Brady and Phillips, 2003). Individual colors can elicit visceral responses, as can the combination of hues on a single screen (Norman, 2004). Poorly selected colors on web sites have been tied to increased perceived difficulty and burden (Novemsky, 2004; Reber and Schwarz, 1999) and to increases in early terminations in web surveys (Pope and Baker, 2005).

All task-oriented information on the control design is set against a semi-transparent blue/green background that is a soft, pastel color, which Norman (2004) describes as viscerally comforting and which have been found particularly appealing to web site users (Lindgaard, 1999). The only other colors are neutral, gray-scale hues, making for a harmonious combination of colors on each screen (Figures 2-4). In the experimental panel, the primary background is a harsh neon-purple hue that is viscerally-displeasing, as are the harsh green, yellow, and red backgrounds of the individual elements (Figures 2-4). Together, those colors fail to follow vision science guidelines for effectively combining hues (Palmer, 1999), reducing harmony.

Screen Organization

Good balance in visual design involves "equal distribution of visual weight along the horizontal and vertical axis" (Lauer and Pentak, 2002:76) and provides websites with a

Table 2: Comparison of control and aesthetically-displeasing screen design characteristics, by design trait

Design Trait	Control Screen Design	Viscerally-Displeasing Screen Design
Grouping of Visual Elements	• Delineates relevant info via placement within blue/green background area • Single, consistent element adds visual comfort, rhythm	• Delineates different info into many different elements, reducing visual comfort • Groups elements via varying background colors, reducing visual rhythm
Shape of Visual Elements	• Visual elements are symmetrical, basic shapes that provide visual comfort	• Visual elements are complex shapes, reducing visual comfort • Elements vary in shape according to text within, reducing visual rhythm
Color Scheme	• Uses pleasing blue/green pastel tone as primary background, adding to visual harmony • Other colors are neutral, avoiding disharmonious color combinations	• All info set against harsh purple background, reducing comfort • Green, red, and yellow also used as background, creating discordant color scheme that reduces visual harmony
Screen Organization	• Main element centered, secondary elements balance each other • Creates balanced layout that adds to visual harmony, comfort	• Varied justifications of individual elements creates unbalanced layout, reducing visual harmony, comfort
Font Selection	• Uses only legible, pleasing Arial font • Maintains text style for all task info, only varying size, adding to visual rhythm	• Varies between Arial and displeasing Times, reducing rhythm and comfort • Uses many text modifications inconsistently, reducing harmony
Whitespace	• Uses whitespace between task and secondary info to visually separate screen, adding visual comfort • Avoids large empty areas	• Eliminates border delineating task information, reducing comfort • Extra whitespace makes overall screen seem empty, reducing visual harmony
Response Options	• Response options are unmodified black text against pleasing blue/green background, adding visual rhythm • White answer boxes use figure/ground contrast to draw focus to task, adding visual comfort	• Response options set against harsh yellow background also used to delineate other info, reducing visual harmony and rhythm • Options are underlined and in a different font than stem, creating clutter and reducing visual comfort

sense of psychological equilibrium and visceral agreement. Unbalanced screens have been found to contribute to reduced overall satisfaction, aesthetic appeal, and perceived usability (Brady and Phillips, 2003).

Here, the control design is well-balanced, with the main visual element (holding the most visual weight) centered vertically and horizontally and secondary elements (the header and footer) balancing each other out by being equally off-center vertically (Figure 3). In contrast, the visual elements of the experimental design are unbalanced, varying between left-, center-, and right-alignment, placing the visual weight off-center and reducing visual harmony (Figure 3).

