Do Looks Matter? The Effect Of Survey Design On Data Quality In Academic Studies On Amazon Mechanical Turk

ELIZABETH G. KAUFER
ST. ANNE'S COLLEGE
WORD COUNT: 11,740

Thesis submitted in partial fulfillment of the requirement for the degree of MSc in Social Science of the Internet at the Oxford Internet Institute at the University of Oxford.

Submitted July 2015.

Abstract

Motivation: Online surveys and participant recruitment through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) have become increasingly popular among academic researchers. While online survey design has been extensively studied, there is little research in the context of crowdsourcing platforms; furthermore, there are certain visual design features such as using the researcher's logo or using a background that thus far have been mostly ignored by online survey methodology literature.

Objective: To examine if the researcher's logo or survey design template affect respondents.

Methods: Online experiment manipulating two visual design factors for online surveys: 1.) using a logo or not using a logo, and 2.) using a dark background or not using a dark background. Participants ($N = 448$) were recruited via MTurk and completed one of the experimental condition surveys on Qualtrics, an online survey design and hosting platform.

Main results: Provides evidence for the effect of the survey background on social desirability bias and emotional effect, while using the researcher's logo had no effect. Using a logo and no background led to significantly lower drop-outs compared to the three other experiment conditions. Demonstrates the high quality of data that can be obtained on MTurk.

Main contribution: Expanded framework for evaluating online survey design factors and effects on respondents. Online survey design recommendations for specific visual design factors, specifically, background and researcher logo. Further evidence for the viability of conducting research and experiments, particularly research on online survey design, on MTurk.

Conclusion: This study demonstrates both the feasibility of conducting research on MTurk and what visual design factors of online surveys may affect the way participants respond. It emphasizes the importance of online survey design and the need for researchers to openly discuss their visual design choices in publication, as these choices may have had an effect on the way participants responded to their surveys.

Acknowledgments

I would like to thank all the staff and faculty at the OII, particularly my supervisor, Dr. Ralph Schroeder, for his ongoing advice and encouragement, and to Dr. Andy Przybylski and Dr. Grant Blank for helping me pretend to be a scientist.

Thank you to the 2014-2015 Msc cohort: I can't imagine a more excellent group of people to learn about the Internet with.

Thank you to everyone who read drafts of my thesis, your feedback was invaluable.

Table Of Contents

List Of Tables And Figures

Chapter 1
Introduction

Besides providing new subjects to research, the Internet has opened up novel methods of conducting research. Now, researchers can recruit participants and run studies through email, forums, and, more recently, crowdsourcing sites and online labor markets, instead of just in person or by mail or telephone. Previously, participants for online research could be recruited via email, forums, or newsgroups (Van Selm & Jankowski, 2006), however, in recent years more academic researchers are using crowdsourcing platforms to find participants and conduct studies as a convenient way to obtain good data (Snow, O'Connor, Jurafsky, & Ng, 2008). For example, Benoit et al. (2015) compared the coding of political texts by experts to non-experts recruited on the crowdsourcing site, Crowdflower, and found the crowdsourced results to be as good as the experts'. They determined crowdsourcing to be a flexible, fast, and relatively inexpensive method for collecting high-quality data (Benoit et al., 2015). Research via crowdsourcing is used across many industries and disciplines, from market research to science: crowdsourcing has proven to be an invaluable tool for large scientific projects, such as Zooniverse, where "citizen scientists" contribute directly to original research projects (Christian et al., 2012). Using crowdsourcing platforms, researchers can recruit a greater number of participants faster and more cheaply than other methods while still obtaining

reliable data. Thus there is a great deal of interest in these platforms, which include sites such as CrowdFlower, Upwork (formerly oDesk), Microworkers, and, perhaps most prominently, Amazon Mechanical Turk.

Amazon Mechanical Turk (MTurk) is one of the most popular crowdsourcing sites for conducting online academic research, particularly online academic surveys. Originally launched in 2005, the original purpose of MTurk was to have workers complete small tasks, called Human Intelligence Tasks or HITs, that computers were not capable of handling and required human judgement. Since then, MTurk has been adopted by academic researchers as a platform for recruiting and paying participants for online experiments and studies. MTurk allows researchers to create or post links to online surveys as if they were any other type of HIT. Workers who are registered on the site can browse or search through lists of HITs and select ones they are interested in. With over 500,000 workers registered on the site (Amazon Web Services, 2015), recruiting for academic studies often goes very quickly. All HITs on MTurk must have a monetary reward, but the cost per participant is a small fraction compared to in-person participant pools (Berinsky, Huber, & Lenz, 2012). Paying participants is very straightforward: a credit card can be used to fund a single study, with researchers using Amazon's existing account and payment infrastructure to approve rewards for individual participants.

Disciplines using MTurk range from psychology to economics to health sciences. There do not appear to be any formal studies on the number of academic studies that are published about MTurk, so it is difficult to say how

many studies have used MTurk to collect data. A search for “Amazon Mechanical Turk” on Google Scholar brings up over 10,000 results for scholarly articles from the last ten years; while this surely includes articles that may only mention MTurk, it still demonstrates how pervasive the platform is across academia.

In combination with user-friendly survey software such as Qualtrics or SurveyMonkey, gathering survey data online has become easier to manage. However, these newer methods of research have also introduced new issues researchers must consider when designing their studies, including issues regarding data quality. As more online survey studies use participants on MTurk, an extensive body of literature has emerged that explores the quality and validity of data that can be obtained from these participants. However, there has been little investigation into the connection between data from crowdsourced participants and the design of online surveys, especially relating to the factor of visual design.

Survey methodology has shown that visual design choices matter: the way a survey looks can have an effect on respondents and therefore the quality of responses. Creating surveys online turns all researchers into visual designers. While many of the design choices are relatively simple to enact with easy-to-use software, they still must choose between many available options. As previous studies have shown, some of these choices may make the survey more or less visually appealing, or make it more difficult for respondents to read and interpret, while other design choices have little to no impact. For example, the readability of a web page depends on the combination of text and background

colors, with black text on a white background usually considered the most readable combination (Hall & Hanna, 2004). Researchers should understand what visual design choices may have an effect on data. Otherwise, this can lead to measurement error or respondent confusion. Further, while there is a wide body of research on online survey methodology, its findings and recommendations should be revisited as survey software and deployment methods change and new recruiting methods, such as crowdsourcing, become popular among academics.

The following study aims to combine research into data quality on MTurk and online survey design, in an attempt to understand how the two are related. Using the experimental method, this study examines the effect of visual survey design on data quality for online academic surveys posted on MTurk. Based on the existing practices of researchers using MTurk, this study investigates the possible effects of two visual design factors of online surveys: the researcher's logo and a dark colored background. Thus, this research covers the following key issues: effects of visual online survey design, an expanded framework for evaluating data quality, and a continued investigation into the viability of using MTurk for academic research. Based on the results of the experiment, its primary contribution is to provide guidance on the visual design of academic surveys for MTurk and how certain features may affect respondents.

Chapter 2
Research Context

2.1 Literature Review

In keeping with the stated aims of this study, the literature review will cover two areas of interest: 1.) research on online survey design and what design elements affect response quality, summarized in Table 1, and 2.) research on MTurk as a participant pool and discussions regarding the quality of data that can be obtained from those participants.

2.1.1 Online Survey Design

Best practices for academic surveys have been extensively studied, including issues particular to online surveys now that it has become much easier and more common for researchers to use online surveys to gather data (Shropshire, Hawdon, & Witte, 2009). Online tools and software offer researchers more options for the structural and visual design of their surveys than paper or phone surveys provide, and the effects of many of these options on data quality have been widely explored in the literature. Ganassali (2008) created a conceptual framework for investigating the connection between survey design features and data quality, which will be further discussed in Chapter Three as part of the experiment design. Design choices are particularly meaningful for online surveys, as the design of the survey is one of the main

ways participants interact with the researcher. As Ganassali (2008) observes, "[I]t is known that within self-administrated surveys, in the absence of an interviewer, the respondent tends to seek information from the instrument itself: the verbal and visual elements of the questionnaire (Schwarz 1996)" (p. 22). The visual design of the survey becomes a key method for understanding its context and establishing its legitimacy with the participant. For online surveys, a larger range of visual design choices are available to the researcher when designing a survey, and are in turn more closely observed by participant. Vehovar and Manfreda (2008) note, "Human-computer interaction research also shows that, compared to paper-and-pencil modes, individuals pay less attention to the text on web pages and more attention to graphical elements (Spool et al., 1999)" (p. 183). Badly designed surveys may stymie participants as they try to interact with the study. Participants notice poor design choices and may become frustrated by them (Sue & Ritter, 2012); furthermore, the design of a survey can affect the results and lead to measurement error if these effects are not anticipated or understood (Dillman, Smyth, & Christian, 2014).

Although there are many studies that examine online survey design elements, the effects of certain design choices are usually not addressed in actual practice. Since researchers generally do not discuss the design of their online surveys, the reasoning behind certain visual design choices is often not explained and the possible effects on data are rarely examined. For example, Kapelner and Chandler's 2010 study on MTurk used an online survey with a black background and white text, which they noted that many participants found difficult to read

or annoying. They acknowledged it may have affected their results; however, Kapelner and Chandler did not discuss why they had chosen to design the survey this way. While this may seem trivial to the researchers, it is clear from the literature that even minor design choices such as background and figure colors can have an effect on respondents. As Oppenheimer writes in his 2008 paper on processing fluency, "[S]eemingly trivial decisions that researchers make when designing their studies can have nontrivial influences on their results" (p. 240).

The research regarding online survey design tends to focus on structural elements such as question and response formats, survey length, interactive features, or recruiting methods (Dillman, Smyth, & Christian, 2014; Fan & Yan, 2010; Funke, 2015; Mohr, Sell, & Lindsay, 2015; Srinilta et al., 2012; Sue & Ritter, 2012). For instance, the way questions are worded and ordered has been shown to have an effect on responses, and technical flaws in the survey's design may lead to higher drop-out rates (Fan & Yan, 2010). Dillman, Smyth, and Christian (2014) go into great detail describing the myriad design choices researchers must make when designing online surveys and give specific guidelines for good design choices: they recommend standardizing the visual presentation of the survey, avoiding visual disarray by reducing the amount of information presented on a single page, and using color and contrast to create visual appeal. However, they do not provide guidelines on using images or logos in the survey design. Additionally, Joinson and Reips (2007) confirmed that the power or status of the researcher, using indicators such as his or her title and

institution, when used in conjunction with a personalized salutation, has a positive effect on the response rate for email survey recruitment.

