ABSTRACT

Title of dissertation: Classifying Mouse Movements and Providing Help in Web Surveys

Rachel Horwitz, Doctor of Philosophy, 2013

Directed by: Professors Fredrick Conrad and Frauke Kreuter,

Joint Program in Survey Methodology

Survey administrators go to great lengths to make sure survey questions are easy to understand for a broad range of respondents. Despite these efforts, respondents do not always understand what the questions ask of them. In interviewer-administrated surveys, interviewers can pick up on cues from the respondent that suggest they do not understand or know how to answer the question and can provide assistance as their training allows. However, due to the high costs of interviewer administration, many surveys are moving towards other survey modes (at least for some respondents) that do not include costly interviewers, and with that a valuable source for clarification is gone.

In Web surveys, researchers have experimented with providing real-time assistance to respondents who take a long time to answer a question. Help provided in such a fashion has resulted in increased accuracy, but some respondents do not like the imposition of unsolicited help. There may be alternative ways to provide help that can refine or overcome the limitations to using response times.

This dissertation is organized into three separate studies that each use a set of independently collected data to identify a set of indicators survey administrators can use to determine when a respondent is having difficulty answering a question and proposes alternative ways of providing real-time assistance that increase accuracy as well as user satisfaction.

The first study identifies nine movements that respondents make with the mouse cursor while answering survey questions and hypothesizes, using exploratory analyses, which movements are related to difficulty. The second study confirms use of these movements and uses hierarchical modeling to identify four movements which are the most predictive. The third study tests three different modes of providing unsolicited help to respondents: text box, audio recording, and chat. Accuracy and respondent satisfaction are evaluated for each mode. There were no differences in accuracy across the three modes, but participants reported a preference for receiving help in a standard text box. These findings allow survey designers to identify difficult questions on a larger scale than previously possible and to increase accuracy by providing real-time assistance while maintaining respondent satisfaction.

CLASSIFYING MOUSE MOVEMENTS AND PROVIDING HELP IN WEB SURVEYS

By

Rachel Horwitz

Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2013

Advisory Committee:
Dr. Frederick G. Conrad, Co-chair
Dr. Frauke Kreuter, Co-chair
Dr. Kent L. Norman
Dr. Stanley Presser
Dr. Roger Tourangeau

Acknowledgements

My path towards this dissertation and PhD inadvertently began my sophomore year of college in a market research class. As a student in that class, I had no clue that survey methodology existed or that I would ever pursue an advanced degree. However, through a series of unrelated events, I stumbled my way into the JPSM Master's program and finally the PhD program. While I never thought I would be here, I am extremely lucky and grateful to the people guided, pushed, and supported me in finding a career that I truly enjoy.

First, I would first like to thank Frauke Kreuter, my advisor during the Master's program and my dissertation co-chair. About a year before I decided to apply for the PhD program, Michael Lemay suggested that I talk to Frauke about it. He also said that no matter what I was thinking at the time, after talking to Frauke I would definitely apply to the program. At that point I was pretty sure I wasn't going to apply, but figured I'd talk to her anyway. Little did I know, Michael was absolutely right. Over the years, Frauke was always available to talk to, answer questions, and provide guidance. She would not often tell me how to do something, but would instead just hand me a book. I didn't necessarily like this approach while wading through what felt like mountains of work, but I learned different analysis techniques that I never would have if she had just given me the answer, and I am very appreciative of that.

I would also like to thank Fred Conrad, my dissertation co-chair. Fred asked the questions that would make me pause and really figure out what I wanted to say in my dissertation. He always found time in his ever-busy schedule to talk to me and provide helpful comments and guidance. I would also like to thank Roger Tourangeau for getting me started on my dissertation. He took a leap and supported an idea that had a pretty good shot of not panning out. I'd also like to thank the rest of my dissertation committee – Stanley Presser and Kent Norman – for their valuable feedback and support through this process.

I am extremely thankful to the Census Bureau, especially Jennifer Tancreto, Mary Frances Zelenak, Mary Davis, and Debbie Griffin for their flexibility and support over the past two years. They have read drafts, listened to presentations, and supported me through moments of frustration. Also at the Census Bureau, I'd like to thank Emily Sherwood, John Studds, and Glenn Eans for programming the survey instrument and providing me with data output. There is no way I would have been able to complete this dissertation without all of their assistance.

I would also like to recognize past teachers and professors that taught me the skills needed to complete this dissertation: Ann Helm taught me how to write, Carl Smith and Rick Vallient were endlessly patient while making me to figure out how to do math on my own, Marcia Kasieta introduced me to the world of surveys, and Mick Couper, taught me all about Web surveys and paradata.

I want to thank all of my friends for their support over the past four years. Special thanks go to Rebecca and Todd, who may be the only people outside of my committee and family who actually know what this dissertation is about, and Jennifer Lee who convinced me to take my first JPSM class. Additionally, I'd like to thank Becca and Jenna from the JPSM program for answering all of my many questions and their "you can do it” attitude as well as Duane; without his IT expertise, the research in this dissertation would not have been possible.

Finally, I’d like to give a very special thanks to my family who listened and supported me through the highs and lows of this journey. They encouraged me when I got bogged down and cheered me on when I reached milestones.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

1.1 Introduction

A goal of the survey interview is to collect reliable and valid data from respondents. Achieving this goal is often difficult because respondents may provide incorrect data intentionally, misinterpret the meaning of a question, satisfice, or answer incorrectly for any number of other reasons. Cognitive interviews and pretests are often used to identify unintentional question errors before production. However, these interviews and tests are generally conducted with a small sample of respondents, so it is unlikely that they identify every type of problem a respondent might encounter. Therefore, even after taking these precautions, there is still room for respondent error in answering survey questions. Some of the most common and systematic errors come from the cognitive difficulties inherent in responding to a question, either because of vague or unfamiliar terms, common words used in a way that differs from a respondent's everyday sense (*misalignment*), or difficulty mapping one's personal experience to the answer categories (*imperfect fit*) (Tourangeau *et al*., 2006). If left uncorrected, these systematic errors could lead to biased estimates. Therefore, we need to find a way to identify respondents when they are facing these types of comprehension problems so we can improve the questions before the survey goes live, revise the questions in real time, or provide assistance.

In traditional survey modes, such as computer assisted telephone interviewing (CATI) and computer assisted personal interviewing (CAPI), interviewers can pick up on signs of respondent difficulty from speech patterns or facial expressions and can provide assistance, in the form of re-reading definitions or repeating the question to give the respondent more time, if needed. However, researchers are facing growing challenges in reaching people through these traditional survey modes. Mail surveys using address-based sampling (ABS) are becoming more popular, and in a mail survey signs of trouble cannot be seen in real time. Another new trend in survey research is administering surveys via the Web. Although Web surveys do not utilize an interviewer, they provide an interactive environment with a vast amount of data, in addition to respondents' answers, that can be collected in real time. These data come at minimal additional cost and can therefore be used in large production surveys.

This dissertation describes three studies, the results of which will provide researchers with a means to identify confused Web survey respondents and provide them with real-time assistance in the best manner possible. The first study identifies a set of specific mouse movements commonly used by respondents while answering survey questions and generates hypotheses as to which may be related to confusion or difficulty. Further, it provides the tools to generate hypotheses about what types of movements are associated with different types of difficulty. The second study statistically tests the hypotheses set forth in the first study to create a model that identifies the mouse movements that are related to difficulty answering a survey question and attempts to link specific movements to specific types of difficulty. The final study determines the best method of providing Web respondents real-time assistance that leads to increased accuracy as well as high respondent satisfaction.