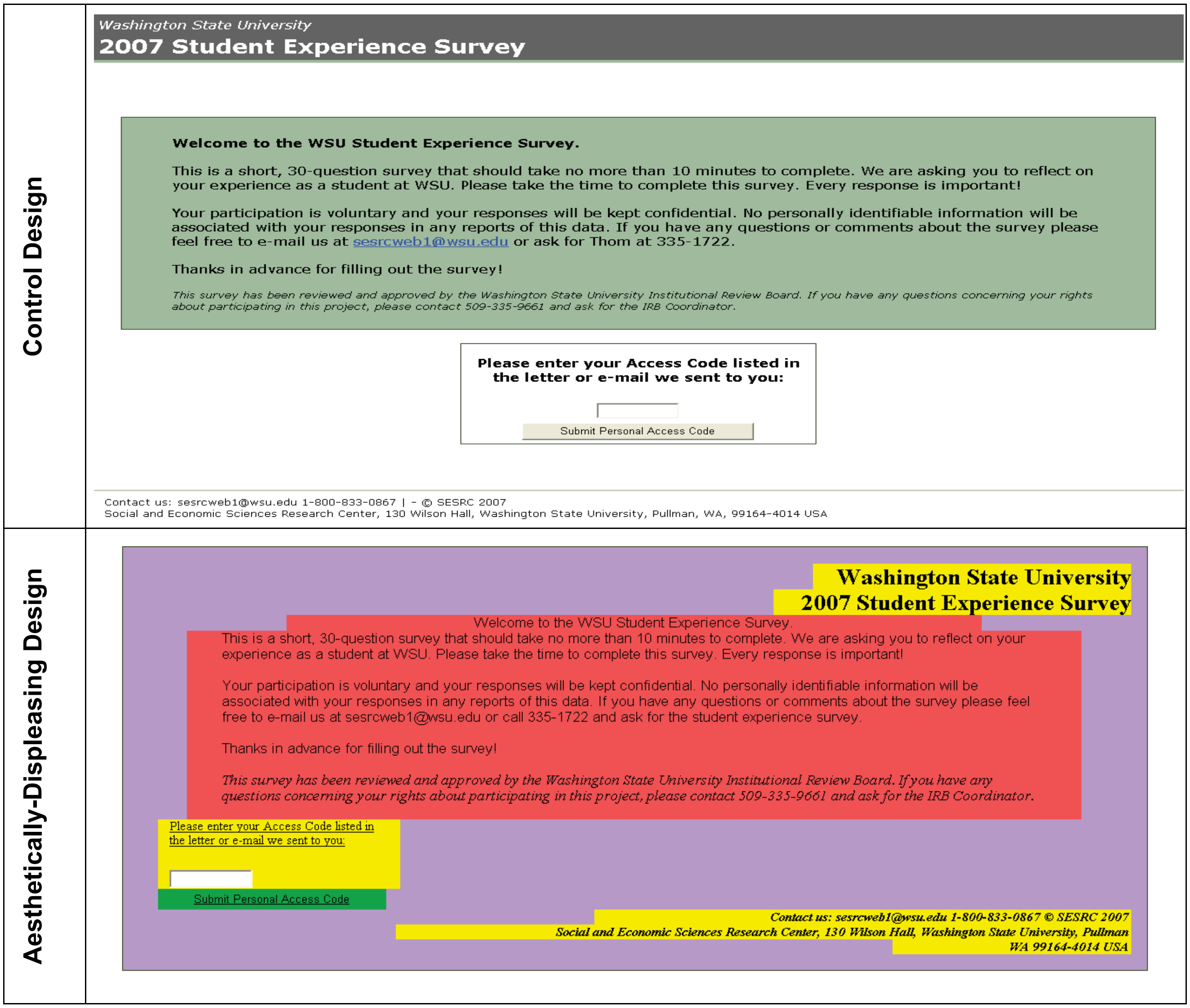
Font selection

Good emotional design delineates different types of information by varying them visually because our eyes naturally segment visual scenes, organizing them into regions based on shared visual characteristics (Palmer, 1999). At the same time, too much variation can lead to a loss of rhythm and add to screen clutter, reducing visual harmony.

In the control design, the text size varies between thematic sections to help separate different types of information, aiding visual rhythm (Figures 2-4). At the same time, it consistently uses Arial font, which has been found to be pleasing for web site users (Bernard et al., 2001), to maintain that rhythm and add comfort. The experimental design instead alternates between Arial and Times New Roman fonts, incorporating multiple modifications of size and style (Figures 2-4). The inconsistent presentation of the text reduces visual rhythm and harmony, while the addition of Times New Roman font reduces visual comfort, as it has been found to be among the least pleasing fonts (Bernard et al., 2001).

Whitespace

The information provided on a web screen can effectively use non-attended-to background space, or ‘whites-


Control Design
Aesthetically-Displeasing Design
Submit Personal Access Code
Submit Personal Access Code

Washington State University
2007 Student Experience Survey

Welcome to the WSU Student Experience Survey.

This is a short, 30-question survey that should take no more than 10 minutes to complete. We are asking you to reflect on your experience as a student at WSU. Please take the time to complete this survey. Every response is important!

Your participation is voluntary and your responses will be kept confidential. No personally identifiable information will be associated with your responses in any reports of this data. If you have any questions or comments about the survey please feel free to e-mail us at sesrcweb1@wsu.edu or ask for Thom at 335-1722.

Thanks in advance for filling out the survey!

This survey has been reviewed and approved by the Washington State University Institutional Review Board. If you have any questions concerning your rights about participating in this project, please contact 509-335-9661 and ask for the IRB Coordinator.

Please enter your Access Code listed in the letter or e-mail we sent to you:

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 | - © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman, WA, 99164-4014 USA

Washington State University
2007 Student Experience Survey

Welcome to the WSU Student Experience Survey.

This is a short, 30-question survey that should take no more than 10 minutes to complete. We are asking you to reflect on your experience as a student at WSU. Please take the time to complete this survey. Every response is important!

Your participation is voluntary and your responses will be kept confidential. No personally identifiable information will be associated with your responses in any reports of this data. If you have any questions or comments about the survey please feel free to e-mail us at sesrcweb1@wsu.edu or call 335-1722 and ask for the student experience survey.

Thanks in advance for filling out the survey!

This survey has been reviewed and approved by the Washington State University Institutional Review Board. If you have any questions concerning your rights about participating in this project, please contact 509-335-9661 and ask for the IRB Coordinator.

Please enter your Access Code listed in the letter or e-mail we sent to you:

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman WA 99164-4014 USA

Figure 2. Screenshots of Log-In Screen for Control and Viscerally-Displeasing Screen Designs

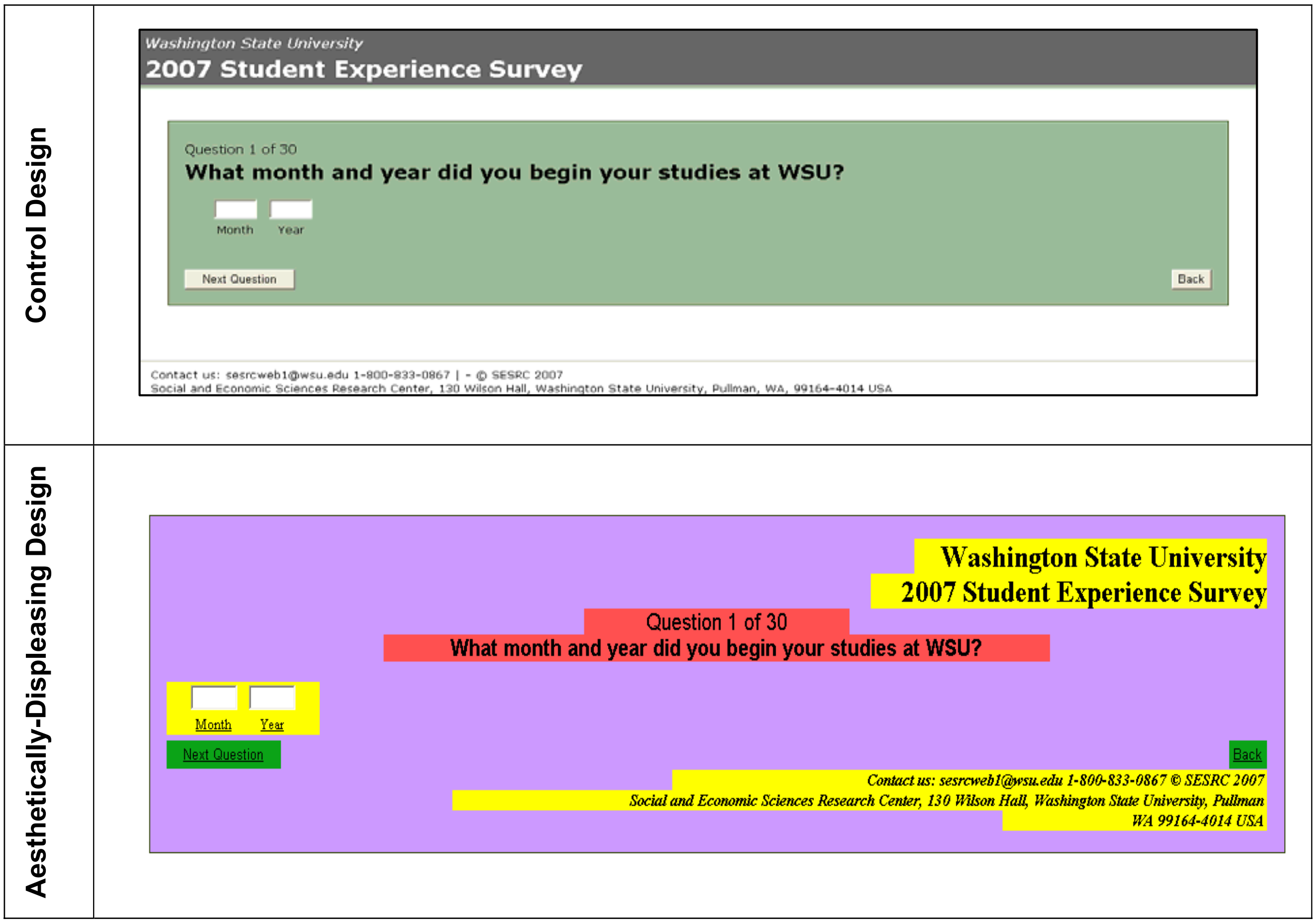
pace', to "organize the material on the page" by separating different functions and tasks (White, 1974:48). Too much whitespace, on the other hand, can give screen layouts an "empty" look (Bernard et al., 2000) and elicit lower perceptions of overall appearance (Spool et al., 1997).