When it comes to visual imagery, the existing research mainly examines the effect of illustrative, supplemental images or image-type questions (such as brand recognition), as opposed to secondary elements like background color or header images. Sue and Ritter (2012) discuss the potential benefits of using supplemental images or graphs to provide context for survey questions. Other advice states that images may be used to help motivate respondents or give them a more pleasing experience while taking a survey (Vicente & Reis, 2010). Couper (2008) breaks images into three separate functions: images that are the question, images that are supplemental to the question, and images that are incidental to the question. Couper states that images located in the header of the survey are likely not meant to be a part of the question content and are rather used for branding or identification purposes to help contextualize survey (2008). However, there has been little further research on the possible effects of header or branding images such as a logo on respondents.

Much of the literature on survey methodology cautions against making an online survey too visually complex as it can affect the rendering of the survey on different browsers or different Internet transmission speeds, which may cause certain visual elements to not load correctly on some respondents' browsers; in other cases, the long loading time may cause them to abandon the survey (Couper, 2008; Fan & Yan, 2010; Sue & Ritter, 2012; Vicente & Reis, 2010). Studies on the aesthetic design of surveys have shown that complex or

distracting images and clashing colors have negative effects on response quality; moreover, high visual appeal has a positive effect on user enjoyment and perceived usability for other types of interfaces (Mahon-Haft & Dillman, 2010). Mahon-Haft and Dillman (2010) found that a displeasing aesthetic design led to negative emotional effects on survey respondents and impacted their responses. Their experiment found that the unappealing design led to a reduction in response effort by participants and affected the answers they received. They write:

> 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 (Mahon-Haft & Dillon, 2010, p. 43).

While their experiment found minor effects to the data quality obtained, they note that aesthetic design does have an impact on responses, and for longer or more complicated surveys this effect could increase.

Aesthetic design appears to be closely related to processing fluency, or, the ease in which a user is able to comprehend a task: visually pleasing surveys seem to be easier to read. Aesthetic features such as simplicity, symmetry, balance, clarity and brightness contrast appear to ease the viewer's processing of an object (Winkielman, Schwarz, Fazendeiro & Reber, 2007). Visual design elements, such as figure-ground contrast or the readability of a particular font, effect processing fluency because they change the way information is presented and therefore influences a user's judgment of an interface (Oppenheimer, 2008). Hall and Hanna (2004) describe the importance of design on user behavior: "Web

design, like most design endeavours is a balance between the functional and aesthetic. Factors such as aesthetically pleasing colour combinations can play an important role in generating positive affect… Aesthetic factors may serve to affect behavioural intention…" (p. 185). For example, using color to enhance the appearance of the survey and its appeal may help motivate a respondent, although strong hues should be used moderately and may appear differently on different computer screens or browsers (Sue & Ritter, 2012). Best practice when utilizing color suggests maintaining the readability of the survey through high text and background contrast (Sue & Ritter, 2012). The general advice in survey methodology is to keep survey designs simple and uncomplicated (Dillman, Smyth, & Christian, 2014; Wang & Doong, 2010). However, even simple designs have a range of choices that may have an effect on respondents, and these have not been sufficiently explored in the existing literature. Changing the color of the survey background is a simple change, but whether this could have an effect has not yet been tested.

However, the literature on survey design sometimes gives contradictory advice. For instance, while Lauer, McLeod, and Blythe (2013) recommend using a progress indicator, Dillman, Smyth, and Christian (2014) oppose it. This seems to be caused by a lack of standardization across research. As Vicente and Reis (2010) observe, it is difficult to compare the effects of survey design on response quality across the literature because much of the empirical research uses different sample populations and different measurements, and often look at factors that do not have a standard definition (for example, how to measure the

length of a survey). Additionally, it is not always clear how different design elements interact with each other. Furthermore, some of the literature may refer to research conducted on older technology or, based on the date of the research paper, on a population sample that was likely less familiar with web interfaces than a sample population on a crowdsourcing site. As Sue and Ritter point out in their guide for creating online surveys (2012), changes in software and the general familiarity of respondents with online surveys requires that survey design guidelines be reviewed and updated to keep pace with these developments.

Therefore, this research will revisit previous survey research and expand on two specific visual design choices with a new population, workers on the crowdsourcing site and online labor market, MTurk.

Table 1. Summary of research on visual design factors and potential effects on respondents.

Design Factor	Potential Effect(s) on Respondent	Research
Visual harmony (e.g. color, contrast, simplicity, clarity)	Visually appealing surveys have a positive emotional effect, appear easier to read and process; confusing or unaesthetic designs appear to have a negative effect on processing fluency and response effort by participants.	Dillman, Smyth, & Christian (2014); Hall & Hanna (2004); Mahon-Haft & Dillon (2010); Oppenheimer (2008); Winkielman et al. (2007)
Visual imagery	Supplemental images may motivate participants or create visual appeal; however too much visual complexity may have a negative effect on respondents.	Couper (2008); Fan & Yan (2010); Sue & Ritter (2012); Vicente & Reis (2010)
Status of researcher	Positive effect on response rate; however research limited to email survey recruitment.	Joinson & Reips (2007)

2.1.2 Data Quality on Amazon Mechanical Turk

Although it is a relatively new recruitment method, there has been a fair amount of research into the feasibility of conducting studies on MTurk, including issues of internal and external validity, and the quality of data obtained from MTurk participants. MTurk has become popular among academic researchers as a participant pool due to its low cost and fast recruiting. As such there is ongoing investigation into how to obtain the highest quality data from MTurk participants.

Understanding the demographics of MTurk is one key area of research. Compared to college samples, samples from MTurk tend to be more diverse and can include workers from outside the U.S.; however, MTurk workers tend to be younger, more educated, and less extraverted than nationally representative U.S. samples (Paolacci & Chandler, 2014). They are also much more likely to use social media than a representative U.S. sample (Kang et al., 2014). Despite these differences, researchers have found that MTurk is appropriate for conducting different types of research, including experiments. Berinsky, Huber, and Lenz (2012) replicated three classic experiments on MTurk and obtained results similar to laboratory results; Paolacci, Chandler, and Ipeirotis (2010) compared experimental results from three different recruitment methods, including MTurk, and concluded that MTurk was a reliable source for running online experiments.

Other researchers have examined what motivates workers on MTurk and how these motivations may be activated when designing an MTurk study.

Despite the relatively low payments (Kaufman, Schulze, and Veit's 2011 study found the median reservation wage, i.e. the lowest amount which a worker would be willing to work for, to be $1.38/hour), workers still appear to be motivated by monetary rewards. Participants, it seems, assess the fairness of the reward in comparison to the difficulty of the task. Some workers, for example, work to earn a specific payment target (Horton & Chilton, 2010). They remain, however, sensitive to pay cuts (Chen & Horton, 2010).

Besides extrinsic motivation, several studies have found that, while monetary payments are an effective motivator for workers on MTurk, intrinsic motivations such as the type of task or the way the task is framed affects output quality. Mason and Watts (2010) found that increasing the monetary reward does not necessarily lead to an increase in work quality. Both Chandler and Kapelner (2012) and Shaw, Horton, and Chen (2011) observed that the way a task on MTurk is framed affects the output. Chandler and Kapelner (2012) saw that framing a task as meaningful (e.g. as helping cancer researchers) produced a higher quantity of work, while tasks that had a low meaning (e.g. working for a corporation) saw a decrease in quality. Thus, workers on MTurk are not solely motivated by monetary factors when deciding how much effort to put into a task.

Research on data quality also looks into ways to structure surveys in order to weed out inattentive respondents by using attention checks. Oppenheimer, Meyvis, and Davidenko (2009) validated a type of attention check called an Instructional Manipulation Check (IMC), which provides evidence to see if survey participants are reading directions or not. While the original paper did

not use MTurk as a participant pool, subsequent studies on MTurk have used the IMC in their survey designs. However, recent research suggests that using IMC questions may prompt participants to pay closer attention to the study in order to avoid being tricked by similar types of questions, and that therefore IMC questions actually change participants' attention rather than just measuring it (Hauser & Schwarz, 2015). Peer, Vosgerau, and Acquisti (2013) conducted a study on using workers' MTurk reputation rather than attention check questions (ACQs) as a way to ensure high data quality. They used several measures of data quality, including failing ACQs such as the IMC and scoring high on a social desirability scale as markers of low data quality. They found that, compared to low reputation workers, the high reputation workers had the lowest social desirability scores and were less prone to fail ACQs; they therefore concluded that using high reputation workers on MTurk was a valid way to obtain high quality data.

However, there has been very little research regarding online survey design and its possible effects on MTurk data quality. A recent study by Mohr, Sell, and Lindsay (2015) on MTurk found that changing the number or size of text response boxes changed the type and elaboration of the responses. Previous research has established that MTurk provides a viable field for academic research and workers react in ways similar to other research samples (Amir & Rand, 2012), so we may expect that MTurk workers will be affected by survey design in a similar way as established in the literature on online survey design. For instance, a study by Komarov et al. (2013) showed that MTurk participants

evaluated GUIs (Graphical User Interfaces) the same way as participants in a laboratory setting. Studies on visual perception on MTurk have also replicated lab effects (Bartneck et al., 2015). However, this does not address online questionnaire design specifically. As previously stated, it is important to revisit this research periodically as technology changes. Additionally, despite surveys being very popular on MTurk, it is a new, relatively untested field for online survey design research.

2.2 Research Questions

Aesthetic design is difficult to study since it is very subjective. Instead, the research design is reframed to look at visual design choices made by academics when publishing surveys or questionnaires. In order to determine which choices to study, academic studies posted on MTurk were systematically reviewed. During a one-week period, screenshots of academic studies posted on MTurk were collected and examined. Any HIT that contained the keywords “survey”, “study”, “experiment”, or “research” and that the researcher’s MTurk worker account qualified for was reviewed; if it was clear based on the description or information sheet that it was academic research, a screenshot was taken of the information sheet or, if this was not included, the first page of the survey. A total of 192 cases were collected and design elements such as font, color, use of images, text alignment, and display polarity were examined. From these elements, two variables, which were seen in at least a quarter of the cases, were

chosen for further study: using the logo of the researcher's institution (used in 29% of cases) and using a dark background (used in 27% of cases). Although these choices are common practice, they are not seen universally across all the surveys.

Based on these results, the research topic was refocused into the following questions, focusing on the two chosen variables:

RQ1: Does the use of the researcher's logo affect respondents?

RQ2: Does the design template of an online survey affect respondents?