1.2 Literature Review

Researchers have studied how to identify respondents who are having difficulty answering survey questions in a variety of modes. Using this information, they have also attempted to provide real-time assistance to respondents they think are experiencing difficulty in order to collect higher quality data. This research provides the basis for the studies described in this dissertation, which intend to introduce a new way to predict respondent difficulty and to determine the best method of providing real-time assistance.

1.2.1 Types of Difficulty

When respondents read a survey question, they do not all necessarily process the meaning of the question in the same way. For example, some respondents may already have an idea of what a certain concept means and answer accordingly without thinking of other potential interpretations. In other cases, the structure of the sentence or the words used may be difficult for respondents to understand. Misinterpretation of survey questions can occur for many other reasons, which have been outlined by several survey researchers (Tourangeau *et al.,* 2000; Suessbrick *et al.*, 2005; Tourangeau *et al.,* 2006).

Tourangeau, Rips, and Rasinski (2000) describe five common reasons for question misinterpretation: structural ambiguity, ambiguous, vague, or unfamiliar terms, presuppositions, syntax, and question pragmatics. Additionally, in 2006, Tourangeau and his colleagues added two additional types of difficulty: imperfect fit and misalignment. While all these types of difficulty can affect survey response, the studies described in this dissertation will focus on three types: imperfect fit, misalignment, and technical or unfamiliar terms. These types of difficulty can occur in government surveys when the structured definitions do not necessarily coincide with the general population’s interpretations of the same words and ideas. Additionally, they are easy to manipulate in a laboratory setting.

Imperfect fit occurs when respondents' answers do not line up with the response options provided. These often involve cases where respondents do not know whether their answer is included in the response options, referred to as borderline cases. For example, in the 2011 American Community Survey (ACS) Internet Test usability testing[1], one respondent had taken several continuing education classes in the past year and he did not know whether that counted as attending school. Another example is whether a floor lamp counts as furniture (Ehlen *et al*., 2005). When there was no definition provided to describe exactly what "furniture" and other everyday concepts cover, only 20 percent of respondents correctly classified borderline cases. As this was a laboratory study, the scenarios induced bias because all incorrect responses were in the same direction. In a real survey environment, we cannot say with certainty that the response errors would be systematic. However, these studies make it clear that there can be great potential for either increased bias or variance if survey administrators are unable to compensate for discrepancies between a respondent's experience and the concepts assumed by the questions.

Misalignment occurs when the definition the respondent applies does not coincide with the definition assumed as provided by the question. Suessbrick and her colleagues (2005) found, in a study about attitudes and behaviors, that respondent's definitions matched the interviewer's in only 44 percent of cases. This misalignment resulted in unreliable answers 14 percent of the time; where respondents changed their answer after being exposed to the interviewer's definition. Similarly, in government surveys, Tourangeau *et al.* (2006) identified two examples of misalignment: "usual residence" and

[1] The 2011 ACS Internet Test usability testing was conducted at the U.S. Census Bureau headquarters in Suitland, MD from September, 2010 through October, 2010.

"disability" as asked in items used by the Census Bureau. Even when respondents received the technical definition for either usual residence or disability, they classified only 49 percent of the scenarios they received correctly. Respondents already had an idea of what these terms meant and these definitions did not necessarily match the Census Bureau's official definitions. This finding suggests respondents were not able to ignore their own idea of what these two terms constituted, regardless of the true definition. Considering the conclusions from these studies, misalignment is particularly troubling because not only do a high proportion of respondents not interpret questions as the designers intend, but many have difficulty relating to a different definition.

Unfamiliar terms can be a stand-alone problem or intertwined with imperfect fit and misalignment. In the most general sense, unfamiliar terms refer to either words or ideas the respondent is not familiar with. If a respondent cannot understand a question, the chances of providing an accurate answer are not high. For example, in an experiment, Conrad and his colleagues (2006) asked respondents how much fatty acid they consumed daily. If respondents do not know what fatty acids are or what foods contain them, they will be unable to answer this question. On the other hand, unfamiliar terms can relate to imperfect fit and misalignment if respondents believe they know what a term means, but they actually do not. When this occurs, due to misinterpretation, they may incorrectly map their response to the answer categories or apply the wrong definition to a word.

1.2.2 Signs of Trouble

Before researchers can provide respondents with assistance, they need to target a set of behaviors that are indicative of misunderstanding or confusion in responding to survey questions. These behaviors change depending on the mode of administration and the level of interviewer involvement. For example, in face-to-face interviewer-administered surveys, the interviewer can notice facial expressions or changing intonations in the respondent's voice that can be signals for help (Smith and Clark, 1993; Brennan and Williams, 1995). In addition, the elapsed time for each question and for the entire questionnaire, verbal expressions, breakoffs, and changed answers, along with many other behaviors, can signal respondent difficulty in understanding questions.

Survey researchers have used this information to determine when respondents are confused by a question. For example, Schober and Bloom (2004) looked at respondents' first utterances in a telephone survey to determine if there were cues, or signals, which suggested respondents were having difficulty with a particular question. They found that midclause or preutterance pauses, fillers (*um, uh, and mm*), repairs (in which speakers correct what they have said), and hedges (*I think, about*), collectively referred to as misalignment cues, were all predictive of respondents needing clarification.

In another study, Smith and Clark (1993) had respondents answer 40 factual questions and rate their own feeling of knowing (FOK). They found the weaker the FOK, the longer participants took to provide an answer, but the faster they provided nonanswers ("Don't know"). Additionally, consistent with Schober and Bloom's findings, weaker FOK resulted in answers with more hedges and fillers. In 1995, Brennan and Williams conducted three follow-up studies to Smith and Clark's. Again, they found respondents produced answers more quickly when they had a higher FOK. Additionally, rising intonation was associated with an incorrect answer 64 percent of the time. Brennan and Williams also had participants rate another persons' feeling of knowing (referred to as "feeling of another's knowing" or FOAK in their paper) based on how the other person responded to a series of questions. FOAK was higher when the answer was followed by a short pause as compared to a long pause. Further, listeners rated FOAK higher when the answer was spoken with falling intonation. Finally, if the respondents filled their pauses with “um” or “uh,” listeners rated their FOAK lower than if the pauses of the same length were unfilled.

Based on these cues, interviewers should be able to identify when a respondent is having difficulty with a question. However, surveys are increasingly automated and self-administered, largely as an attempt to eliminate increasing contact costs in interviewer-administered modes. Therefore, a new method of predicting cognitive difficulty is needed. To respond to this challenge, Ehlen, Schober, and Conrad (2005) used the misalignment cues identified in prior studies to create a model to be used in automated telephone studies for predicting difficulty. They found the best predictor of misaligned responses is response latency – the time elapsed between the end of the interviewer’s question and the beginning of the respondent’s answer. Specifically, they identified a “Goldilocks range,” defined as a window of time in which respondents who understand the question will respond. This definition suggests that when questions are answered either too quickly or too slowly it indicates misalignment.