The single primary element and simple screen organization of the control panel make judicious and consistent use of whitespace, adding visual harmony and rhythm (Figures 3, 4). The experimental screen instead sets all visual elements against a single background, creating excessive "whitespace" (actually purple, here) that reduces harmony, while the varied placement of those elements within the whitespace reduces rhythm (Figures 3, 4).

Response options format

It has been suggested that, due to their task importance, response options should be visually highlighted using figure/ground contrast and minor text modifications to help focus respondents' attention, increasing comfort and rhythm (Dillman et al., 2005). The control design does this, highlighting the white answer spaces (figure) against the blue/green background (ground) and using smaller, Arial font than in question stem (Figure 4).

At the same time, excessive variations in the response options can create screen clutter, which reduces visual harmony (Brady and Phillips, 2003). While lacking figure/ground contrast to focus attention, the response options on the experimental design are visually delineated into a separate element set against a displeasing yellow background



Washington State University
2007 Student Experience Survey

Question 1 of 30
What month and year did you begin your studies at WSU?

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 | - © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman, WA, 99164-4014 USA

Washington State University
2007 Student Experience Survey

Question 1 of 30
What month and year did you begin your studies at WSU?

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman
WA 99164-4014 USA

Figure 3. Screenshots of Question 1 for Control and Viscerally-Displeasing Screen Designs

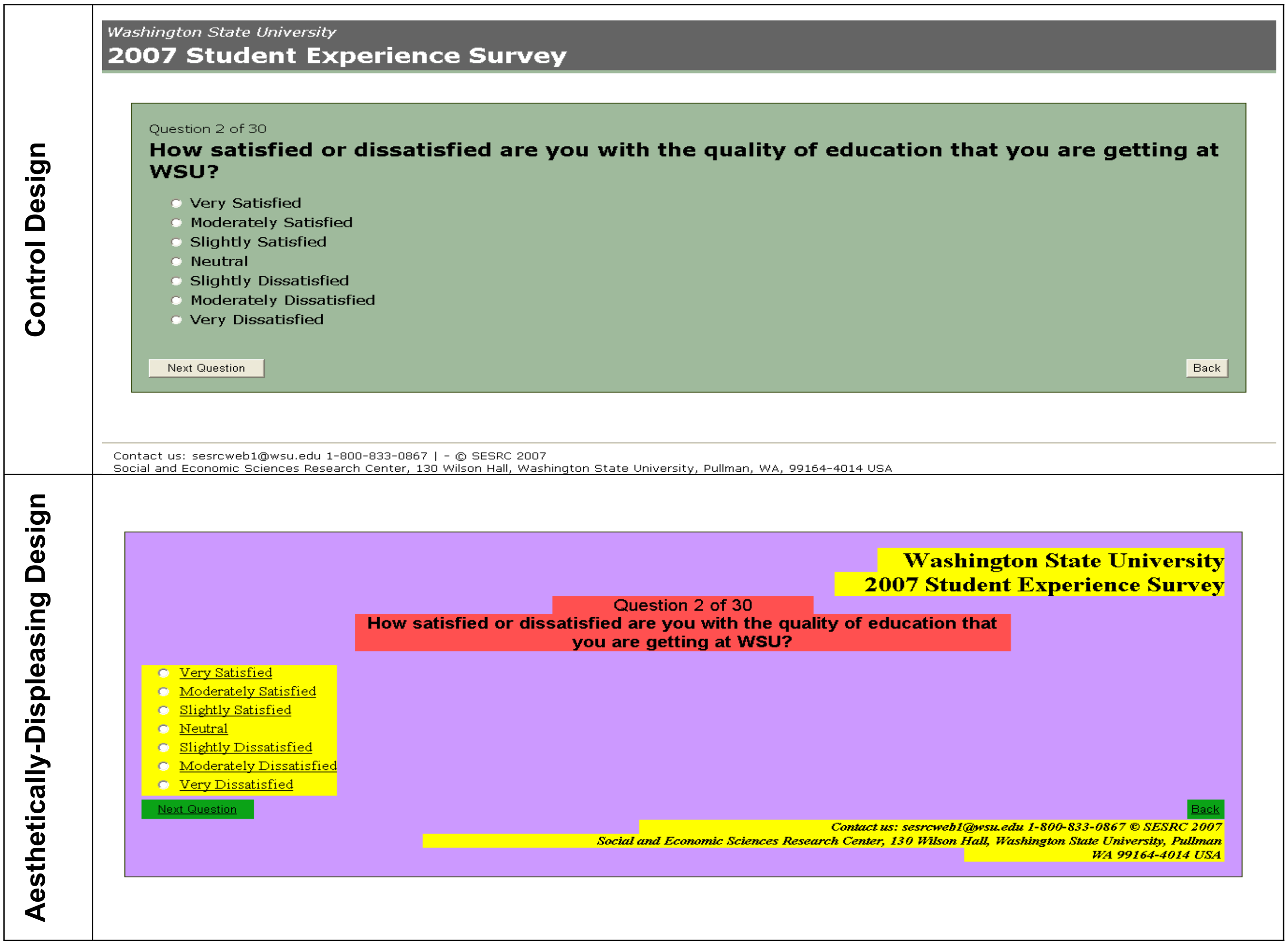
color that is also used for other thematic sections, reducing rhythm and harmony. The text of the options also differs greatly from that of other elements, creating uncomfortable clutter (Figure 4).