Using a modification of Ganassali's 2008 framework (discussed in more detail in Chapter Three), the effect on respondents can be measured by looking at data quality and emotional response. The significance of each factor is described below.

2.2.1 Logo

In addition to serving as a design choice, the logo is a visual signal of the status or power of the researcher. Previous research on survey methodology has shown that a logo fosters trust in respondents and reminds them of the institution's reputation. In Srinilta et al.'s qualitative study of survey design (2012), they found a high preference for using a company's logo as part of the design, as the logo served as a reminder of the company's products or services and inspired trust in its reputation. A company logo is also perceived as a signal of legitimacy for marketing surveys (Wang & Doong, 2010). Mahon-Haft and Dillman's (2010) experiment indicated that surveys from a more legitimate

source appear to inspire more effort on behalf of respondents. It seems reasonable that the logo of an academic institution may have a similar legitimizing or trustworthy effect on respondents, and therefore affect data quality.

A logo may also have a positive effect on the response rate. Joinson and Reips (2007) speculated that personalized survey recruitment emails that came from a high status requester had a higher response because it inspired socially desirable behavior in the participants (i.e. responding to the survey). Joinson and Reips used text to indicate requester status, but a logo is similarly an indicator of status and power. Based on this, it seems that a logo, as a visual reminder of the requester's status, may influence participants to respond in a socially desirable manner. While a logo may prompt respondents to pay more attention and fill in more of the answers in the survey, it may also lead them to give more socially desirable responses. The effect of a logo on a respondent's social desirability score has not been addressed in the literature.

Therefore, the following hypotheses related to RQ1 are proposed:

H1-a: Having a logo on every page of the survey will lead to better response quality.

H1-b: Having a logo on every page of the survey will lead to higher social desirability score (lower data quality).

H1-c: Having a logo on every page of the survey will inspire trust and lead to a more positive emotional response.

2.2.2. Background

The contrast of the background to the foreground (the body of survey) may have an effect on respondents even though it is relatively unimportant to the content of the survey. Ganassali (2008) terms a visual element such as the background as incidental imagery; however, this does not mean participants would not notice it. Dillman, Smyth, and Christian (2014) describe the usefulness of using lightly colored backgrounds and black text in question boxes in order to draw participants' attention to the questions and response options. As described in the review of MTurk studies, dark backgrounds are not uncommon in academic online studies. When using popular online survey software such as Qualtrics, survey designers often have many templates to choose from when creating a survey, and several of these may include dark backgrounds. For example, of the 30 survey templates available on Qualtrics, eight have dark backgrounds as their default setting. In any case, changing the background color in one of the Qualtrics templates is very straightforward and does not require advanced coding skills, so even if a chosen template did not have a dark background, a researcher could change it with little effort.

A dark background may also help respondents focus their attention on the questions. Dillman, Smyth, and Christian (2014) describe the way participants take in and parse the page of an online survey. First, respondents process the basic page layout and take in visual design clues (such as colors, contrast, size, shape) and begin to understand what elements are important: by using visual boundaries to organize the page, the contrast of the two colors separates the

figure from the background and informs users what to focus on. The dark background condition may be perceived as more visually pleasing to participants since it demonstrates greater figure-ground contrast (Reber, Schwarz, & Winkielman, 2004) and an attractive color scheme (Hall & Hanna, 2004).

From this evidence, the following hypotheses relating to RQ2 are proposed:

H2-a: A dark background is more visually appealing and will lead to better response quality.

H2-b: A dark background is more visually appealing and will lead to more positive emotional response.

Chapter 3
Methodology

3.1 Conceptual Framework

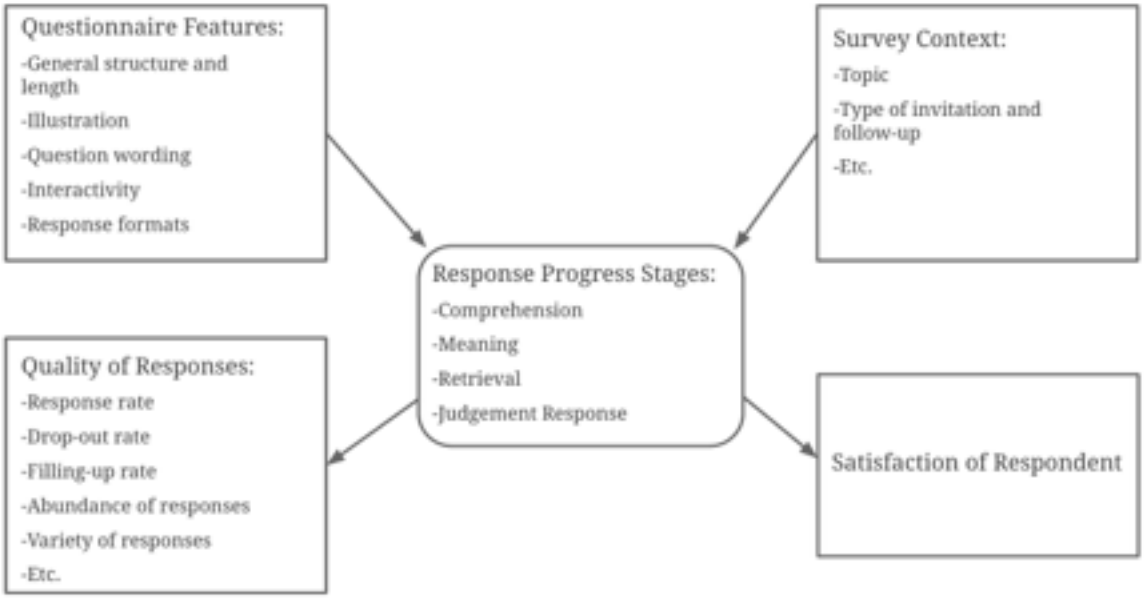

Figure 1. Conceptual framework for online questionnaire design (Ganassali, 2008).

Ganassali's 2008 conceptual framework on online questionnaire design (Figure 1) is used to structure the research design. It was chosen because it consolidates the literature on online survey methodology into a single framework and contributes a clear, standardized method for analyzing response quality. The framework breaks an online survey into these components: Questionnaire Features & Survey Context, which affect the respondents' progression through the survey and can be measured in the Quality of Responses and the Satisfaction of the respondent. Particularly useful for this research is the way in which data

quality is operationalized: response quality includes aspects of the data collected other than just the actual responses, namely the drop-out proportion, the variety of responses (do participants just select one or the default answer for all or many questions), and the filling-up proportion (how many questions are completely answered).

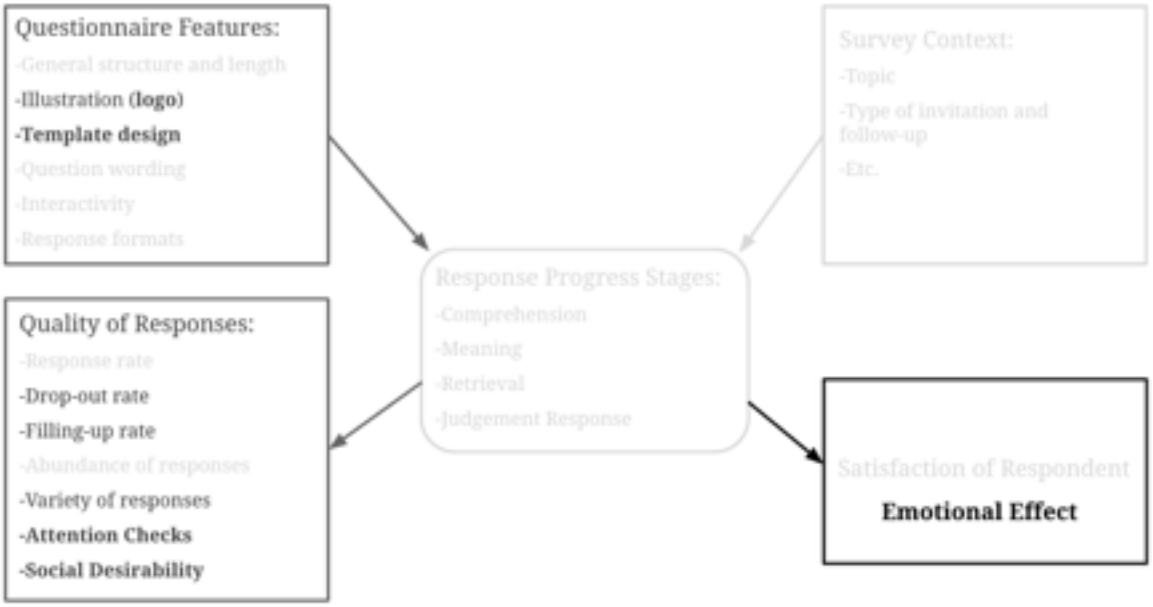

Figure 2. Modified framework (researcher's creation).

Figure 2 shows a modification of this conceptual model, which highlights the factors which will be analyzed in this research paper. It expands the existing framework to include additional factors for analysis (in bold) in Questionnaire Features and Quality of Responses, and replaces Satisfaction of Respondent with Emotional Effect.

The data quality component is modified based on an experiment carried out on MTurk by Peer et al. (2013), where the researchers used failure to heed

attention checks and a high score on a social desirability scale as indicators of low data quality. Adding their outcome measures gives us a better idea of the way respondents react to the survey, if they are paying attention, and if they modify their responses to give themselves a better impression to researchers.

The satisfaction of respondent, which in Ganassali's study was a single question asking the respondent to rate the quality of the survey, has been replaced with an emotional effect scale because research indicates there is a connection between emotion and design (Mahon-Haft & Dillman, 2010). The emotional affect scale replaces Ganassali's single satisfaction question in order to investigate the effect of the survey on respondents' emotional state and see if there is any notable difference between conditions.

3.2 Experiment Design

In order to detect if these two design features affect data quality, an experiment using an online survey was carried out on MTurk. An experimental approach was used in order to demonstrate if a relationship exists between these variables (Brewer & Crano, 2014). The two conditions are based off of actual practice by academics on MTurk, and the experiment is designed to look like a typical academic study on MTurk, so the experiment has strong ecological validity. The experiment used a 2x2 design with two conditions: logo (with or without) and background (dark background or no background). All other factors (such as content, interaction, display, usability, etc.) remain constant; therefore the experiment should have sufficient internal validity. Since the researcher is

situated at the University of Oxford, that institution's logo was used in the experiment; the logo was downloaded directly from the Oxford University logo pack for web (http://www.ox.ac.uk/public-affairs/branding-toolkit) and added as a header to the survey. For the dark background condition, the background of the Qualtrics survey template was changed from white to a dark shade of blue in order to match the Oxford University logo and present a harmonious color scheme. Qualtrics software was used to design, build, and host the survey, rather than using the built-in MTurk interface. Qualtrics was selected because it is a popular survey tool and used at many academic institutions, and allows for greater control over the survey design and permits multi-page surveys compared to MTurk's native survey software. The content and structure of the survey was the same across all conditions: only logo and background were altered between conditions. The four experiment conditions are included in Figure 3. Note that, while only a small portion of the background is visible in Figure 3, the background color fills the entire web page.