Although the response latency model was originally created for integrated voice response systems (IVR), other researchers have extended the model to Web surveys. In 2003, Heerwegh used a Web survey to replicate a Bassili and Fletcher study (1991) that looked at response latencies and their relationship to attitude stability. They asked respondents their opinion on a question, and once that question was answered, the respondents were presented with a counter argument. They suggested that respondents who changed their answers after being exposed to the counter argument possessed unstable attitudes. As expected based on Bassili and Fletcher's findings, Heerwegh found that the average response times for the respondents who did not change their answers were significantly lower than the average for the respondents who did change their answer across two waves of a survey. Heerwegh also looked at response latencies to knowledge questions (2003). He asked respondents three different knowledge questions. For each question, the average response time for respondents who answered correctly was significantly shorter than the average for respondents who answered incorrectly. These results suggest that when respondents do not know the answer to a question, response times increase as they try to generate an answer instead of simply retrieving it from memory.

While the research in response latencies is helpful in identifying when respondents are having trouble answering a question, we do not know whether it is because they are confused by the words, the interpretation, the mapping, or if they understand the question but simply do not know the answer. In interviewer-administered surveys, the interviewer can engage in dialogue to determine why a respondent is taking a long time to answer a question. However, in automated surveys, additional indicators are needed to understand the respondent's thought process.

1.2.3 Eye Tracking

The Web provides a unique opportunity, unlike other automated data collection modes, to collect additional information while respondents are completing the survey. One use of these data is in pretesting to evaluate question wording and learn about how respondents answer survey questions. It is for this purpose that eye tracking can be used. Eye tracking has been used for decades in a number of fields, most notably in psychology to study cognitive processing and visual attention (Rayner, 1983; Rayner, 1992; Duchowski, 2007). More recently, however, it has been used in Web design to determine where to display information on Web sites based on where users focus their attention (Pan *et al.*, 2004; Tatler, 2007).

In the survey literature, eye tracking has been used to supplement other information. For example, while a long response time could suggest a respondent does not know the answer to a question, eye-tracking data can show whether there is a specific word the respondent is having trouble with or if the respondent does not understand the response categories (Redline *et al*., 2009). Specifically, Redline and her colleagues experimented with how to present long lists of response options. Their experiment grouped the options conceptually, with and without headings. They found longer response times for questions with headings, but they did not know why. However, after looking at the eye tracking data, they discovered respondents were circling back and forth between the question and the response options. Their conclusion, drawn by examining all the available paradata, was that respondents were confused by the headings and thought they should provide a response for each group of response options. Eye tracking in surveys has also been used to identify questions that are difficult to comprehend (Graesser *et al*., 2006). For items that include unfamiliar technical terms, Grasser and his colleagues found longer fixation durations on those terms, longer reading times, and more fixations. These studies indicate that eye tracking can provide valuable information about how respondents both answer questions and behave when they are confused.

In another study, Galesic and her colleagues (2008) conducted an eye tracking experiment where they measured how long respondents fixated on definitions in Web surveys. There was a “rollover” condition, in which the respondent could roll the mouse cursor over a particular word and a definition would pop up, and an “always on” condition, in which the definition was always displayed. While not many respondents in the rollover group requested definitions, those who did spent significantly more time reading them than those in the “always on” condition. This might suggest that when question text gets too long, respondents tend to not process all the information. Additionally, Galesic and her colleagues found the more time respondents spent reading the definition, the more the definition affected their answers.

Although eye tracking can provide valuable information, it cannot be used in production surveys because expensive hardware and software must be installed on every user’s computer. Therefore, this technology and the inferences researchers make from it are limited to laboratory studies. This information is still very valuable to have in pretesting, however it cannot be used to provide respondents with assistance in the field.

1.2.4 Providing Real-Time Assistance

Once survey administrators know that respondents are having trouble with a question, they can intervene and provide the respondents with assistance. In traditional interviewer-administered surveys, respondents can ask interviewers to clarify unfamiliar words or to explain how they are supposed to map their experience onto the answer categories. While strictly standardized interviews do not allow interviewers to answer these questions, conversational interviewing encourages it. Schober and Conrad (1997) argued that conversational interviewing, or specifically allowing interviewers to answer respondent questions, would increase the number of correct responses to questions involving complex mapping. They conducted their study over the telephone, using questions from ongoing government surveys. Respondents answered questions based on fictional scenarios they had received prior to the interview, and there was a known correct answer for each question. As expected, Schober and Conrad found significantly more correct answers to questions involving complicated mappings when interviewers were allowed to answer the respondent's questions.

While Schober and Conrad were able to show the value of conversational interviewing, in self-administered modes there is no interviewer to help the respondent. Therefore, self-administered survey designers spend considerable resources selecting the ideal question wording and developing definitions to help respondents. Ensuring respondents see definitions is important because accuracy can be drastically increased if respondents see and read them (Lind *et al*., 2001; Schober and Conrad, 1997). In these studies, respondents had the option to view definitions, and those that did, answered more accurately. However, only a very small percentage of respondents opted to view the help. Suessbrick and her colleagues (2005) point out that errors due to misunderstanding arise when either respondents do not receive help or the help does not address their specific reason for confusion. Therefore, an extension of this is the hypothesis that if researchers can determine why respondents are confused, they can provide tailored help that will increase the accuracy for the set of individuals that receive it.

The research by Conrad, Schober, and their colleagues, has shown respondents who read definitions consistently report more accurately than respondents who do not. Further, when respondents are not able to obtain assistance, they report lower satisfaction and indicate they would have preferred help to be provided (Conrad *et al*., 2007). Therefore, survey administrators have tried several tactics to encourage respondents to see the information, typically definitions, provided in help links. To date, help in Web surveys has come in one of three forms:

- Always on – Additional text after the question on the original screen
- Respondent initiated help – A link that can be clicked and a window will appear with text
- Model initiated help – An algorithm determines whether a respondent needs help and it appears in a separate window

Each form has benefits and drawbacks, which will be discussed in the subsequent sections.

Always on

Researchers need to balance providing respondents with information they need to answer a question accurately and getting them to notice this information, while not drastically increasing response times. The first option is to simply include any necessary definitions in the text of the question. This way everyone receives the same stimuli so the survey is standardized and, theoretically, everyone will understand exactly what the question is asking. Research suggests respondents do read much of the additional text in the "always on" condition because that condition consistently leads to higher accuracy and longer response times as compared to other conditions. For example, Conrad and his colleagues (2007) found responses were accurate 70 percent of the time when definitions always appeared, as compared with only 58 percent of the time for the next best condition in their study, in which help was provided based on a model that took age into account.

While always providing respondents with definitions increases accuracy, it also increases response times and, in turn, survey length. Specifically, Lind and her colleagues (2001) saw response time increase from 23.4 seconds to 33.5 seconds when they included parts of definitions in the question text. Additionally, in a CATI survey, Schober and Conrad (1997) reported response times (the time it took the participant to hear the question and provide an answer) increased from 16.1 seconds to 60.9 seconds for younger respondents. When definitions were included, response times increased even more for older respondents.