Evaluation of Screen Designs for Carrying-out the Design Objectives

The validity of this experiment hinges upon the experimental and control designs eliciting the intended emotional reactions from respondents. Theoretically, the control design exemplifies good emotional design, with visual traits known to elicit positive visceral responses, while the experimental design includes multiple viscerally-displeasing traits known to influence negative emotional reactions. Additionally, the control design has also been successfully used in 10 previous web surveys, including five previous iterations of this Student Experience Survey, in which it elicited virtually no negative reactions from participants.

However, the emotional reactions expected to produce data quality differences would not be directly evident in the data collected, so we sought to further verify the designs' validity by conducting 30 cognitive interviews with students from the same population as the sample was drawn. Participants completed both the control and experimental versions, and their unprompted reactions were noted before a series of retrospective comparative questions were asked.

Results from those interviews, which will be analyzed more thoroughly elsewhere (Mahon-Haft, in progress), verified that the screen designs indeed produced different emotional reactions. Evidence of negative emotional reactions was common in response to the aesthetically-displeasing design, both physically and verbally during completion of the form and during the retrospective questioning (see Table 3). This was not the case for the control design, which elicited very few negative responses and many more positive reactions (Table 3). In reporting their preferences between designs, 26 of 30 respondents chose the control panel as the 'better screen design' and only four chose the viscerally-deficient version. Additionally, 11 respondents reported they would be 'less likely to begin' the experimental version and 14 reported they would likely 'provide less effort', compared to only one person said the same about the control version to each question. The experimental version was described by some as 'unprofessional', 'distracting', and 'ugly,' while the



Washington State University

2007 Student Experience Survey

Question 2 of 30

How satisfied or dissatisfied are you with the quality of education that you are getting at WSU?

- ○ Very Satisfied
- ○ Moderately Satisfied
- ○ Slightly Satisfied
- ○ Neutral
- ○ Slightly Dissatisfied
- ○ Moderately Dissatisfied
- ○ Very Dissatisfied

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 | - © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman, WA, 99164-4014 USA

Washington State University
2007 Student Experience Survey

Question 2 of 30
How satisfied or dissatisfied are you with the quality of education that you are getting at WSU?

- ○ Very Satisfied
- ○ Moderately Satisfied
- ○ Slightly Satisfied
- ○ Neutral
- ○ Slightly Dissatisfied
- ○ Moderately Dissatisfied
- ○ Very Dissatisfied

Contact us: sesrcweb1@wsu.edu 1-800-833-0867 © SESRC 2007
Social and Economic Sciences Research Center, 130 Wilson Hall, Washington State University, Pullman WA 99164-4014 USA

Figure 4. Screenshots of Question 2 for Control and Viscerally-Displeasing Screen Designs

control panel was repeatedly described as 'professional' and 'organized', while (Mahon-Haft, in progress). Collectively, the cognitive interviews provided evidence that the control design constituted 'good' emotional design, and that the experimental version was viscerally-displeasing and might detrimentally impact respondent behavior.

Data Collection Procedures

We fielded two visually-distinct, but otherwise identical, panels of a web survey, obtaining separate data from the aesthetically-displeasing and control forms. The instrument was a 31-question web-based Student Experience Survey assessing undergraduate student experiences at Washington State University, conducted between February and April of 2007. Each panel initially included 600 undergraduate students randomly selected from a list of all undergraduates (obtained from the university) using a random number generator.

Students that were part of the sample were originally contacted via a postal letter containing a $2 incentive, explaining the survey's purpose. It mentioned that the survey was being conducted by the Social and Economic Sciences Research Center, a widely recognized survey research department on campus that annually does numerous surveys of undergraduates. This letter also provided the URL for the log-in page and a personal access code required for entry. That letter was followed by an email with similar verbal content sent three days later to non-respondents for whom official university email addresses were available. Thereafter, remaining non-respondents were contacted three more times, twice by email and once more by postal mail. The overall response rate for these two panels of the survey was 56.3% (640 of 1200).

All of the question screens were constructed using HTML tables where proportional widths were programmed in order to maintain a consistent visual stimulus regardless of individual user screens. Cascading Style Sheets were used to automatically adjust font size and style to accommodate varying user screen resolutions and achieve similar images on the computer screens of all respondents. The same host server was used to administer and collect data for both versions.

Table 3: Frequencies of different types of emotional reactions expressed during cognitive interviews, by form

	Aesthetically Displeasing				Control Form			
Type of Emotional Reaction	Think Aloud	Retrospective	Physical	Any Type	Think Aloud	Retrospective	Physical	Any Type
General Discontent	23	26	17	26	0	1	4	4
Distraction	8	13	7	19	0	0	0	0
Burdened	9	15	14	18	0	0	0	0
Aesthetic Discomfort	18	24	11	26	0	0	0	0
Boredom	0	0	0	0	0	0	0	0
Positive Reactions	5	5	0	5	12	26	0	26
Total	63	83	49	94	12	27	4	30

For both versions, client-side paradata were also collected, tracking every key stroke and navigational movement along with the length of time between the moment the screen is loaded on the respondent's computer and the respondent submits an answer. This tool, initially used by Heerwegh to analyze web survey responses (2003), provides more details about the way in which surveys are completed, step by step, and has been recently used to deepen analysis in the field (see, for example, Stern et al., 2006). By providing detailed information on survey access and the time spent responding, paradata was used to supplement response data in understanding data quality.

Findings

With such a wide spectrum of predictions, analysis of the data collected required the use of multiple statistical techniques. Whenever possible, we used binary logistic regression to compare the likelihood of dichotomous outcomes, with the control form as the reference category. Those maximum likelihood predictions provide us with p-values to test for significant differences in outcome likelihood, as well as odds ratios reflecting the magnitude of any observed differences. For predictions where dichotomous outcomes were inappropriate measures, we utilized one-sided chi-squared tests to compare respondent distributions and one-sided t-tests to compare differences in means.

Response Rates

Response rates are a primary indicator of data quality, prompting tests for differences in non response patterns between designs. Both panels had virtually identical rates of initial participation, with 55.5% (333 of 600) of respondents beginning the control form and 57.0% (342 of 600) beginning the experimental form. Using chi-squared tests to compare the distribution of non-responses, early terminations, and completions we found that, contrary to expectations, completion rates and early terminations did not differ between forms (see Table 4).

Additionally, measurement error results when respondents skip questions, so we compared the likelihoods that respondents would skip items when completing each form. As predicted, respondents completing the experimental form omitted items significantly more often (67.2%, vs. 58.7%), and binary regression tests showed that they were 1.5-times as likely to do so (see Table 5).

Substantive Response Data

To test for a potential negative skew on opinion ratings, the questionnaire included four scalar opinion questions, in which respondents rated the satisfaction and desirability of various aspects of their student experience. Using binary logistic regression, we tested for differences in the likelihood of selecting the two most favorable subjective categories.