Figure 3. The four experiment conditions.

The experiment and results, however, may not be generalizable to the general U.S. population or perhaps even the Internet-using population. MTurk workers tend to be younger and more educated than the general U.S. population (Berinsky, Huber, & Lenz, 2012; Paolacci and Chandler, 2014), and this may influence how they respond to online interface design. Furthermore, MTurk workers appear to be extremely familiar with online questionnaires and academic surveys (Chandler et al., 2015; Rand et al., 2013). Many studies have established the benefits of using MTurk for academic research, however recently some researchers have raised some concerns about the possibility that the MTurk participant pool may be overused or possibly exploited; this may prove to be an issue with this type of experiment as well.

3.3 Outcome Measures

The questions on the survey and three of Ganassali's factors of response quality were used to measure the effects of the experiment conditions. The content of the survey is included in Appendix C.

3.3.1 Pew Social Media Omnibus Survey 2014

The first page of survey questions related to social media use, as this was expected to be of interest and relevance to MTurk workers. A total of 22 questions on social media were taken from the topline questionnaire from Pew Research Center's Social Media Omnibus Survey 2014 (Duggan et al., 2015) and asked about respondents' usage of specific social networking sites such as Facebook and Twitter. This measure was also used as a manipulation check, since it asked about nonsensitive user behavior and therefore should not be influenced by the experimental factors.

3.3.2 Social Desirability Scale

Reynolds' (1982) Short Form C of the Marlowe-Crowne Social Desirability Scale was used, which is a 13-item version of the Marlowe-Crowne scale scored from 0-13 ($\alpha = 0.80$). As in Peer et al.'s study (2014), a high social desirability score was considered an indication of low data quality. A high social desirability score indicates strong social desirability bias, in which respondents change their answers to make their behavior or feelings appear more favorably and present a

better impression to the researchers. Social desirability bias can negatively affect response quality, especially for surveys that ask about sensitive information or behavior, since respondents may not report truthfully.

3.3.3 Emotional Affect Scale

This study utilized the PANAS-X (Positive and Negative Affect Scale - Expanded) scale as it should capture a more nuanced picture of the respondent's emotional state than Ganassali's single question on satisfaction. The scale consists of 60 emotion words and asks participants to rate how aptly each word describes his or her emotional state at that moment (Watson & Clark, 1999). Participants selected from a five-point scale, from 1 "Very slightly or not at all" to 5 "Extremely." The PANAS-X is broken into 13 scales for analysis (see Appendix B). Each was found to be highly reliable in post-collection analysis: Positive Affect (10 items, $\alpha = 0.93$); Negative Affect (10 items, $\alpha = 0.94$); Fear (6 items, $\alpha = 0.92$); Hostility (6 items, $\alpha = 0.92$); Guilt (6 items, $\alpha = 0.93$); Sadness (5 items, $\alpha = 0.89$); Joviality (8 items, $\alpha = 0.95$); Self-Assurance (6 items, $\alpha = 0.88$); Attentiveness (4 items, $\alpha = 0.82$); Shyness (4 items, $\alpha = 0.856$); Fatigue (4 items $\alpha = 0.90$); Serenity (3 items, $\alpha = 0.87$); and Surprise (3 items, $\alpha = 0.83$). As previous research shows, emotions and design appear to be linked; therefore, the PANAS-X scale is used to show if either design factor has an influence on the respondent's emotional state.

3.3.4 Attention Check Questions

Three attention check questions (ACQs), modeled on Peer et al. (2014), were utilized. ACQs are a common feature of surveys, designed to weed out respondents who are not paying attention to the questions or seem to be answering randomly. The first was a fake social networking site called ChatFace inserted into a matrix question about how frequently the respondent uses common social networking sites. Any answer other than "Don't Know" or "Don't Use" was coded as failing the attention check. The second attention check was the last question of the demographic section, which was embedded in a question about how many hours a week the respondent spends on MTurk; it instructed the respondent to ignore the question and proceed with the survey. Any answer to this question was coded as failing. The final attention check was a slight modification of the Instructional Manipulation Check (IMC) used by Peer et al. (2014), as the last question of the survey. The question was a paragraph of text that contained directions on how to answer the final question, "What was this survey about?" The IMC told participants to select "Other" and type "Internet" into the text box. Any answer that did not follow the instructions was coded as failing. This question was inserted at the end of the survey in order to avoiding the possible effects IMC questions have on participant attention (Hauser & Schwarz, 2015b). Since participants who miss multiple ACQs are typically eliminated from respondent data, failing ACQs was also used as a measure of data quality.

3.3.5 Ganassali's Factors (Drop-Out, Filling-Up, Variety of Responses)

Three of Ganassali's factors of response quality were also investigated: drop-out proportion (respondents who fail to finish the survey), filling-up proportion (proportion of completed questions on the survey), and variety of responses (how varied or differentiated responses to scale questions are). These were chosen because they can be easily processed and analyzed using online survey software or popular data analysis programs like Excel. These are used as indicators of data quality by providing contextual information about how respondents interacted with the survey. Researchers want a low drop-out rate because it means they have a larger sample size; drop-out rates can be as high as 15-20% for online surveys (Ganassali, 2008). A high filling-up rate means that respondents have answered many or most of the questions on the survey, which gives researchers more data to work with. A high variety of responses for scale questions shows that respondents are answering questions thoughtfully and with real effort, whereas a low variety of responses indicates they are making the same answer choice for multiple questions (e.g. only choosing the middle option for scale questions).

3.3.6 Pilot Study

A plain version of the survey was piloted with members of the Oxford Internet Institute to check for content, spelling, or logical errors. The pilot survey received 16 responses. Based on participants' feedback, changes were

made to the wording of several of the questions on social media and attention checks. The final attention check (IMC) was also added to the final survey design after the pilot study.

3.4 Participants

A total of 465 MTurk workers participated in the survey. Respondents who answered none of the survey questions were removed, as well as those who took more than 20 minutes to complete the survey (N = 448; 41% male, 58% female; 37% 25-34 years old, 25% 35-44 years old; 79% White). A full breakdown of participant demographics is included in Appendix A. The participants were limited to U.S. workers on MTurk who had a 98% approval rating and had completed at least 500 HITs on MTurk. Since it is expected that experienced and high reputation workers will give high quality data (Peer et al., 2014), many academic studies on MTurk use high thresholds such as 98% or at least 100 HITs completed (Staffelbach et al., 2014). Therefore, for purposes of ecological validity, this experiment also used high reputation, high experience workers.

3.5 Ethics

The research design was reviewed by and received ethics clearance through the University of Oxford Central University Research Ethics Committee (approval number OII C1A 15-018). Workers on MTurk are anonymous: they are only identified by their ID numbers, so participants' privacy is ensured. Only

basic demographic information was requested in the survey, and these questions were all optional. There were no questions on sensitive issues. The first page of the survey was a detailed information sheet which contained information about the survey content, what could be expected, and the contact information of the researcher and the Ethics Committee. Before participants could proceed with the survey, they needed to check off a box confirming that they were 18 years of age and had read the information sheet. The final page of the survey contained a short debriefing section and repeated the contact information for the researcher.

3.6 Procedure

A HIT containing a link to the survey was posted to MTurk under the title "Social Media Use Survey" and described as an academic study. After participants accepted the HIT, they clicked on a link that redirected them to the third-party Qualtrics site. The link randomly assigned them to one of the four experiment conditions on Qualtrics. After completing the survey on Qualtrics, they received a code that they would input on MTurk to receive credit for taking the survey and be approved for payment. Participants received $0.25 for completing the survey, based on previous MTurk studies (Crump, McDonnell, & Gureckis, 2013; Goodman, Cryder, & Cheema, 2013; Mohr, Sell, & Lindsay, 2015) and the availability of researcher funds. They were paid within 36 hours of submitting their completion codes. No submissions were rejected for any reason (such as incompleteness) since response quality was the outcome measurement for the experiment. The HIT was posted for 14 days until at least 400

submissions had been collected on MTurk in order to obtain at least 100 observations in each cell. A statistical power analysis was performed in order to determine sample size for the experimental factors. The effect size was considered to be large using Cohen's *d* (0.80). With an alpha = 0.05 and power = 0.80, the projected sample size was approximately $N = 30$ for one factor. Therefore, the obtained sample size should be sufficient for this study.

Chapter 4
Results

Partial responses (at least 20% of survey completed) are included in the analysis as they are also a measurement of data quality. Of the included participants, 23 (5.13%) reported taking the survey on a tablet and 6 (1.34%) reported taking the survey on a mobile device. The mean completion time for all included participants was 5.8 minutes.

4.1 Manipulation Check

The Pew Internet survey questions on social media served as a manipulation check, as the experimental manipulations should not affect reported Internet usage. A two-way analysis of variance (ANOVA) was conducted for each item. ANOVA is used to detect differences between means of the dependent variables (outcome measures). The experimental factors showed no significant effect on reporting of the 22 questions on social media usage (p-values ranged from 0.087 to 0.992). Therefore, we can assume that differences between the experimental groups are due to the manipulations rather than some other factor.

4.2 Social Desirability Scale

The social desirability scale was scored according to the scale's instructions. The sum of each participant's responses was calculated to obtain their overall social desirability score (with 0 being the lowest possible score and 13 being the highest possible score). A two-way ANOVA was run ($N = 447$) to see if there were any effects of background or logo on participants' social desirability scores (see Table 2). No significant interaction was found for logo and background ($F(1, 443) = 0.72$, $p = 0.40$). There was also no significant difference in the logo condition ($F(1, 443) = 0.11$, $p = 0.74$). Thus, H1-b is not accepted: it appears that using a logo does not produce more socially desirable answers from participants.