The increase in response time could lead some respondents to abandon the survey. Studies have shown that both survey length and the difference between expected length and actual length can increase breakoffs (Hogg and Miller, 2003; MacElroy, 2000; Peytchev, 2009; Manfreda and Vehovar, 2002). To simulate survey length, Peytchev (2009) and Conrad and his colleagues (2005), in separate studies, used a progress indicator that showed slow progress at the beginning of a survey and compared it to a progress indicator that moved quickly at the beginning. In both studies, respondents who saw the slow progress indicator were significantly more likely to abandon the survey as compared to those who saw the fast progress indicator and the control group. In another study, Crawford (2001) and his colleagues compared informing respondents that a survey would take 8-10 minutes to complete with telling them it would take 20 minutes to complete. While the 8-10 minute condition had a higher initial response rate, it also had a higher breakoff rate than the 20-minute group because the survey took longer than respondents were expecting.

While this type of drop out is a different type of mechanism than what Peytchev and Conrad examined, both types are important to consider when deciding whether to include definitions. Conrad and his colleagues (2005) noted that respondents generate expectations based on their first experience with the survey, and these impressions are generally not changed even if the experience changes. Therefore, if there are a lot of definitions at the beginning of a survey, respondents may perceive it to be lengthy and therefore drop out. On the other hand, many surveys provide respondents with an estimated time to complete. If respondents read definitions carefully, the total time it takes them to complete the survey could be longer than they are expecting or willing to commit, also leading them to break off.

Respondent-initiated

Since there are drawbacks to providing respondents with help whether they need it or not, another option is to let them request help if they feel they need it. There are several ways of allowing respondents to access help. For example, Conrad and his colleagues (2006) compared definitions obtained with a single click, two-click, click and scroll, and roll-over. They found that the more effort required to access a definition, the less likely a respondent was to request it. The rollover condition resulted in 22.4 percent of respondents requesting definitions compared to only 17.4 percent in the single click condition, and even fewer in the two-step conditions. In another study, respondents were asked to answer four questions. When access to definitions only required one click and the help was useful, respondents requested 3.67 out of four definitions. However, when two clicks were required, respondents rarely requested more than one definition, regardless of whether the help was useful or not (Conrad *et al*., 2006). Additionally, response accuracy increased when respondents requested definitions, which suggests they read the text and found it useful. Specifically, Conrad and his colleagues saw accuracy increase from 24 percent to 35 percent for complex mappings when respondents accessed definitions. Additionally, when respondents did access the help, their response times were significantly longer than when they did not access help.

While this research is encouraging in that providing useful definitions to respondents can increase response accuracy, it is even more discouraging that so few choose to consult the definitions. Some respondents that do not choose help most likely do not realize they need assistance. However, others may realize they need the help, but still decide not to access it. One explanation is that help is not on the critical path; that is, it is not an essential step in answering the question. In this situation, respondents would rather move through the survey as quickly as possible and assume their answers are good enough, even if that decision actually leads to reduced accuracy. Another explanation is that people that use the Web for browsing are accustomed to using the Internet to obtain information, while in surveys, respondents provide information. As a result, they may be less likely to seek clarification of concepts than they would in other settings (Schober *et al.*, 2003). Whichever explanation is correct, it is clear from the inaccuracies when respondents are not exposed to help, the low click-through rate is cause for concern, especially if the questions are technical or vague.

Model-Initiated

Those respondents who realize they need help but are not willing to expend the additional effort to get it may display cues that they are having trouble answering a question. With this in mind, researchers have attempted to model respondents' behaviors to predict when they are having trouble answering. In Web surveys, response latencies seem to be the best available predictor of difficulty (Schober and Conrad 1997; Ehlen *et al.*, 2005; Conrad *et al.*, 2006). Using the average completion time for a question, researchers can pinpoint a threshold, which when exceeded indicates the respondent is having trouble answering a question. When a respondent exceeds this threshold, a definition will appear on the screen. This method leads to higher accuracy than respondent-initiated help because it increases the proportion of individuals who see the definition. For example, Ehlen and his colleagues (2005) found, for complicated questions, accuracy increased from 20 percent when no clarification was available to 28 percent when respondents could request information to 64 percent when the model identified those who needed assistance. In addition, when the authors used age-dependent thresholds for activating the help, accuracy increased even more to 71 percent. In another study, Conrad and his colleagues (2006) found accuracy levels of 35 percent for the user-initiated condition, 48 percent for model-initiated, and 58 percent for model-initiated within age groups.

Although model-initiated help improves accuracy, respondents do not necessarily welcome the imposition. Respondents in Conrad and his colleagues' 2007 study were asked to provide satisfaction ratings for the survey. Respondents gave the respondent-initiated help a rating of 3.40 out of four points. However, the model-initiated help only received a rating of 2.89 points out of four. Conrad and his colleagues presume that respondents did not like the system-initiated help because it was unsolicited. Another possible explanation comes from educational psychology. When students receive unsolicited help, they perceive the teacher to believe they have lower ability than other students who did not receive any help. In turn, this leads to a self-perception of low ability (Graham and Barker, 1990). By providing unsolicited help to respondents, survey administrators could be offending respondents by implying they are not capable of completing the survey task. Whatever the reason for the lower satisfaction score, it is of concern because people who do not enjoy the survey-taking experience may be more likely to break off.

1.2.5 Humanization

Increasing satisfaction in model-initiated help is very important to data quality since few respondents request help on their own. While respondents prefer requesting help to receiving system-initiated help, they did report liking the ability to interact with the instrument while taking surveys (Conrad *et al.*, 2007). If model-initiated help were presented in a different way, it is possible respondents would see it as more like a human interaction, which could lead to higher satisfaction ratings and fewer breakoffs. While little research has been done to vary the presentation of the help feature from standard text, there has been research on the effect of varying the presentation of Web surveys as a whole. Specifically, researchers have measured the impact on respondents' perceptions and answers after humanizing Web surveys using audio and graphic displays.

There are two definitions of humanization discussed in the literature: making interfaces more humane, in the sense of easier and more comfortable to use (Laurel, 1990; Shneiderman, 1987) and "humanifying" in the sense of embodying such human-like attributes as speech, speech recognition, and social intelligence (Eichenwald, 1986; Katunobu *et al.*, 1992; Binnick *et al.*, 1989). For the purposes of this dissertation, the second definition will be used, which makes the human-computer interaction more similar to a human-human interaction. This interaction can be enhanced by including features in Web surveys such as animated agents, video, or audio. As these features are more conversational in nature, they should make respondents feel more engaged (Johnston, 2007).

In addition to looking and sounding more human-like than a text interface, these features can also include customized feedback which can enhance a user's experience. Research on using a humanized interaction to provide feedback has focused on education, specifically in motivation and perception of the learning experience. Barlow and his colleagues (1997) used animated agents to provide problem-solving advice to middle school students. They found that the presence of an agent, even if it is not expressive, can have a positive effect on student's perception of their learning experience. Not surprisingly, however, the more lifelike the agent is, the more helpful, credible, and entertaining it is perceived as being.

This type of effect comes from the social agency theory, which is creating a human-like interaction in a multimedia environment (Mayer *et al.,* 2003; Moreno *et al*., 2001). The theory posits that multimedia learning environments can be designed to encourage learners to operate under the assumption that their relationship with the computer is a social one. By using verbal and visual social cues in computer-based environments, people will consider their interaction with the computer to be similar to what they would expect if interacting with another person.