As expected, respondents completing the aesthetically-displeasing form were less likely to report particularly favorable opinions for all four questions, suggesting negatively skewed opinions related to screen design. However, the differences were only significant at the .05 level for two questions (Q7 and Q10), for which the likelihood of highly favorable ratings was approximately 0.7-times as likely to occur (see Table 6).

Data from the list-style open-ended question (Q11), which asked respondents what businesses they wished to have nearby, was coded to reflect both response volume, as indicated by the likelihood of that each of the three text boxes was used, and response depth, indicated by the number of businesses listed, boxes filled in, and characters typed (overall and in the final box), as well as the amount of time spent on this question screen.

As seen in Table 7, there was trend towards reduced volume on the experimental form, on which respondents were less likely to use each of the text boxes. However, our prediction is only partially confirmed, as the difference is only significant in the third text box, where respondents were again about 0.7-times as likely to answer. Notably, the odds ratios grow progressively smaller with each successive text box, suggesting that the effect is greater where more effort is required.

The pattern remains much the same in terms of response depth, where all three measures reflect less thorough responses on the experimental form, two of which are statistically significant (see Table 8). On average, of those who filled in at least one text box (total n = 279), those filling out the aesthetically-displeasing form (n=144) reported in 9% fewer boxes (1.66 vs. 1.82, p= .03) and spent about 18%

Table 4: Completion Rates and Experimental Design Effects, by Panel

	Potential Respondents		Experimental Design Effects		
Completion Status	Control (%)	Experimental (%)	% change	χ^2	p value
Non-Responses	267 (44.5%)	258 (43.0%)	-1.5%	0.27	0.60
Early Terminations	20 (3.3%)	15 (2.5%)	-0.8%	1.40	0.22
Completions	313 (52.2%)	327 (54.5%)	+2.3%	0.44	0.51
Total	600 (100%)	600 (100%)			

Overall χ^2=3.539, p=0.316, d.f.=2

Table 5: Likelihood of Respondents Omitting Items, by Panel

	Respondents		Experimental Design Effects		
Likelihood of . . .	Control (%) n=312	Experimental (%) n=326	b	odds	p value
Skipping a Question	183 (58.7%)	219 (67.2%)	.367	1.443	.02

Overall χ^2=4.975, 2 Log Likelihood=836

less time completing this question screen (52.8 vs. 64.1 seconds/response, p= .00). With a similar, though statistically-insignificant, trend towards businesses provided in those responses, these results generally support the expected reduction in response depth.

Data from the optional open-ended comments screen at the end of the survey (Q31) were also coded to reflect whether comments were offered, as well as their conceptual themes (visual, elaboration of earlier responses, content suggestions, etc.) and their tone (positive or negative). Table 9 reports on the volume and depth of those comments, demonstrating that those completing the two designs were equally likely to provide comments and to offer responses with more than one theme. Thus, these indicators of open-ended response of volume and depth disconfirm our predictions.

Deeper analysis of the content of the comments that were offered, however, suggests that the aesthetically-displeasing design may still have had a detrimental effect on comment depth. Among those providing comments (total n=141), the likelihood of providing at least one theme that was not strictly a visual design criticism was much lower on the experimental panel (10 percent as likely, see Table 10). At the same time, experimental panel respondents were 9.4 times more likely to provide negative feedback strictly related to the visual appearance (p=.001). This implies that the volume and depth predictions disconfirmed above may have been skewed when respondents who would otherwise not offer comments did so simply to attack the form's appearance.

Analysis of the tones of the optional comments can also be found in Table 10, confirming our prediction that the worsened mood expected of respondents completing the experimental would translate into comments with a negative tone. Comments from experimental panel participants were indeed 2.3 times more likely (p= .01) to include remarks with a negative tone, an impact echoed by their decreased likelihood (though it is not quite statistically significant) to provide comments with a positive tone.

Respondent Commitment

Based on expectations that the experimental design would reduce the amount of commitment provided by respondents, we predicted that satisficing would increase, leading to increased evidence of primacy. To test this prediction, we used t-tests to compare the mean number of times respondents selected "top" and "bottom" categories, or those options representing the quartiles of responses highest and lowest on the lists.

These tests supported our predictions, demonstrating that the responses of those completing the experimental form demonstrated a greater tendency towards primacy (see Table 11). They selected top categories 10.6 times per respondent, four percent more often than those completing the control form (p= .03). Additionally, the experimental form produced, on average, five percent fewer bottom category selections per respondent (Table 11). This evidence of increased primacy further confirms that their response distribution was skewed towards the top of the list. More broadly, increased primacy indicates more satisficing, confirming the predicted reduction in respondent commitment provided.

Compounding the loss of commitment provided, we expected an increase in the commitment required to complete the aesthetically-deficient design. Using paradata to record response times on each screen, we predicted lengthier response times when completing the experimental form, overall and for early questions and question formats requiring greater effort. Other factors, such as connection speeds and multi-tasking while on-line, can impact response times, so we excluded from our data the extreme outliers (response times greater than two standard deviations from the mean completion time for that particular question screen).

According to paradata and cognitive interview subjects, even the experimental screen design does not keep this survey from being particularly quick and easy to complete. Collectively, all respondents needed only seven minutes to log

Table 6: Likelihood of selecting two most favorable responses to opinion questions, by question, by panel

	Selecting Most Favorable Options		Experimental Design Effects		
Likelihood of Selecting Options for...	Control (%) n=324	Experimental (%) n=331	b	odds	p value
Q2- Sat w/ Education	270 (87.3%)	261 (78.9%)	-.293	.746	.14
Q4- Sat w/ Advising	148 (46.1%)	146 (41.3%)	-.194	.824	.22
Q7- Sat w/ Classes	228 (72.2%)	210 (63.8%)	-.384	.681	.02
Q10- Desirability of Pullman	207 (65.3%)	190 (58.1%)	-.305	.737	.05

Table 7: Likelihood of providing list-style open-ended responses (Q11), by panel

	Respondents Using Text Box		Experimental Design Effects		
Likelihood of Responding to Q11 using...	Control (%) n=324	Experimental (%) n=331	b	odds	p value
Text Box #1	271 (83.6%)	269 (81.0%)	-.180	.835	.38
Text Box #2	190 (58.6%)	177 (53.3%)	-.216	.805	.17
Text Box #3	132 (40.7%)	109 (32.8%)	-.341	.711	.03

in and complete all 31 questions, none of which were particularly sensitive and only two of which even potentially involved any typing. Within that context, paradata demonstrated that the experimental design indeed required greater respondent commitment (see Table 12). The aesthetically-displeasing design produced overall mean response times that were only slightly longer (17.3 vs 15.2 sec/screen), but the difference is meaningful in this survey setting (13.8% growth) and is significant (p= .00, total n=19,865).