Table 2. Two-way ANOVA for social desirability score for experiment condition.

	d.f.	**SS**	**MS**	***F***	***P***
Logo	1	1.31	1.31	0.11	0.74
Background	1	60.61	60.61	5.32	0.022*
Logo x Background (interaction)	1	8.16	8.16	0.72	0.39
Error (within)	443	5,046.31	11.39		
Total	446	5,116.56			

* Significant at the 0.05 probability level.
N = 447

However a significant relationship was found between the social desirability score and background ($F(1, 443)$, $p = 0.022$). Respondents in the dark background had a lower mean social desirability score ($M = 5.70$, $SD = 3.30$)

compared to those in the no background condition ($M = 6.44$, $SD = 3.43$). This gives partial support to H2-a.

Controlling for demographic variables (age, gender, education, location, and ethnicity) did not produce any significant results.

4.3 Attention Check Questions

For the three ACQs, 45.6% of respondents passed all three, 40.7% failed one, 11.9% failed two, and 1.8% failed all three ($N = 448$). Since these were categorical variables, cross-tabulations and chi-square tests were conducted for the ACQ measures.

The first ACQ asked participants to report how often they used the Internet to go on common social networking sites. A fake social networking site, ChatFace, was included in the list and those who reported using the site failed the ACQ; 7.1% of respondents failed and 92.9% passed ($N = 444$). There were no significant effects for logo ($X^2(3, N = 444) = 0.09$, $p = 0.75$), background ($X^2(1, N = 444) = 2.16$, $p = 0.14$), or experiment condition ($X^2(1, N = 444) = 2.34$, $p = 0.51$).

The second ACQ, located at the end of the demographic questions, required participants to ignore the question and continue with the survey; 44.9% failed and 55.1% passed ($N = 448$). A technical flaw with this ACQ made it so that if a respondent accidentally selected an answer, he or she could not unselect it (as the instructions required): nine participants emailed to notify the researcher of the flaw. This could account for why a much larger proportion of respondents failed this ACQ compared to the other two, although it is unclear if

more workers also had this problem as there is no way to discern who may have wanted to uncheck their selection. There were no significant effects for logo ($X^2(1, N = 448) = 0.00$, $p = 0.99$), background ($X^2(1, N = 448) = 0.34$, $p = 0.56$), or experimental condition ($X^2(1, N = 448) = 0.79$, $p = 0.85$).

The final ACQ was the last question in the survey. It was an Instructional Manipulation Check (IMC) which required participants to read a paragraph of text, which included instructions on how to correctly answer the question; 10.3% failed and 89.7% passed ($N = 414$). There were no significant effects for logo ($X^2(1, N = 414) = 0.11$, $p = 0.74$), background ($X^2(1, N = 414) = 0.11$, $p = 0.74$), or experiment condition ($X^2(3, N = 414) = 0.32$, $p = 0.97$). Overall, the experimental factors appear to have no significant effect on the respondents' performance on the ACQs.

4.4 Drop-Out

All observations were included in the analysis of drop-out ($N = 465$). Participants who reached the last page of the survey and submitted their responses were marked as completed, while those who began the survey but did not finish were marked as drop-outs. Each cell had close to the same number of participants, however the no background/logo condition had much fewer drop-outs compared to the other three. The number of completed responses is shown in Table 3.

Table 3. Total versus completed responses for each condition.

	Total Responses	Completed Responses
No Background/No Logo	115	101
Dark Background/No Logo	115	101
No Background/Logo	118	114
Dark Background/Logo	117	102

There were no significant results for logo ($X^2(1, N = 465) = 2.14$, $p = 0.15$) or background ($X^2(1, N = 465) = 2.92$, $p = 0.09$). A chi-square test on experiment condition was statistically significant ($X^2(3, N = 465) = 7.89$, $p = 0.048$): the experiment condition with no background and a logo had the lowest proportion of drop-outs (see Figure 4). This interaction gives partial support to H1-a and H2-a.

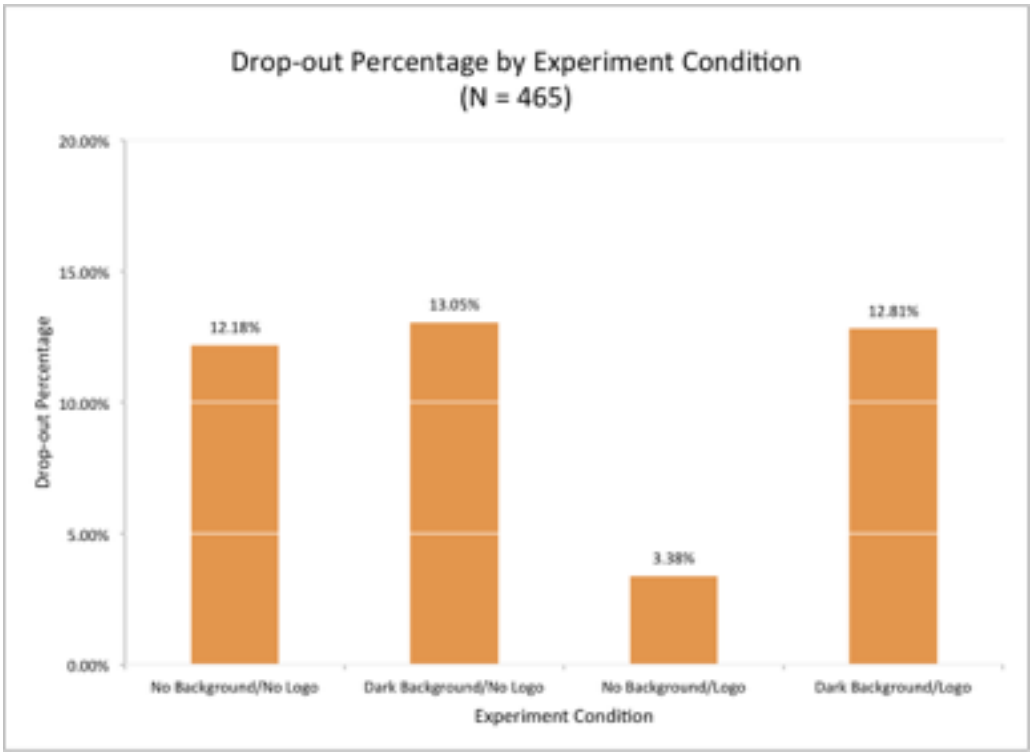

Figure 4. Percentage of drop-outs by experiment condition.

4.5 Filling-Up

In order to measure how many of the questions respondents chose to complete (i.e. filling-up proportion), none of the questions in the survey except the informed consent question were required. To measure the filling-up proportion, the sum of the number of items answered was taken for each respondent, with the lowest possible score being 0 (no questions answered) and the highest being 101 (every question answered). A two-way ANOVA ($N = 448$) showed no significant effects for logo ($F(1, 444) = 1.01$, $p = 0.32$), background ($F(1, 444) = 0.25$, $p = 0.62$), or interaction ($F(1, 444) = 1.42$, $p = 0.23$).

4.6 Variety of Responses

The PANAS-X question was also used to measure the variety of responses, because it is a matrix question with a five-point scale and 60 items. It was the longest question on the survey and therefore required more effort to answer. The standard deviation of responses was calculated for each respondent in order to gauge the range of responses per item. A standard deviation of 0 would indicate the respondent picked the same answer for every item, while a standard deviation of 1 and up was taken to indicate more response variety; the distribution for respondents' SD was normally distributed ($M = 1.10$, $SD = 0.35$). A two-way ANOVA ($N = 445$) showed no significant effects for logo ($F(1, 441) = 0.13$, $p = 0.72$), background ($F(1, 441) = 0.76$, $p = 0.38$), or interaction ($F(1, 441) = 0.73$, $p = 0.39$).

4.7 Emotional Effect

The PANAS-X breaks its 60 emotional items into several scales, which were each analyzed separately. A full breakdown of each scale and its relevant items is included in Appendix B. The mean of each respondent's answers to the scale items was calculated to obtain the score for each scale. Table 4 shows the means and standard deviation for each emotional scale.

Table 4. Means (standard deviation) for PANAS-X scales by logo and background.

	Logo		Background	
	No Logo	**Logo**	**No Background**	**Background**
Positive Affect	2.69 (0.97)	2.69 (0.95)	2.79 (0.93)	2.59 (0.98)
Negative Affect	1.39 (0.72)	1.33 (0.60)	1.38 (0.69)	1.35 (0.62)
Fear	1.37 (0.73)	1.30 (0.61)	1.36 (0.69)	1.31 (0.65)
Hostility	1.37 (0.71)	1.30 (0.61)	1.39 (0.75)	1.27 (0.55)
Guilt	1.47 (0.85)	1.36 (0.69)	1.39 (0.72)	1.44 (0.82)
Sadness	1.69 (0.97)	1.59 (0.83)	1.62 (0.88)	1.67 (0.92)
Joviality	2.29 (1.09)	2.34 (1.05)	2.39 (1.05)	2.24 (1.07)
Self-Assurance	2.16 (0.90)	2.14 (0.97)	2.26 (0.95)	2.04 (0.91)
Attentiveness	3.29 (0.96)	3.38 (0.94)	3.46 (0.89)	3.20 (0.98)
Shyness	1.47 (0.75)	1.44 (0.74)	1.47 (0.76)	1.45 (0.72)
Fatigue	1.96 (0.97)	1.98 (0.99)	1.99 (0.92)	1.95 (1.03)
Serenity	3.38 (1.07)	3.46 (1.07)	3.52 (1.03)	3.32 (1.09)
Surprise	1.53 (0.85)	1.54 (0.89)	1.57 (0.89)	1.51 (0.85)

The means for each scale were compared using a two-way ANOVA (N = 445); for simplicity's sake, only the p-values from the two-way ANOVA tests are included in Table 5.

Table 5. P-values from two-way ANOVA tests for PANAS-X scales.

	Logo	Background	Logo x Background (interaction)
Positive Affect	0.933	0.029*	0.259
Negative Affect	0.319	0.582	0.254
Fear	0.361	0.552	0.361
Hostility	0.319	0.051	0.267
Guilt	0.138	0.544	0.311
Sadness	0.277	0.611	0.262
Joviality	0.677	0.122	0.368
Self-Assurance	0.767	0.015**	0.729
Attentiveness	0.331	0.005**	0.079
Shyness	0.658	0.814	0.925
Fatigue	0.776	0.644	0.300
Serenity	0.526	0.047*	0.091
Surprise	0.903	0.461	0.726

* Significant at the 0.05 probability level.
** Significant at the 0.01 probability level.
N = 445

As illustrated in Table 5, no significant results were found for logo or interaction. Therefore, H1-c is not supported. However, significant effects were found in the following emotional scales for background: positive effect (F(1, 441) = 4.77, p = 0.029); self-assurance (F(1, 441) = 5.91, p = 0.015); attentiveness (F(1, 441) = 7.84, p = 0.005); and serenity (F(1, 411) = 3.96, p = 0.047). Participants in the no background condition reported higher overall positive emotional effects,

were more attentive, self-assured, and serene, than participants in the dark background condition. Thus, H2-b is not supported.