This theory does not require animated agents to be seen. Just adding a voice feature to text can also impact perceived learning. For example, Mayer and his colleagues (2003) compared the social appeal of a human voice to that of a computerized voice. They found participants in the human voice group scored significantly higher on learning performance tests than the computerized voice group and they also rated the speaker more positively. Similarly, Atkinson and his colleagues (2005) experimented with adding either a human voice or a computerized voice to an animated agent designed to teach students a math lesson. The students rated the agent with the human voice more positively than the computerized voice and they also performed better on the practice problems.

Additional research has shown users tend to like environments with agents more than environments without agents (Person *et al.*, 2007). Similarly, Sproull (1996) found that adding more human-like attributes makes interacting with an interface more satisfying, although the exact reason is not known. Further, Tourangeau and his colleagues (2003) conducted an experiment where respondents were either exposed to a picture of the investigator who provided personalized feedback or to no picture and generic feedback. The version with both a picture and feedback was viewed by respondents as most like a conversation and least like interacting with a machine. Taken together, these findings suggest that, in a learning environment, the more human-like an interaction is, the more positively it is perceived and learning itself can be enhanced.

While respondents report enjoying a more human interaction with computers, there could be some negative consequences, especially if the purpose of the interaction is to provide respondents with assistance. Research in human-computer interaction has suggested that computers that exhibit more human-like behaviors are subject to greater social expectations from their users than those with less human-like features. As a result, when computers violate rules of social etiquette, act rudely, or simply fail to work at all, users may become frustrated and angry with the computer (Miller, 2004; Mishra and

Hershey, 2004; Nass, 2004; Reeves and Nass, 1996). In a survey setting, this reaction can lead to increased item nonresponse, satisficing, or breakoffs.

In order to gain cooperation from respondents and make their experience with the computer as pleasant as possible, the communications literature has outlined computer etiquette to enhance users' experience. For example, having an automated response to a computer user will probably not enhance the quality of the interaction, but having the system deliver a response appropriate to the specific user will (Person *et al.*, 2008). In other words, systems that are able to interpret users' needs and exhibit appropriate responses are likely to result in increased usability and productivity. While this is not always possible, small things like personalizing conversations by addressing people by name can also enhance the computer interaction (Mishra and Hershey, 2004).

Web surveys have the capability to engage in a more polite and interactive dialogue with the respondent than paper self-administered questionnaires. Furthermore, they can process information in real-time, which gives them the ability to tailor messages to each individual respondent, depending on his or her needs. Therefore, it is expected that the more human-like the interaction is, the more closely respondents will pay attention to the stimuli provided and the more they will enjoy the overall experience. If this theory is supported and respondents can receive model-initiated help that directly addresses their question, they should be more satisfied with the model's "imposition."

1.2.6 Multi-Modal Presentation

Providing respondents with a more human-like interaction may have additional benefits than just increased satisfaction. Research in multi-modal presentation suggests when respondents are exposed to information using different sensory inputs, such as visual and audio, comprehension can be increased. Additionally, it may also help respondents complete a task more quickly and with more ease.

Most of the research in multi-modal presentation comes out of the education discipline. However, its roots are based in psychology and working memory (Miller, 1956; Baddeley and Hitch, 1974). Due to the constraints of working memory, there is a limit to the amount of information that can be processed at any one time. If a person tries to process too much information, working memory can become overloaded, which can decrease the effectiveness of processing (Miller, 1956).

Despite this limitation of working memory, Baddeley and Hitch (1974) and Penney (1989) proposed models of working memory that could allow more information to be processed at one time. According to Baddeley and Hitch, people take in information using different cognitive storage systems. Specifically, there is a phonological loop that codes vocabulary acquisition, learning to read, and language comprehension. In addition, there is a separate visuo-spatial "sketch pad." This sketch pad processes and stores visual and spatial information. Finally, there is a central executive, which is a control system that coordinates the information coming from the visuo-spatial sketch pad and the phonological loop subsystems. According to Baddeley and Hitch, information can be held in each subsystem separately and simultaneously. Similarly, Penney's model suggests there are auditory and visual processors that independently process incoming information.

In education, information is often presented to students in a way that requires them to split their attention between two pieces of information that are presented in the same mode. This is known as the split-attention effect, which suggests inferior learning occurs when one's attention has to be divided between two information sources within one modality (Ayers and Sweller, 2005). For example, in math classes, students may see a diagram with text that describes it. This presentation may overload working memory because the student needs to go back and forth between the text and the diagram, holding additional information in memory each time.

In order to avoid this split-attention effect, Kalyuga (1999) used Baddeley and Hitch's and Penney's working memory model to hypothesize that the amount of information that can be processed using both auditory and visual channels might actually exceed the processing capacity of a single channel. Specifically, Kalyuga presented three groups of subjects with instructional materials. The first group received visual text and audio, the second received only visual text, and the third received only audio. The findings indicate the participants in the audio and visual text group had a lower number of reattempts in interactive exercises, higher performance scores, and reported a lower cognitive load than the other two groups.

With new computer technologies, these same principles can be extended to help with working memory overload in Web behaviors on computers and even mobile devices. Specifically, Fang and his colleagues (2006) experimented to see whether short audio narrations interfered with visual processing of textual information on a Web site. Prior research using computer interfaces focused on presenting the same information through audio and visual channels (Archer *et al.*,1996). Fang and his colleagues, on the other hand, were interested in how additional information presented in a different mode, either helpful or irrelevant to the task at hand, was processed.

In their experiment, Fang and his colleagues had participants browse a Web site containing course curriculum. There were three experimental conditions: visual presentation only, visual and audio, and visual and assistive audio. Participants were to browse the page and find the information necessary to answer a series of questions. Their first hypothesis was that if participants were exposed to irrelevant audio, they would successfully receive the auditory information and it would not negatively influence their processing of the visual information. As expected, they found participants were able to receive the auditory information without hurting overall processing. Specifically, participants answered just as many questions correctly when exposed to the visual and irrelevant audio as they did with just the visual. In addition, the participants did successfully receive the audio message while completing the task, which was measured by recognizing words from the audio.

In addition to measuring how people handle irrelevant additional information from a different sensory mode, Fang and his colleagues also hypothesized helpful information presented through the auditory channel during the Web-browsing process can be received by users and would improve their performance on browsing tasks. They found participants in the visual and assistive audio group had higher accuracy and also spent significantly less time on each question as compared to the visual group.

Web browsing, education, and survey taking are very different tasks. However, all involve searching for specific information from a list of text. Therefore, it is plausible these same principles of dual modality can be extended to the survey methodology field, specifically in the form of providing help. Since help is currently displayed in a text format, respondents may be experiencing a split-attention effect as they go back and forth from the help text to the question trying to understand what is being asked. However, if the help were auditory, respondents would be able to process both the words in the question and the help at the same time, thus decreasing the burden on working memory and leading to greater understanding and accuracy.

1.2.7 Summary of the Literature

Long before survey methodology became a discipline, researchers had studied how we communicate with each other. This research has included our facial expressions, levels of eye contact, the questions we ask, and our speech patterns. Originally, this information was used in communications and education to determine whether a conversation partner was confused about some aspect of the exchange. As survey methodologists came onto the scene, they adopted this literature and expanded it to assess when respondents were having difficulty answering a question.