Additionally, comparisons of response times for specific types of questions show that the experimental design increased mean response times in every format except the open-ended question (Q11). Even scalar questions demonstrate a small (5%), but significant (p= .00), increase in completion times for the experimental panel, and the check-all-that-apply questions that require greater attention reflect a larger increase of 22.8% (p= .00). Also as expected, the increased commitment required was exaggerated early in the form (Table 12), when respondents had not yet adjusted to the harsh and unexpected appearance of the aesthetically-displeasing design. The open-ended question showed the opposite effect, with average response times that were 11 seconds less per screen among those completing the experimental form. That reversal was also expected, and its magnitude suggests that the increased effort and commitment required to complete the experimental form would be larger in a more challenging survey setting.

Subjective Survey Experience

The last four questions of both designs asked respondents to subjectively rate their experience completing the questionnaire in terms of overall satisfaction, perceived difficulty, perceived focus, and visual appeal. Expecting perceptions to be negatively skewed on the experimental form, we expected that respondents would be less likely to report in the most positive categories for each question.

However, Table 13 shows only the visual appearance of the experimental design was significantly less likely to be rated as such (odds ratio= .267, p= .00). While the experimental form was slightly less likely to be rated in the highest categories in terms of overall satisfaction (odds ratio= .907) and perceived focus (odds ratio= .854), those differences were minor and statistically insignificant. Most surprisingly, the aesthetically-displeasing design was actually 1.65 times more likely to be rated as a ‘1’ or ‘2’ on the difficulty scale, contradicting our predictions and much previous research (eg., Lindgaard, 2007). Though that effect was not quite significant at the .05 level (p= .08), it potentially suggests that the experimental design did not increase perceived burden.

Discussion and Conclusions

This web survey experiment was designed to examine the possible data quality effects stemming from people’s innate emotional reactions to aesthetic qualities of screen design. Four potential avenues of impact were analyzed: response rates, substantive response behavior, respondent commitment, and subjective survey experience. As suggested by Norman’s (2004) ideas on the importance of emotional design, the viscerally-displeasing design impacted data quality in each of those areas. In all, 28 indicators of data quality were examined, 15 of which resulted in less desirable results on the experimental version (Figure 4). Importantly, there was not a single test where the control version produced significantly less desirable results.

However, the impact of the aesthetically-displeasing screen design was not as widespread as expected. Nonresponse rates were only slightly affected, as the completion rates and early termination rates were not significantly different and the only impact was an increased likelihood of item omissions. In addition, subjective interpretations of the survey were barely impacted, with the only significant change

Table 8: Indicators of Response Depth for List-Style Open-Ended (Q11), by Panel

	Mean Values (std. dev.)		Experimental Design Effects[a]		
	Control	Experimental			
Response Depth Indicators	n=144	n=135	% change	t-test	p value
Text Boxes Used	1.82 (1.1)	1.66 (1.1)	-9%	1.84	.03
Business Names & Types	2.01 (1.5)	1.91 (1.7)	-5%	.858	.19
Response Time (sec)	64.1 (108.0)	52.8 (57.6)	-18%	2.582	.00

[a] d.f.=277 for all tests

Table 9: Likelihood of Providing Optional Open-Ended Comments (Q31), by Panel

	Respondents Providing Comments		Experimental Design Effects		
	Control (%)	Experimental (%)			
Likelihood of providing...	n=324	n=331	b	odds	p value
Any Comments	70 (21.6%)	71 (21.4%)	-.009	.991	.96
More than One Theme	37 (11.4%)	37 (11.1%)	-.024	.977	.92

Table 10: Likelihood of Different Types of Optional Comments (Q31), by Theme, by Panel

	Respondents Providing Comments		Experimental Design Effects		
	Control (%)	Experimental (%)			
Likelihood of providing...	n=70	n=71	b	odds	p value
Non-Visual Comments	69 (98.6%)	62 (87.3%)	-2.304	.100	.03
Negative Visual Design	3 (1.0%)	21 (7.1%)	2.239	9.380	.00
Negative Tone (All Themes)	32 (45.7%)	47 (66.2%)	.844	2.326	.01
Positive Tone	19 (6.2%)	12 (4.0%)	-.605	.546	.14

Table 11: Selections of Top and Bottom Categories, by Stage, by Panel

	Mean Values (std. dev.)		Experimental Design Effects[b]		
	Control	Experimental			
	n=324	n=331	% change	t-test	p value
Top Category Selections	10.2 (2.7)	10.6 (2.9)	+4%	1.885	0.03
Bottom Category Selections	6.4 (2.4)	6.1 (2.4)	-5%	-1.675	0.04

[b] d.f.=654 for both tests

Table 12: Mean Response Times and Experimental Design Effects, Overall and by Question Type

	Mean Response Times (std. dev.)		Experimental Design Effects[c]		
Question Type	Control	Experimental	% change	t-test	p value
Early Questions (Q1-10)	12.0 (12.3)	13.4 (17.5)	+11.6%	3.61	.00
Scalar	7.7 (5.8)	8.1 (9.0)	+5.2%	3.54	.00
Check-All (Q's 3,5,13)	22.8 (9.8)	28.0 (25.6)	+22.8%	3.29	.00
Open-Ended (Q11)	64.1 (108.0)	52.8 (57.6)	-17.6%	-2.58	.00
Total- All Stages (Q1-30)	15.2 (22.2)	17.3 (51.3)	+13.8%	-3.84	.00

[c] Due to different numbers of questions of each type and elimination of outliers, n values vary by test, but all are over 1,000 due to separate observations for each question screen

Table 13: Likelihood of Selecting Two Most Desirable Subjective Survey Experience Options, by Question, by Panel

	R's Selecting Desirable Options		Experimental Screen Effects		
Likelihood of Selecting Options for...	Control (%) n=324	Experimental (%) n=331	% b	odds	Prob
Overall Satisfaction (Q27)	164 (50.6%)	160 (48.2%)	-.097	.907	.53
Perceived Difficulty (Q28)	290 (89.5%)	310 (93.4%)	.502	1.65	.08
Perceived Focus (Q29)	224 (69.1%)	218 (65.7%)	-.158	.854	.34
Aesthetically Pleasing (Q30)	153 (47.2%)	64 (19.3%)	-1.321	.267	.00

being in ratings of visual appeal; the difficulty ratings for the experimental form actually showed an insignificant trend towards better ratings.