Controlling for demographic variables (age, education, ethnicity, location, gender) added significantly in the following scales: positive affect (gender, ethnicity, age), self-assurance (gender, ethnicity, age), attentiveness (gender, age), and serenity (age).

However, one should note that with the high number of statistical comparisons being performed, it is possible that a significant result is just the result of random chance (Hsu, 1996). This study should be repeated in the future to investigate the replicability of the findings.

Chapter 5
Discussion

The results of the experiment show mixed results for the hypotheses, which are summarized in Table 6. Besides providing recommendations for the visual design of online surveys, the study also demonstrates a more detailed method for evaluating the response quality of surveys and the usefulness of using MTurk for survey research.

Table 6. Summary of research hypotheses, results, and conclusions.

Hypothesis	Finding	Conclusion
H1-a: Having a logo on every page of the survey will lead to better response quality.	Partially supported.	No background/logo condition had significantly lower drop-outs.
H1-b: Having a logo on every page of the survey will lead to higher social desirability score (lower data quality).	Not supported.	Logo does not affect the social desirability bias of respondents.
H1-c: Having a logo on every page of the survey will inspire trust and lead to a more positive emotional response.	Not supported.	Logo does not affect the emotional response of respondents.
H2-a: A dark background is more visually appealing and will lead to better response quality.	Partially supported.	Dark background condition had lower mean social desirability score; no background/logo condition had significantly lower drop-outs.
H2-b: A dark background is more visually appealing and will lead to more positive emotional response.	Not supported.	Respondents in the no background condition reported higher positive emotion effects, including attentiveness, self-assurance, and serenity.

5.1 Data Quality on MTurk

Unlike other MTurk studies, this study provides a detailed method for evaluating data aside from the actual content of the responses. By combining Ganassali's 2008 framework and Peer et al.'s (2014) data quality outcome measures, a more complete picture of survey data quality can be evaluated. In this experiment, data quality was measured along several variables: social desirability score, attention check questions, drop-out proportion, filling-up proportion, and variety of responses. These measures provide nuanced information about the response quality since they look at the overall characteristics of the responses: how many respondents finished the survey (drop-out), how much effort participants put into responding (attention check questions, variety of responses), how much of the survey they completed (filling-up), and if they appeared to show social desirability bias (social desirability score). By looking at these factors as well as the content of responses, researchers can better understand how respondents interacted with the survey and can make sure they are measuring responses correctly. If respondents have high social desirability scores or a very low variety of responses, their other responses cannot necessarily be taken at face value.

These evaluation criteria are particularly important for online surveys, since the environment where the survey is being taken cannot be controlled by the researcher and participants cannot be directly monitored. Researchers who use MTurk for their academic surveys may also be concerned that participants are inattentive or rush through the questions. Despite these concerns, previous

research has shown that the quality of data obtained on MTurk is high and respondents appear to pay close attention to the research tasks, particularly if they have a high approval rating on MTurk.

Since only high reputation and experienced participants were recruited, overall high data quality could be expected. The percentage of drop-outs for each condition was under 20%, the filling-up rate was high across conditions, and under 13% of participants failed two or more ACQs. This is in line with other studies on MTurk, where high reputation workers have produced high data quality. For these high reputation workers, the effects of the survey's visual design appear to be limited. The hypotheses on data quality, H1-a and H2-a, are only partially accepted.

There was no effect of the experimental factors on passing attention check questions, variety of responses, or filling-up proportion, but a significant effect was found for drop-out and experiment condition. The combination of no background and logo led to significantly lower drop-outs than the other three conditions. The drop-out proportion in the no background/logo condition was 3.38%, compared to approximately 12-13% in the other conditions. Drop-outs can lead to uneven cells or partial responses, which can be very problematic for some surveys (Ganassali, 2008). Therefore, if drop-outs are a concern for researchers, they may opt to use a logo and no background when designing a survey. Additionally, H1-b is not accepted: using a logo had no effect on respondents' social desirability score. For researchers who want to use a logo, they do not need to worry about it leading respondents to change their answers to seem more

socially acceptable to the researchers. However, the background seems to have had an effect on respondents' social desirability bias. Respondents in the no background condition had a higher social desirability score than those in the dark background condition. This was an unexpected result: it is unclear why there would be a relationship between background and social desirability score. The effect may be just a chance finding and should be reexamined in subsequent studies.

These findings demonstrate what kinds of the visual design choices researchers must make when creating online surveys, which will be discussed in the next section.

5.2 Online Survey Design

Online survey software design tools make it much easier for anyone to create surveys. Templates can be used out-of-the-box, or modified easily, and the range of possible visual design choices is extensive. The results of this study providence guidance for academics who want to understand what visual design choices they should worry about and what might affect their survey data. As Paolacci and Chandler (2014) attest, "The potential for arbitrary design choices to influence sample composition suggests that researchers should be transparent in the materials used in their studies and the methods used to recruit and exclude participants" (p. 187). Based on the results of this research, several design recommendations for researchers using online surveys can be made.

Using a logo does not appear to have a particularly strong effect on its own, but used in combination with no background significantly reduced the proportion of survey drop-outs. Using a logo also does not appear to affect the way participants respond either emotionally or increase their social desirability score. However, researchers need to pay attention to what background is used because it appears to have an effect on respondents' emotions and their social desirability score.

Regarding participants' emotional responses, based on the findings we fail to accept H1-c, but do accept H2-b. For emotional effects, using a logo does not change the way participants respond emotionally to the survey. However, the survey background can potentially change mood. Participants in the no background condition had higher means for positive emotion affect, self-assurance, attentiveness, and serenity, than participants in the dark background condition. One should note that, while participants in the no background condition reported feeling more attentive, since there was no significant difference in the ACQ measure between background conditions, this appears to only reflect participants' feelings rather than actual actions. These qualities may be desirable to some survey researchers, since respondents appear to be more assured and calm, and feel more positively overall.

Participants in the no background condition, however, also had higher social desirability scores, which indicates low data quality and flags participants' responses as potentially less genuine. This has important implications for researchers studying human emotions, morals or values, or those trying to collect sensitive or social information on MTurk. In these cases, socially desirable

reporting or changes in mood brought on by the survey design might affect the way participants respond and introduce measurement error. While qualities like attentiveness and self-assurance may be desirable traits that researchers want to inspire in participants, if the information they are investigating is socially undesirable, the background may have a negative effect. Depending on what is being measured or studied, researchers may decide that avoiding high social desirability bias is more important than promoting positive emotional effects in respondents. Critically, there were no significant effects in either experimental condition for the negative affect emotional scales (fear, hostility, guilt, and sadness), so researchers need not be concerned with accidentally disconcerting their respondents if they choose any of these conditions. Regardless of what information they are trying to collect or what phenomena they want to measure, researchers need to think carefully about how their survey instruments look, and should address not only their research design choices, but also specifically their visual design choices, when discussing their studies, and acknowledge that the visual design may affect data.

This study also provides further confirmation that MTurk is viable for conducting academic experiments, and further expands the type of experiments that can be conducted. By examining participants' interactions with online survey interfaces, it contributes to our understanding of user behavior online. With one exception (Mohr, Sell, & Lindsay, 2015), no other research thus far has investigated online survey design through using MTurk; this experiment demonstrates the usefulness of this type of research and provides a research

design that could be used to investigate other visual design features for online academic surveys. Additionally, MTurk participants could be used to examine other aspects of life online, such as social media use, online dating, e-government services, or other kinds of Internet usage. Demographic research has demonstrated that Internet users are systematically different from non-Internet users (Paolacci & Chandler, 2014), but research on MTurk has not investigated whether MTurk workers are typical or atypical of the Internet-using population. Besides providing new areas for future research, this could have implications for the way MTurk workers react to certain online survey design features compared to those who are less familiar with online surveys or with online interfaces in general.

5.3 Limitations and Future Work

The experiment has several limitations that may account for the absence of strong effects and also affect its generalizability to other, non-MTurk populations. With its highly experienced population, MTurk workers may be too proficient at certain kinds of academic research tasks and therefore fail to show expected effects (Rand et al., 2013). They also appear to be extremely comfortable with online interfaces. Even though effects were found, it is possible that other sample populations who are less familiar with online surveys and interface design may be affected by the experimental factors differently.

Recent research has shown that MTurk workers have become very experienced with academic studies: in a 2013 self-report questionnaire, the

median amount of academic studies taken by MTurk workers was 300, compared to a median of 15 for subjects in a lab (Rand et al., 2013). Participants may have been members of MTurk for several years. Because of this, participants are likely very familiar with certain scales, questions, or paradigms, and therefore may not answer intuitively, or need less time to deliberate answers to questions (Rand et al., 2013). A study by Chandler et al. (2015) demonstrated that prior exposure to tasks or questions can reduce the effect size of findings. Rand et al. (2013) also showed a decrease in the effect size for a common cooperation game from MTurk participants from 2011 to 2013. Since common and well-established scales (i.e. Marlowe-Crowne Short Form C, PANAS-X, IMC) were used to measure the effect of the experimental manipulations in this research study, the effect sizes may have been smaller than if the same experiment was given to naive participants. Some participants may have been focused only on getting through the task as quickly and efficiently as possible, and therefore may not have truly noticed or fully absorbed the design factors of the study; if they are answering research items that they have answered many times before, they can get through the survey very quickly. Similarly, Hauser & Schwarz (2015a) suggest that MTurk participants appear to learn how to look for ACQs, and this high level of attentiveness may influence not only how they respond to ACQs but to the survey overall.

However, sample populations who are new to online academic surveys may take more time to complete questionnaires and may take notice of the design in different ways. This experiment should be replicated on other

crowdsourcing sites or with other recruitment methods (such as email or forums) to see if other populations are affected by these design factors in different ways. For example, the current study only surveyed U.S. participants, but participants in other countries and who have different design traditions may react differently. As shown in the experiment results, demographic factors such as age, gender, and ethnicity appear to affect emotional response to the survey; therefore, it seems reasonable that other nationalities could exhibit distinctive or unique effects to the visual design of online survey.