Using these cues, interviewers were able to determine when respondents were having difficulty answering a question and could provide them with the appropriate help if trained to do so. However, with new technology and increased costs associated with interviewer-administered questionnaires, survey administrators are moving away from interviewer-administered modes, requiring new ways to assess when respondents are confused by a question.

With Web surveys becoming more popular, survey researchers have attempted to identify indicators of difficulty to use in lieu of speech disfluencies available in interviewer-administered surveys. Thus far, researchers have focused on response latencies and eye movements. While response latencies allow administrators to provide real-time assistance as in the interviewer-administered modes, they do not provide any information regarding specifically why a respondent is having trouble. Additionally, following eye movements requires hardware to be installed on a computer, limiting its usefulness outside of a laboratory setting.

While providing assistance to respondents increases accuracy, it is not clear if respondents like receiving help, especially when it is unsolicited. A handful of studies have looked at different ways to provide respondents with help (always on, respondent-initiated single click, respondent-initiated two-click, respondent-initiated click and scroll, generic model-initiated, etc.) and different ways to humanize the survey taking experience (animated agents, audio, pictures). While respondents may enjoy taking surveys with these features, there is no evidence the added humanization helps meet survey goals.

Although researchers have independently shown that respondents do not like unsolicited help and do like humanized experiences, none of the research has focused on how respondents are provided with help, which is always presented in a standard text format. Additionally, the goals for providing help are different than those for completing an entire survey. If respondents like the heightened interactivity and personalization, it should make them pay more attention to the help and hopefully more tolerant of the unsolicited imposition. Together, this will ideally result in higher accuracy and respondent satisfaction.

Chapter 2: Identifying Mouse Movements People Make while Completing Surveys

2.1 Background

Researchers have used different methods over the years to identify respondents that are having trouble answering questions in interviewer administered and automated surveys. However, this study attempts to find a new way to identify these respondents in Web surveys. In Web surveys, as discussed in Chapter 1, looking at data such as response latencies can help researchers detect respondent problems and identify questions with which respondents are struggling. Although not practical in a production environment, eye tracking data are among the most valuable data that can be collected to help understand what respondents are thinking. However, its limited use has led researchers to seek other technologies that may be able to provide the same information on a larger scale for less cost.

Web designers that use eye tracking are faced with a similar issue as survey researchers; they can examine eye movements in the lab, but not while real users are browsing their Web sites. Therefore, Web designers started looking at mouse movements as an inexpensive alternative; tracking mouse movements only requires JavaScript, which is simply implemented as an integrated component of the Web browser on the user's computer. Additionally, the mouse cursor is the most widely used instrument to browse Web pages, and while it is relatively difficult to control gaze position, controlling cursor position is quite easy (Leiva, 2011).

Preliminary research in mouse tracking has focused on the different movements people tend to make. In general, Web browsers tend to either use the mouse as a reading aid or move the cursor to a blank or scrolling area of the screen (Arroyo *et al.*, 2006).

Additionally, they found the trajectory of mouse movements can have importance; direct movements indicate purpose and meaning, whereas slow, arched trajectories indicate more uncertainty. A similar result was also seen by Huang and his colleagues (2011). They examined good and bad abandonment of Web searches; good abandonment being when the excerpt below the result link provided adequate information and bad abandonment being when none of the results were satisfactory. Compared to good abandonment, users that left due to bad abandonment traveled an overall longer distance with the mouse, they spent more time moving the mouse, and they moved it at a slower speed.

The Web browsing literature has also identified specific types of mouse movements users exhibit while looking at a Web page. This literature differentiates between informational and navigational queries (Guo and Agichtein, 2008). With navigational queries, users tend to go directly to the result they are interested in and do not spend much time browsing the other options. On the other hand, with informational queries, users spend more time reading through the other options, which results in more total mouse movements. The research in this study focuses on informational queries, as they are the most similar to surveys in that respondents sort through the response options for information that best fits their situation.

By tracking mouse movements and eye movements simultaneously, researchers can see that in informational queries few users use the mouse to follow along with their eye to read horizontally; rather a larger proportion use the mouse to read vertically (Rodden *et al.*, 2008). That is, the mouse acts as a type of pointer to help the user keep track of what line they are on while the eyes continue to read horizontally. This behavior of using the mouse to track focus of attention was also seen when users were multi-tasking between different browser tabs. Users tended to highlight the last sentence they were looking at so they would know where to come back to (Leiva, 2011).

In addition to using the mouse to keep track of where a user is on a page, it can also be used to indicate interest. Researchers at Microsoft examined the implication of hovers on search engine results pages (Huang *et al.*, 2011). They saw considerable mouse activity on options that we not clicked. Additionally, prior to clicking a result, users hovered on surrounding results. They suggest this could mean the other results were being considered because observing one or more unclicked hovers dramatically increased the likelihood the considered link would be clicked at some point. This interpretation is supported by research on extracting relevant text from search results (Hijikata, 2004). By tracking vertical reading, markers, text selection, and clicks, Hijikata was able to extract more keywords the user was interested in as compared to a random extraction.

A similar result was seen by Rodden and her colleagues (2008). They found respondents used the mouse as a marker to note something they wanted to come back to while they continued browsing with their eyes. Google is using this information to improve their search relevancy scores by predicting what a respondent's second choice would be while they are performing a search (Guo and Agichtein, 2008). While informational Web browsing and survey taking are very different, a list of response options is similar to a list of search results. Respondents need to skim through the list until they find the most suitable option. Therefore, this type of information could be used to determine whether a respondent is having difficulty deciding between two response options.

Since none of these actions have been studied in a survey setting, I tracked participant engagement in a set of indicators[2] in usability testing for the ACS Web instrument. During November and December of 2010, 19 participants came to the Census Bureau's usability lab to complete a Web version of the ACS. I watched recordings of these sessions and noted each time a respondent displayed any of the following actions:

- Horizontal reading – using the cursor to follow their eyes while reading
- Vertical reading – keeping the cursor aligned to the left while reading to keep track of which line they were on
- Hover – holding the cursor over any of the following for an extended period of time: question text, previous button, or next button
- Highlighting – clicking, then dragging the cursor over a word or phrase and then releasing
- Marker – holding the cursor over one of the response options for an extended period of time
- Regressions[3] or backward movements – moving the cursor from one response option to another or moving the cursor from the response options to the next button, question text, or white space and back again

While reading the questions, all but two respondents kept the cursor in white space and then moved it towards the response options when they were ready to answer. The other two respondents read each question horizontally with their mouse. For very simple and straightforward questions, such as "Do you live at 100 Main Street?", respondents did not exhibit any of the behaviors I was looking for. They simply moved the cursor from its resting place (which was either in the white space or over the 'Next'

[2] This list was generated from Rodden *et al.*, 2008 and Mueller and Lockerd, 2001
[3] In a survey context I will refer to regressions as any instance moving back and forth with the mouse between lines that have already been processed (as compared to Liversedge and Findlay, 2000, who define regressive eye movements as rereading an earlier sentence)

button from the previous screen) directly to the response option they selected. However, for questions with many response options, they were more likely to use their mouse to read vertically through the options. Further, respondents tended to use their mouse as a marker while they scanned other options and also used it to go back and forth between options they seemed to be considering. While the usability testing was not a formal study, it does offer support that the same actions seen in Web browsing also appear in survey taking.