Still, there were some potentially important differences, as the experimental design had a strong impact on many data quality indicators, particularly those related to response behavior and respondent commitment (see Table 14). Where effort was required to answer open-ended questions, respondents to the experimental design were less likely to offer complete responses on a list-style question and the responses provided on both questions were less thorough, reducing the quality of the data obtained. The negative effects of the viscerally deficient design was also reflected in the data obtained, on both scalar and open-ended questions. Responses to scalar opinion questions were negatively skewed (two of four comparisons were statistically significant) and open-ended responses were more likely to include comments with a negative and less likely to include those with a positive tone. Finally, there was strong evidence that respondents provided less commitment when completing the experimental form, where there was evidence of increased satisficing resulting in a primacy effect. At the same time, this form required more commitment from respondents, as demonstrated by the lengthier completion times. Together, these differences suggest that the experimental design increased measurement error in a number of ways.

However, these results must be interpreted within the context of the college-student population studied, the administrators' existing legitimacy and persistence, and the ease and innocuous content of the survey instrument. They can only be projected to apply in similar survey contexts, and these limitations seem to suggest that the aesthetically-deficient screen design may have a greater impact on data quality in other contexts.

Strong impacts were observed on open-ended responses and the effort provided by and required of respondents, suggesting that when greater effort is required the effects of screen design will be stronger. This implies that less computer-savvy populations, who will be more burdened by web surveys, are likely to be even more strongly impacted by poor screen design. Similarly, since legitimate survey sources improve people's motivation to respond (Groves et al., 2000; Dillman, 2007) and to provide full effort (Krosnick, 1991), a questionnaire from a less legitimate source would likely increase the impact of aesthetically-poor screen design. Finally, a questionnaire that is not as easy to complete (this one having taken an average of just over seven minutes per respondent) or asks sensitive questions, both of which factor into respondent motivation (Krosnick et al., 1996), seems likely to increase the impact of screen design. Considering this study's population, administration, and content, these findings also suggest that screen design may have less effect than in other circumstances.

In this survey context, our findings provide limited support for application of Norman's (2004) contention that emotional reactions can impact user behavior to a web survey setting. Collectively, these findings suggest that the aesthetically-displeasing design led to negative emotional reactions that impacted some aspects of data quality, but not all. The unattractive design consistently reduced respondent effort, increased burden, and skewed the answers we received, increasing measurement error. Additionally, all observed differences suggested less valuable data from the experimental version. However, it had only minor effects on non-response error and respondents' perceptions of the survey, suggesting that people will still complete web surveys, even when they are ugly. Hence, visual appeal mattered.

Overall, this study provides limited evidence suggesting that aesthetically-displeasing screen design can detrimentally impact respondents' behavior, supporting recent theoretical contentions that appearance alone can potentially impact data quality (Dillman et al., 2005). If emotional reactions to a short, easy survey with an aesthetically-displeasing design can have this demonstrable of an impact among a young, computer-proficient population, how else might reactions to screen design impact survey data in other contexts? Therefore, even the relatively minor harm to data quality observed here suggests that survey designers need to consider aesthetics when designing web surveys.

References

Bernard, M., Chaparro, B. S., & Thomasson, R. (2000). *Finding information on the Web: Does the amount of whitespace really matter?* Usability News 2.2 [Online]. Retrieved from /usabilitynews/2W/whitespace.htm.

Bernard, M., Liao, C., & Mills, M. (2001). *Determining the best online font for older adults.* Usability News 3.1 [Online]. Retrieved from../usabilitynews/3W/fontSR.htm.

Bornstein, R. F. (1992). Subliminal mere exposure effects. In R. F. Bornstein & T. S. Pittman (Eds.), *Perception without awareness: Cognitive, clinical, and social perspectives* (p. 191-255). London Press: The Guilford Press.

Bowker, D., & Dillman, D. A. (2001). *An Experimental Evaluation of Left and Right Oriented Screens for Web Questionnaires.*

Table 14: Summary of Predicted Data Quality Impacts and Experimental Design Effects (odds ratios, χ^2, and t values with associated probabilities) for all Data Quality Indicators

Avenues of Potential Impact	Predicted Experimental Data Quality Effects	Experimental Design Effect Size	p value
Respondent Cooperation	Increased Non-Response		
	Reduced Completion Rate	0.44	.51
	More Early Terminations	1.40	.22
	Greater Likelihood of Skipping Items	1.443*	*.02*
Substantive Responses	Reduced Likelihood of Favorable Opinion Ratings		
	Quality of Education	0.746	.14
	Quality of Advising	0.824	.22
	Quality of Classes	0.681*	*.02*
	Desirability of Pullman	0.737*	*.05*
	Reduced Likelihood of Completing Open-Ended Questions		
	List-Style- Used Text Box #1	0.835	.38
	List-Style- Used Text Box #2	0.805	.17
	List-Style- Used Text Box #3	0.711*	*.03*
	Optional- Provided Comments	0.991	.96
	Reduced Depth of Open-Ended Responses		
	List-Style- Fewer Boxes Used	1.840*	*.03*
	List-Style- Fewer Business Provided	0.858	.19
	List-Style- Less Time Responding	2.582***	*.00*
	Optional- Provided >1 Theme	0.977	.92
	Optional- Provided Non-Visual Comments	0.100*	*.03*
	Negative Tone on Open-Ended Comments		
	More Negative Visual Design Comments	9.380***	*.00*
	More Negative Comments (All Themes)	2.326**	*.01*
	Fewer Positive Comments	0.546	.14
Respondent Commitment	Less Commitment Provided (More Satisficing/Primacy)		
	More Likely to Select Top Categories	1.890*	*.03*
	Less Likely to Select Bottom Categories	-1.680*	*.04*
	More Commitment Required (Greater Burden)		
	More Time per Screen, Overall	-3.840***	*.00*
	More Time per Screen, Early Q's	3.610***	*.00*
	More Time per Screen, Check-all Q's	3.290***	*.00*
Subjective Survey Experiences	Reduced Likelihood of Rating Survey Most Desirably		
	Overall Satisfaction	0.907	.53
	Perceived Difficulty	1.650	.08
	Perceived Focus	0.854	.34
	Visual Appeal	0.267***	*.00*

2000 Proceedings of American Association for Public Opinion Research, American Statistical Association, Alexandria, VA.

Brady, L., & Phillips, C. (2003). *Aesthetics and usability: a look at colour and balance.* Usability News 5.2 [Online] retrieved from usabilitynews/51/aesthetics.htm.