Online survey design and academic studies on MTurk offer many avenues for further research. Similar experiments could be carried out that manipulate other simple design changes and investigate the effect on data quality. For example, does changing or mixing font types or having a photograph or textured background affect respondents? While an academic institution's logo appears to have little effect on this sample population, would the logo of a corporation have a different influence? As discussed in Chapter One, MTurk workers tend to put more effort into tasks that are framed as academic or non-profit research compared to those framed as corporate research. Thus, the logo of a corporation may elicit different findings. What about if institution's reputation is positive, negative, or unknown? The University of Oxford is a global, highly reputable institution; an academic institution with less prominence or an unfavorable reputation could demonstrate a different kind of effect on data quality or emotional response.

This research also shows that survey design researchers and methodologists need to revisit and update previous findings on visual design choices as potential survey participants become more used to web interface design features and new types of interfaces are introduced. Certain design aspects may matter more or less depending on how familiar different populations interact with online surveys. This same research design could easily be used for other, more noticeable design factors in order to see how other participant samples react. Recalling Kapelner, and Chandler's 2010 study, using a survey design with light text on a dark background may produce much stronger effects than logo or background. Their paper indicated that participants noticed and were bothered by this design, but recreating this study's experiment using Kapelner and Chandler's visual design factors could formalize what these effects actually are and how it may influence responding.

Furthermore, evolving online interfaces introduce new design features, question types, and methods for taking surveys that need to be investigated for their potential effects on data. For example, as mobile devices are becoming more and more popular, more academic surveys may be completed on these devices. Mobile interface design can be entirely different from web interfaces, but there is relatively little research on mobile survey design.

5.3.1 Mobile

Only a small percentage of the surveys in this study were taken on a mobile device, but future online surveys could have higher rates of mobile usage.

While mobile-optimized versions of a survey can be created, the experience may be different: the response format changes and the literal interaction with the survey changes (the text and images are smaller, respondents touch and interact intimately with the device). Depending on the question formats and structure of the survey, the mobile version may be a completely different survey. For instance, matrix questions are completed in a different format on mobile than on a desktop or laptop computer, as shown in Figure 5: each item has its own scale and must be opened separately in order to answer the question for that item.

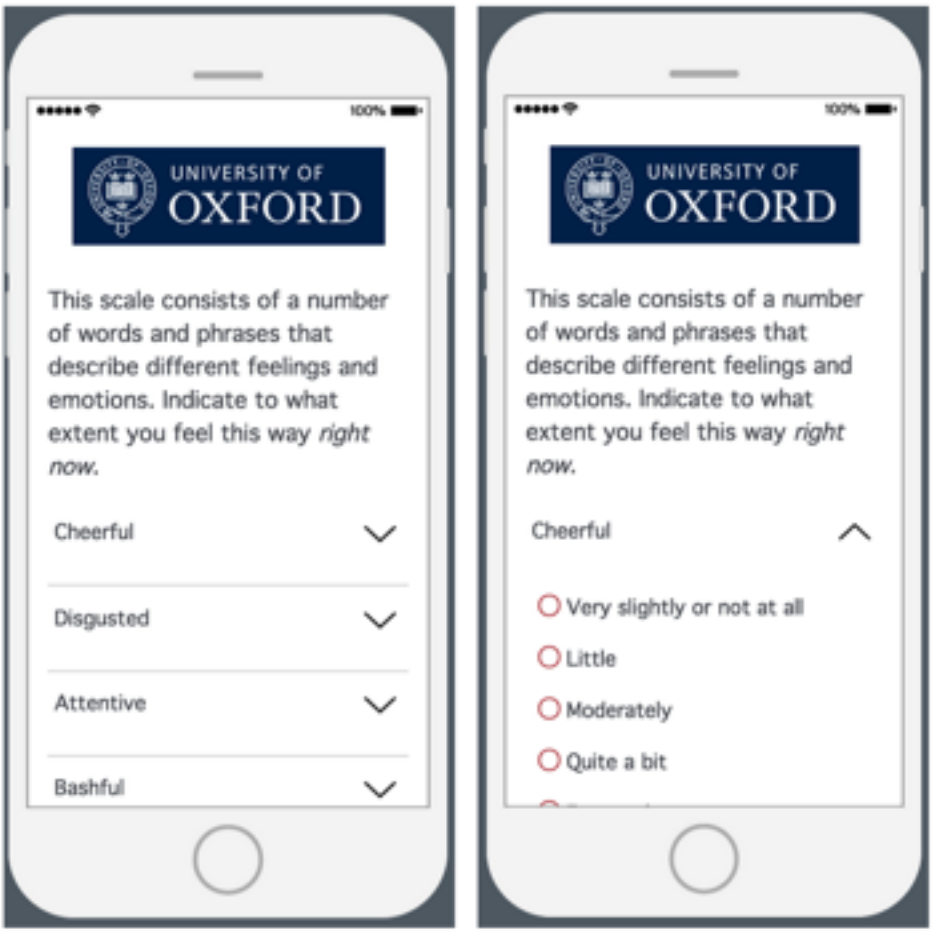

Figure 5. Mobile version of the survey (logo/no background condition), showing the optimized version of the PANAS-X matrix question.

Visual design factors may effect users differently on mobile surveys. The mobile format itself may also introduce new research factors or tasks, for example, asking participants to use certain mobile features like the GPS or camera to contribute data to the survey, that must be studied for their effects on participants as well as their overall effectiveness for research and data collection.

Surveys that are not responsive or optimized for different browsers or formats may contain errors or missing elements which affect the participant's ability to respond correctly (Sue & Ritter, 2012). In this study, Qualtrics automatically created a mobile version of the survey. A survey can also be manually optimized for mobile, but there is less literature on mobile survey design to use as a guideline (Sue & Ritter, 2012). Allowing participants to take surveys on mobile devices also changes when and where participants can respond (Dillman, Smyth, & Christian, 2014). Depending on their surroundings, it may be more difficult for them to concentrate. Because of these differences, responses collected on a mobile device may need to be analyzed separately, especially if they make up a significant proportion of the responses collected (Sue & Ritter, 2012). Academic studies should find out how the survey was taken and compare the responses collected on different devices to see if there are any significant differences in the data quality. More research is needed in the area of mobile survey design in order to determine how participants interact with surveys on mobile devices, and if it creates any difference in the way they respond to survey questions.

Chapter 6
Conclusion

With the growth of online surveying and crowdsourced work, the importance of understanding and accounting for the effects of visual design also increases. Failure to take the effects of visual design choices into account may introduce error into data analysis and lead researchers to incorrect conclusions about their data. Online surveys and crowdsourcing platforms are useful and convenient tools for conducting research, but researchers should assume that even simple design choices like the background color of the survey could have an effect on respondents.

As seen in other MTurk studies, the MTurk crowdsourcing platform is appropriate for conducting academic research, including experiments, and that recruiting from high reputation workers produces good quality data. This study presents an expanded framework for designing online surveys and assessing data quality, and situates the collected data within the context of the survey's design and recruitment method. There is more to response quality than just the answer content: by analyzing this study's data along several dimensions, it provides other researchers with a guide for what to expect from typical MTurk surveys and how to evaluate the data obtained. The results of this study have discovered that the answer to the research question, *Does the design template of an online survey and use of the researcher's logo affect the data quality of respondents?* is both yes and no. The researcher's logo affects data quality in a

limited way (i.e. on participant drop-out, when combined with no background), and the survey background affects participants' social desirability bias and certain emotional responses. Through a review of the overall characteristics of responses (e.g. variety of responses, filling-up proportion) and emotional effect in addition to the content of the survey, researchers can glean a fuller picture of how participants reacted to and interacted with their survey. The framework also highlights what design factors should be taken into account when preparing an online survey.

This study contributes to the literature on online survey methodology by examining two common, though often overlooked, visual design choices. The results of this experiment give academic researchers guidance on picking certain simple visual design elements of surveys, in this case using researcher logos or dark backgrounds. Using a logo does not appear to have any significant effect on its own, so researchers may decide whether or not to use it without great concern for its effect on data quality or respondents' emotions. In contrast, the survey's background does appear to affect some aspects of data quality and emotional response. Having no background seems to increase both social desirability bias and respondents' positive emotions. Researchers who need to measure certain social biases or emotional dimensions of their participants should take the results of this research under advisement when they design an online survey. Crucially, they should understand how their visual design choices may affect responses to their survey, and discuss these choices when disseminating their work.

Crowdsourcing research is a popular and useful method that allows researchers to collect more responses more easily than other methods. While a crowdsourcing site like MTurk may not be appropriate for every type of research, MTurk remains an important participant recruiting pool and multiple studies have demonstrated the high quality of data which can be collected from the site.

Since online surveys and crowdsourcing platforms are cheaper, faster, and more convenient, more research across many disciplines is being done this way. A keyword search in Scopus for "online survey" produced over 1,000 results for articles published between 2012 and 2014, and "crowdsourcing" returned over 500 results. These are clearly prominent ways that research will be conducted in the future; however, if the reporting on how these types of studies are designed is not transparent, it becomes difficult to determine how reliable or generalizable these studies truly are.

When designing research studies, nothing can be taken for granted. Researchers ought to think as carefully about the design of the survey as about the content of the questions they ask, as the two are closely intertwined: a survey cannot correctly measure a phenomenon if the method of measurement introduces error or noise. This research study demonstrates how data quality from MTurk academic surveys can be analyzed and how even simple design factors can influence respondents.

Bibliography

Includes all works cited in this thesis

Amazon Web Services. (2015). Service summary. *Amazon Mechanical Turk*. Retrieved from https://requester.mturk.com/tour

Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one, 7*(2), e31461.

Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PloS one, 10*(4).

Benoit, K., Conway, D. Lauderdale, B. E., Laver, M., and Mikhaylov, S. (2015) Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368.

Brewer, M. B. & Crano W.D. (2014) Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed.). (pp. 11-26). Cambridge: Cambridge University Press.

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization, 90*, 123-133.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological science*, 0956797615585115.

Chen, D. L., & Horton, J. J. (2009). The wages of pay cuts: Evidence from a field experiment. Available at SSRN 1443526.

Christian, C., Lintott, C., Smith, A., Fortson, L., & Bamford, S. (2012). Citizen science: Contributions to astronomy research. *arXiv preprint arXiv: 1202.2577*.

Couper, M. P. (2008). *Designing effective web surveys* (Vol. 75). New York: Cambridge University Press.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one, 8*(3), e57410.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

Duggan, M., Ellison, N.B., Lampe, C., Lenhart, A,. and Madden, M. (2015). Social media update 2014. *Pew Research Center*. Retrieved from: http:// www.pewinternet.org/2015/01/09/social-media-update-2014/

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior, 26*(2), 132-139.

Funke, F. (2015). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, 0894439315575477.