Another observation I made, which was also noted by Arroyo and his colleagues (2006), is the speed at which respondents move their mouse. On straightforward questions for which respondents know the answer and do not need to think about, they moved their mouse quickly and directly to the proper response option. However, on questions where they were considering multiple response options, respondents were more likely to move the cursor much more slowly and less directly to the response option. Using the speed of movement as an indicator could shed more light on when respondents are having difficulty.

While researchers have identified a set of movements that Web users make with their mouse, we still do not know exactly what they mean. Researchers at the Massachusetts Institute of Technology have been able to predict, with between a 65 percent and 75 percent success rate, what a Web browser's second choice would be in scanning a search query (Mueller and Lockerd, 2001). However, the model they used to generate their predictions is not public information, so it is not clear if they used more information than just the mouse movements. To make the information obtained from mouse movements more complete, we need to combine them with the eye-tracking data, as was done in Redline and her colleagues' 2009 experiment to determine exactly what respondents are focusing on when displaying these behaviors. Therefore, this study aims to identify a set of mouse movements in which respondent engage and form hypotheses regarding which are related to difficulty.

2.2 Study Design

This study involved 30 participants answering 20 questions from the ACS. Participants with unique living situations were recruited to ensure some questions would be difficult to answer. We identified difficult questions by asking participants to rate each question's difficulty and also through debriefings after they completed the survey. Each participant's session was recorded so we could later review the recordings and code a set of predetermined mouse movements. This section describes the individuals that participated in the study, the data collected, and the procedures used to collect the data and code the mouse movements.

2.2.1 Participants

The recruitment goal for this study was to interview 40 participants from the Washington, DC metropolitan area. Participants were recruited using both flyers posted around the city and print advertisements on Craigslist in September, 2011 (Appendix A). In order to capture a wide range of individuals, the following criteria were used as a guide for recruiting:

- 12 participants age 18-34
- 12 participants age 35-49
- 12 participants age 50-64
- 4 participants age over 65
- 6 participants in a same-sex co-habitating couple
- 5 participants in a same-sex co-habitating registered domestic partnership
- 2 participants in a legally married same-sex couple
- 15 participants living in a building with more than 5 units

- 8 participants with a child in a joint custody arrangement
- 8 participants that have taken a short course in the past 2 months

This variety in participants was desired for several reasons. First, it has been noted in response latency studies (Ehlen *et al*., 2005) that response times increase with age. It is also possible movements differ by age as well. This could be because younger people have been using computers since they were very young, so they interact with them differently than older respondents, or it could merely be a function of speed decreasing with age. Therefore, recruiting participants with a wide range of ages will help control for these differences. Secondly, the questions ask participants about their own lives, so it cannot be guaranteed they will have trouble answering any of the survey questions. Therefore, the other criteria attempt to identify individuals with atypical living situations to ensure difficulty on at least some of the questions. These atypical situations were identified from ACS pre-testing and usability testing.

After two weeks of recruiting, the final set of participants consisted of 30 individuals[4]. These participants ranged in age from 22 to 64 and their education ranged from some high school to a post graduate degree. Table 2.1 provides a breakdown of the demographic characteristics of the participants. Each participant answered a series of screener questions to determine eligibility based on the criteria outlined at the beginning of this section (Appendix B).

All of the participants were asked to come to the Joint Program of Survey Methodology (JPSM) office in College Park, Maryland to complete the survey and they were reimbursed $30 for their efforts.

[4] We were unable to recruit 40 participants with the required criteria within the two-week time frame under which we were working.

Table 2.1 Demographic Characteristics of the Participants

Gender	**Number**	**Percent**
Male	15	50.0
Female	15	50.0
Age		
18-34	14	46.7
35-49	6	20.0
50-64	10	33.3
Education		
Some High School	1	3.3
High School Graduate or equivalent	3	10.0
Some college	9	30.0
Associate's Degree	3	10.0
4-year college Degree	8	26.7
Some Graduate school	3	10.0
Post Graduate Degree	3	10.0
Race		
Black	14	46.7
White	11	36.7
Other	5	16.7

2.2.2 Web Survey

The data for this study consisted of 20 questions taken from the Web version of the ACS, which was being tested in 2011 and 2012 at the US Census Bureau (Appendix C). These questions were selected because in the usability tests for the 2011 ACS Internet Test they elicited a greater number and a larger variety of mouse movements than other questions and in debriefings some respondents reported being confused by the questions. Additionally, some of the questions are similar to those used in Schober and Conrad (1997), which have already been tested for the purposes of manipulating question difficulty. The questions in this study used the same wording as the ACS instrument; however we did not include the additional instructions and help content. The lack of clarifying information should make the questions more challenging for the participants, which should lead to more confusion and ideally more mouse movements.

In addition to the 20 ACS questions, we asked participants a series of demographic questions, such as gender, age, race, education, and computer experience. Since participants answered autobiographically, the researcher did not know which questions might be problematic. Therefore, following each ACS question, we had participants rate the difficulty of the ACS question on a one to five scale (these rating questions will be referred to as probing questions). Finally, participants were debriefed after they had completed the survey to determine whether they had difficulty answering any of the questions. The debriefing protocol can be found in Appendix D. The debriefing reports, along with the difficulty ratings, were used to identify questions that respondents had difficulty answering.

2.2.3 Data

Eye tracking and mouse movement data were collected on a computer equipped with Tobii's 1750 eye tracker and the corresponding Studio Professional Edition software. This software collects data points every 16.7 milliseconds and has a frame rate of 50 Hz, which means it takes 50 pictures of the eye per second. The Tobii software is very diverse and has tools to assist researchers with test design, recording, and analysis. Researchers can define areas on the screen that are of interest and will receive summary statistics, including time to first fixation, fixation duration, visit count, percentage fixated, time from first fixation to next mouse click, time to first mouse click, and mouse click count, for each defined area. Researchers also have the ability to define and code their own events that are not part of Tobii's default settings. By defining a coding scheme, specific actions of interest to the researcher can be output to an events log. This coding scheme was used to capture and log all of the specific mouse movements.

Tobii Studio organizes data within projects. Once the project is complete, all of the data that Studio collects can be exported to an Excel file. The file summarizes all the eye movements for each eye, including a latitude, longitude, and time stamp for each gaze, fixation and mouse click. Additionally, any data that were separately coded are also included in the export, along with any notes that accompany the data.

In cases where the detailed statistics miss the big picture, Tobii also has the ability to visually summarize these movements for researchers. Specifically, it can replay each participant's session with or without an overlay of the eye movements, which is how the individual mouse movements were coded.

2.2.4 Procedure

Data Collection

Eligible participants came into the JPSM office and sat in front of the Tobii eye tracking monitor to take the survey. They listened to an introduction and background to the study (Appendix D) and then read and signed a consent form and answered a brief questionnaire covering their demographic information and computer and Internet use (Appendix E). Finally, their eyes were calibrated to the eye tracker. After the calibration, the researcher logged participants into the survey and told them to answer the questions as if they were at home by themselves. The researcher then left the room while the participants completed the online survey. From a different computer, the researcher was able to view participants' computer screens while they completed the survey, so she was able to take note of any irregular behaviors to ask about during the debriefing.