Chaparro, B. S., Shaikh, A. D., & Baker, J. R. (2005). *Reading Online Text with a Poor Layout: Is Performance Worse?* Usability News 7.1 [Online] retrieved from usabilitynews/71/page_setting.htm.

Christian, L., & Dillman, D. A. (2004). The Influence of Symbolic and Graphical Language Manipulations on Answers to Paper Self-Administered Questionnaires. *Public Opinion Quarterly*, *68*(1), 57-80.

Coates, D. (2004, August). *Online Surveys: Does One Size Fit All?* Paper presented at RC33 6th International Conference on Social Methodology: Recent Developments and Applications in Social Research Methodology.

Couper, M. P. (2008). *Designing Effective Web Surveys*. New York: Cambridge University Press.

Couper, M. P., & Miller, P. V. (2008). Web Survey Methods Introduction. *Public Opinion Quarterly*, *72*(5), 831-835.

Couper, M. P., Traugott, M., & Lamias, M. (2001). Web Survey Design and Administration. *Public Opinion Quarterly*, *65*, 230-254.

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P. J. (2000). *Brain potentials in affective picture processing: covariation with autonomic arousal and affective report.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Montréal, Quebec.

Damasio, A. R. (2000). A second chance for emotion. In R. D. Lane & L. Nadel (Eds.), *Cognitive Neuroscience of Emotions.* New York: Oxford University Press.

Dillman, D. A. (2007). *Mail and Internet Surveys: The Tailored Design* (2nd ed.). New York: John Wiley.

Dillman, D. A., & Christian, L. M. (2005). Survey Mode as a Source of Instability Across Surveys. *Field Methods*, *17*(1), 30-52.

Dillman, D. A., Gertseva, A., & Mahon-Haft, T. (2005). Achieving

Usability in Establishment Surveys Through the Application of Visual Design Principles. *Journal of Official Statistics*, *21*(2), 183-214.

Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for the Construction of Web Questionnaires.* Technical Report 98-50 of the Social and Economic Sciences Research Center, Washington State University, Pullman, Washington.

Draper, S. W. (1999). Analysing fun as a candidate software requirement. *PersonalTechnology*, *3*, 1-6.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, *69*(3), 370-392.

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, *22*(2), 313-328.

Goleman, D. (1996). *Emotional Intellegence: Why It Can Matter More Than IQ.* London: Bloomsbury.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and illustration. *Public Opinion Quarterly*, *64*, 299-308.

Hassenzahl, M., Beu, A., & Burmeister, M. (2001). Engineering joy. *IEEE Software*, *18*(1), 70-76.

Heerwegh, D. (2003). Explaining Response Latency and Changing Answers using Client Side Paradata from a Web Survey. *Social Science Computer Review*, *21*, 360-373.

Jennings, M. (2000). *Theory and models for creating engaging and immersive ecommerce Websites.* Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research (New York: ACM Press), pp. 77-85.

Knutson, J. F. (1998). *The Effect of the User Interface Design on Adoption of New Technology.* Dissertation Abstracts International: Section B: The Science and Engineering 59(3-B), 1399.

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., & Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement. *Public Opinion Quarterly*, *51*, 201-219.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in Surveys: Initial Evidence. *New Directions for Program Evaluation*, *70*(2), 29-44.

Kurosu, M., & Kashimara, K. (1995). *Apparent usability vs. inherent usability.* In Proceedings of the CHI 1995 conference companion on human factors in computing systems (pp. 292-293). New York: ACM.

Lauer, D. A., & Pentak, S. (2002). *Chapter 5: balance. Design Basics*. Australia: Wadsworth.

Ledoux, J. (1994). Cognitive-emotional interactions in the brain. In P. Ekman & R. J. Davidson (Eds.), *The Nature of Emotion* (p. 73-118). Oxford: Oxford University Press.

Ledoux, J. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* New York: Simon & Schuster.

Lindgaard, G. (1999). *Does emotional appeal determine perceived usability of web sites?* Hawthorne: University of Technology, School of Information Technology.

Lindgaard, G. (2007). Aesthetics, Visual Appeal, Usability and User Satisfaction: What Do the User's Eyes Tell the User's Brain? *Australian Journal of Emerging Technologies and Society*, *5*(1), 1-14.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers*, *15*(3), 429-452.

Lindgaard, G., Dudek, G., Fernandes, G., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology*, *25*, 115-126.

Mahlke, S. (2006). *Studying user experience with digital audio players.* Submitted to ICEC2006, 5th International Conference on Entertainment Computing, 20-22 September, Cambridge, UK.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Quarterly journal of experimental psychology. *Confirmation bias in a simulated research environment: An experimental study of scientific inference*, *29*, 85-95.

Niedenthal, P. H., Halberstadt, J. B., & Innes-Ker, A. H. (1999). Emotional response categorization. *Psychological Review*, *106*, 337-361.

Norman, A. D. (2004). *Emotional Design: Why We Love (Or Hate) Everyday Things.* Cambridge, MA: Basic Books.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2004). *The Effect of Preference Fluency on Consumer Decision Making.* Working paper, Yale University Graduate School of Business.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: Bradford Books/MIT Press.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web Survey Design: Paging Versus Scrolling. *Public Opinion Quarterly*, *70*(4), 596-607.

Pope, D., & Baker, R. (2005). *Experiments in Color for Web-Based Surveys.* Paper presented at 2005 FedCASIC Conference, Washington, D.C.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338-342.

Singer, E. (2006). Introduction: nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 637-645.

Sleep, D., & Pulleston, J. (2008). *Measuring the value of respondent engagement: Summary of research findings*. London: Engage Research.

Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1997). *Web Site Usability: A Designer's Guide, User Interface Engineering*. North Andover: MA.

Stern, M. J., & Dillman, D. A. (2006). Community Participation, Social Ties and Use of the Internet. *City and Community*, *5*(4), 409-424.

Toepoel, V. (2008). *A Closer Look at Web Questionnaire Design.* Ph.D. dissertation. Universiteit van Tilburg, Netherlands.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: interpretive heuristcs for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368-393.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*(1), 91-111.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*, 127-145.

Tullis, T. S. (1984). *Predicting the Usability of Alphanumeric Displays.* Ph.D. dissertation, Rice University, Texas.

van der Heijden, H. (2003). Factors influencing the usage of websites: The case of a generic portal in the Netherlands. *Information and Management*, *40*, 541-549.

Weller, D. (2004). *The Effects of Contrast and Density on Visual Web Search.* Usability News 6.2 [Online]. Retrieved from http://psychology.wichita.edu/surl/usabilitynews/62/density.htm.

White, J. V. (1974). *Editing by design.* New York: Bowker Company.

Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151-175.