Ganassali, S. (2008, March). The influence of the design of web survey questionnaires on the quality of responses. In *Survey Research Methods, 2*(1), 21-32.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213-224.

Hall, R. H., & Hanna, P. (2004). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology, 23*(3), 183-195.

Hauser, D. J., & Schwarz, N. (2015a). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 1-8.

Hauser, D. J., & Schwarz, N. (2015b). It's a trap! Instructional Manipulation Checks prompt systematic thinking on "tricky" tasks. *SAGE Open, 5*(2), 2158244015584617.

Horton, J. J., & Chilton, L. B. (2010, June). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce* (pp. 209-218). ACM.

Hsu, J. (1996). *Multiple comparisons: theory and methods.* CRC Press.

Joinson, A. N., & Reips, U. D. (2007). Personalized salutation, power of sender and response rates to Web-based surveys. *Computers in Human Behavior, 23*(3), 1372-1383.

Kang, R., Brown, S., Dabbish, L., & Kiesler, S. (2014, July). Privacy attitudes of Mechanical Turk workers and the US public. In *Symposium on Usable Privacy and Security (SOUPS)*.

Kapelner, A., & Chandler, D. (2010). Preventing satisficing in online surveys: A 'kapcha' to ensure higher quality data. In *The World's First Conference on the Future of Distributed Work (CrowdConf2010)*.

Kaufmann, N., Schulze, T., & Veit, D. (2011, August). More than fun and money: Worker motivation in crowdsourcing - a study on Mechanical Turk. In *AMCIS, 11*, 1-11.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013, April). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 207-216). ACM.

Lauer, C., McLeod, M., & Blythe, S. (2013). Online survey design and development: A Janus-faced approach. *Written Communication, 30*(3), 330-357.

Mahon-Haft, T. A., & Dillman, D. A. (2010, May). Does visual appeal matter? Effects of web survey aesthetics on survey quality. In *Survey Research Methods, 4*(1), 43-59.

Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter, 11*(2), 100-108.

Mohr, A. H., Sell, A., & Lindsay, T. (2015). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, 0894439315588736.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237-241.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making, 5*(5), 411-419.

Paolacci, G., & Chandler, J. (2014). Inside the turk understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184-188.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023-1031.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*.

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*(4), 364-382.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe–Crowne Social Desirability Scale. *Journal of clinical psychology, 38*(1), 119-125.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011, March). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 275-284). ACM.

Shropshire, K. O., Hawdon, J. E., & Witte, J. C. (2009). Web survey design: Balancing measurement, response, and topical interest. *Sociological Methods & Research, 37*(3), 344-370.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast---but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). Association for Computational Linguistics.

Srinilta, W., Anuntavoranich, P., Malisuwan, S., & Taraphokakul, S. (2012). A study on preference of interface design techniques for Web survey. *International Journal of Computer Science Issues (IJCSI), 9*(5).

Staffelbach, M., Sempolinski, P., Hachen, D., Kareem, A., Kijewski-Correa, T., Thain, D., ... & Madey, G. (2014). Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with Amazon Mechanical Turk. *arXiv preprint arXiv:1406.7588*.

Sue, V. M., & Ritter, L. A. (2012). *Conducting online surveys*. Sage.

Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity, 40*(3), 435-456.

Vehovar, V., & Manfreda, K. L. (2008). Overview: online surveys. The *SAGE handbook of online research methods*, 177-194.

Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review, 28*(2), 251-267.

Wang, H. C., & Doong, H. S. (2010). Nine issues for Internet-based survey research in service industries. *The Service Industries Journal, 30*(14), 2387-2399.

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule-expanded form.

Winkielman, P., Schwarz, N., Reber, R. and Fazendeiro, T. A. (2003). Cognitive and affective consequences of visual fluency: When seeing is easy on the mind. In L. M. Scott & R. Batra (Eds.), *Persuasive imagery: A consumer response perspective.* (pp. 75–89). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Appendix A: Participant Demographics

Table 7. Participant age.
Original categories (55-64 years old, 65-74 years old, 75 years or older) collapsed into a single category, 55 years or older.

18-24 years old	48
25-34 years old	159
34-45 years old	106
45-54 years old	66
55 years or older	47
Total	426

Table 8. Participant gender.

Male	176
Female	249
Other	1
Total	426

Table 9. Participant ethnicity.
Original categories (Hispanic or Latino, Black or African-American, Native American or American Indian, Asian/Pacific Islander, Other) collapsed to single category, Nonwhite.

White	337
Nonwhite	89
Total	426

Table 10. Participant education.
Original categories collapsed (Some high school, no diploma, High school graduate, diploma or equivalent = High school or less); (Trade/technical/ vocational training, Associate degree = Associate or technical degree); and (Master's degree, Professional degree, Doctorate degree = Graduate degree).

High school or less	45
Some college, no degree	116
Associate or technical degree	64
Bachelor's degree	155
Graduate degree	41
Total	421

Table 11. Participant location.
Original answer choices (individual states) collapsed into the four U.S. Census Regions.

Northeast	61
Midwest	95
South	170
West	100
Total	426

Appendix B: Item Composition of the PANAS-X Scales

Table 12. Item composition of the PANAS-X scales (Watson & Clark, 1999).

General Dimension Scales	
Negative Affect (10)	afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed
Positive Affect (10)	active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong
Basic Negative Emotion Scales	
Fear (6)	afraid, scared, frightened, nervous, jittery, shaky
Hostility (6)	angry, hostile, irritable, scornful, disgusted, loathing
Guilt (6)	guilty, ashamed, blameworthy, angry at self, disgusted with self, dissatisfied with self
Sadness (5)	sad, blue, downhearted, alone, lonely
Basic Positive Emotion Scales	
Joviality (8)	happy, joyful, delighted, cheerful, excited, enthusiastic, lively, energetic
Self-Assurance (6)	proud, strong, confident, bold, daring, fearless
Attentiveness (4)	alert, attentive, concentrating, determined
Other Affective States	
Shyness (4)	shy, bashful, sheepish, timid
Fatigue (4)	sleepy, tired, sluggish, drowsy
Serenity (3)	calm, relaxed, at ease
Surprise (3)	amazed, surprised, astonished

Appendix C: Survey Content

Please answer the following questions regarding your social media use and habits.

Q. Do you ever use the Internet to… (Scale: Did yesterday, Have done this before, Have never done this, Don't know)

Use Twitter

Use Instagram

Use Pinterest

Use LinkedIn

Use Facebook

Q. Thinking about the social networking sites you use, about how often do you visit or use… (Scale: Several times a day, About once a day, A few days a week, Every few weeks, Less often, Don't know, Don't use)

Twitter

Instagram

Pinterest

LinkedIn

ChatFace

Facebook

Q. Thinking about your use of Facebook, approximately how many total Facebook friends do you have? (Free text)

Q. Approximately how many of your total Facebook friends do you consider actual friends? (Free text)

Q. How often do you share, post or comment on Facebook, as opposed to reading or viewing content?

Frequently

Sometimes

Hardly ever

Never

Don't know

Prefer not to answer

Q. Thinking about who is in your Facebook network, are you Facebook friends with… (Scale: Yes, No, Doesn't apply, Don't know, Prefer not to answer)

Your parents

Your children

Other family members

Work colleagues

Friends from the past, such as high school or college

Current friends

Neighbors

People you have never met in person

[Page break]

Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether the statement is True or False as it pertains to you personally.

Q. It is sometimes hard for me to go on with my work if I am not encouraged.

True

False

Q. I sometimes feel resentful when I don't get my way.

True

False

Q. On a few occasions, I have given up doing something because I thought too little of my ability.

True

False

Q. There have been times when I felt like rebelling against people in authority even though I knew they were right.

True

False

Q. No matter who I'm talking to, I'm always a good listener.

True

False

Q. There have been occasions when I took advantage of someone.

True

False

Q. I'm always willing to admit when I make a mistake.

True

False

Q. I sometimes try to get even rather than forgive and forget.

True

False

Q. I am always courteous, even to people who are disagreeable.

True

False

Q. I have never been irked when people expressed ideas very different from my own.

True

False

Q. There have been times when I was quite jealous of the good fortune of others.

True

False

Q. I am sometimes irritated by people who ask favors of me.

True

False

Q. I have never deliberately said something that hurt someone's feelings.

True

False

[Page break]

Q. This scale consists of a number of words and phrases that describe different feelings and emotions. Indicate to what extent you feel this way right now. (Scale: Very slightly or not at all, Little, Moderately, Quite a bit, Extremely)

Cheerful	Sad	Active	Angry at self
Disgusted	Calm	Guilty	Enthusiastic
Attentive	Afraid	Joyful	Downhearted
Bashful	Tired	Nervous	Sheepish
Sluggish	Amazed	Lonely	Distressed
Daring	Shaky	Sleepy	Blameworthy
Surprised	Happy	Excited	Determined
Strong	Timid	Hostile	Frightened
Scornful	Alone	Proud	Astonished
Relaxed	Alert	Jittery	Interested
Irritable	Upset	Lively	Loathing
Delighted	Angry	Ashamed	Confident
Inspired	Bold	At ease	Energetic
Fearless	Blue	Scared	Concentrating
Disgusted with self	Shy	Drowsy	Dissatisfied with self

[Page break]

Please answer the following demographic questions.

Q. Please specify your gender

Male

Female

Other

Prefer not to answer

Q. Please specify your age

18-24 years old

25-34 years old

35-44 years old

45-54 years old

55-64 years old

65-74 years old

75 years or older

Prefer not to answer

Q. Please specify your ethnicity

White

Hispanic or Latino

Black or African American

Native American or American Indian

Asian / Pacific Islander

Other

Prefer not to answer

Q. Please specify your location (Drop-down list of U.S. states/territories)

Q. What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

Less than high school

Some high school, no diploma

High school graduate, diploma or the equivalent (for example: GED)

Some college credit, no degree

Trade/technical/vocational training

Associate degree

Bachelor's degree

Master's degree

Professional degree

Doctorate degree

Prefer not to answer

Q. How did you fill out this survey?

Desktop or laptop

Tablet

Mobile phone

Q. How many hours do you spend on MTurk each week? Please ignore this question and simply click the >> button to proceed.

Less than 5 hours a week

6-10 hours a week

11-20 hours a week

21-30 hours a week

Over 30 hours a week

[Page break]

Thank you for participation, you are almost done. Most modern theories of decision-making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. In order to facilitate our research on decision-making we are interested in knowing certain factors about you, the

decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read the instructions, please select Other and type "Internet" in the text box.

Q. What was this survey about?

Social Media

Psychology

Design

Other