Upon completion, the researcher returned to the room to debrief the participant. Each participant was asked if there were any questions he or she had difficulty answering, if there were any questions where there was a response option he or she expected to see but was not there, or whether there were any questions he or she had trouble choosing between two response options. Additional probes were provided for participants who said none of the questions were difficult. In order to facilitate recall, the researcher presented a PowerPoint presentation with all of the survey questions so the participant could scroll through and refer back to specific questions instead of trying to pull from memory. Finally, in cases where the researcher took note of a particular behavior, she asked the participant to engage in a retrospective think-aloud[5].

To do the retrospective think-aloud procedure, the researcher opened the Tobii software replay and had the participant watch themselves respond to specific questions. They were asked to describe or explain what they were thinking about as they answered the question. When participants fell silent, the researcher probed to gain a description of their actions. None of the participants had a problem with this exercise and all were able to explain what they were thinking about while answering the questions.

Mouse Movements

Prior to bringing participants into the lab, 11 different mouse movements were identified that were of interest to the researcher (Table 2.2). These movements were based on the research conducted by Rodden *et al.* (2008), Guo and Agichtein (2008), Arroyo *et al.* (2006), Huang *et al.* (2011), Leiva *et al.* (2011), and also from the 2011

[5] Retrospective think-alouds were only captured for the second half of the respondents. The original idea was to have a sample of five to ten respondents engage in a typical think-aloud while taking the survey. However, there was a concern that this behavior would negatively impact their engagement with the mouse. Therefore, the retrospective think-aloud was used, but the idea was not formulated until part way through the data collection.

ACS Internet Test usability sessions. Each movement was added to Tobii Studio's coding scheme feature with a shortcut key to identify it. For example, a Hover was defined as F1 in Tobii Studio. Any time F1 was clicked during the replay mode in Tobii, a Hover was recorded.

Table 2.2 Description of Coded Mouse Movements

Definition	**Description**
Click	Mouse button is clicked on a word (not associated with a highlight, answer selection, or missing the radio button)
Highlight	Mouse is used to click and then highlight all or part of the question or response answer text
Horizontal reading	Mouse follows along with the eye to read along from left to right
Hover	Mouse stays over the 'Next' button or the question text for more than 2 seconds
Intent	Mouse movement from question resting place to response option, or from response option to 'Next' deviates from typical behavior (slower, hesitation, looping, indirect)
Marker	Mouse stays over a radio button or response option text for more than 2 seconds
Response-to-Response	Mouse moves back and forth between two response options.
Response-to-Question	Mouse moves from response option to question text and back again one or more times
Response-to-Next	Mouse moves from response options to 'Next' button and back again one or more times
Response-to-Space	Mouse moves from response options to white space and back again one or more times
Vertical reading	Mouse is used to follow the eye from top to bottom or bottom to top

Most of these actions involve specific movements with the mouse that can be clearly defined. However, Intent is more subjective. The idea of Intent comes from research by Arroyo and his colleagues (2006), who suggest the more direct a movement is, the more certainty the user has. Therefore, Intent is intended to measure when respondents behave more hesitantly, or less directly, in answering a question than they normally do.

In order to track each time one of these actions occurred, two students (one undergraduate student and one Master's student, both at the University of Maryland) were hired. The students watched the recordings from the Tobii Studio replay mode and logged, using the shortcuts, each time each movement occurred. Additionally, they could provide notes to the researcher, such as how long a Hover or Marker lasted. Two coders were used so there would be a measure of reliability. Once the students completed their coding, the researcher analyzed each instance where the coders did not agree and made a final decision. In borderline cases, the two students and the researcher sat down together to discuss the movement in question to come to an agreement.

2.3 Analysis

The crux of this research is identifying what movements people make with the mouse when responding to surveys online. Therefore, our analysis starts with determining the reliability of the coders. We then describe the techniques used to measure the different movements observed throughout the survey and which might be related to difficulty. We also describe how we examined participant's focus of attention while engaging in different movements and how long it took them to respond when movements were and were not present. Finally, we use the information from the debriefings to link specific movements to types of difficulty.

2.3.1 Reliability

To measure reliability, we assigned all instances where the coders matched a value of '1' and all disagreements a value of '0.' We then used the Test Kappa function in PROC FREQ in SAS® to compute the kappa values which can be used to measure the strength of agreement between the two coders. Specifically, Landis and Koch (1977) proposed the following kappa values as standards for strength of agreement: ≤ 0 = poor, .01-.20 = slight, .21-.40 = fair, .41-.60 = moderate, .61-.80 = substantial, and .81-1.00 = almost perfect. We compared the strength of agreement between the two coders overall and by movement.

2.3.2 Movements Observed

This study was intended to be exploratory in order to understand the types of mouse movements people engage in while answering survey questions online. The ACS Internet Test usability sessions identified two general responding behaviors: Typical and Readers. Typical respondents kept the cursor in white space while they read the question and did not move it until they were ready to answer the question. Readers used the mouse to follow their focus of attention. In other words, using the mouse, they followed along horizontally as their eyes read the question text and response options. We assessed both types of general behavior since they influence what other movements can occur. For example, if a Reader is answering a question, it is more difficult for them to engage in other mouse movements because they are always moving the mouse with their visual focus.

Once we gained a general understanding of how people answer questions, we focused on what specific mouse movements they used. To begin with, frequencies were calculated for each movement. Frequencies provide information on which of the expected movements participants actually engaged in. Additionally, they provide preliminary insight into which movements might be related to difficulty. For example, movements that are very common may not be as related to difficulty as more uncommon movements because it is unlikely participants had difficulty with every question. In addition to the raw frequencies, we also calculated the percent of screens, across all participants, on which each movement occurred.

To understand the significance of the different movements, it is important to examine the probing questions and ACS questions separately. This can help differentiate between general behaviors and behaviors associated with difficulty or confusion. Specifically, we assumed the probing question was not hard for respondents to answer, especially since they saw it 20 times throughout the survey. Therefore, we suspect that actions that occurred regularly for the probing questions are less likely to be indicative of confusion. For example, Response-to-Response is defined as moving the cursor back and forth between two response options. At first glance, this movement might reflect difficulty deciding between two response options. However, in the case of the probing question, it is more likely the respondent is just determining the best fit and there is no actual confusion. Therefore, there are at least two different reasons respondents might engage in this behavior, one indicative of difficulty and the other not. To help identify the movements that are more likely to be associated with difficulty, we separately calculated the percent of screens on which each movement occurred for ACS questions and probing questions.

It is also important to identify who is making these different movements and how often. It is possible only one participant displayed one of the movements, but they did so 20 times, making the movement seem more important and predictive than it actually was. Multiple participants need to engage in each movement for it to have greater meaning to future research. Therefore, we analyzed the number of participants that engaged in each movement, the average number of times each participant engaged in each movement and the maximum number of screens on which a single participant engaged in each movement.

Once we know what movements are being made, we need to understand who is making them. Specifically, we need to know if all of the participants are making each of the movements or if certain demographic groups make specific movements. Therefore, we compared the frequency of movements that participants made by gender, age, race, and education to see if any group was more likely to engage in specific movements than another group. The frequency was calculated by dividing the total number of instances of each movement in each group by the number of participants in each group that engaged in the movement.

A final measure we examined to understand the movements participants made was where in the instrument the movements occurred. It is possible participants engaged in more movements at the beginning of the survey as they learned how to select answers and proceed through the instrument. However, once they grew accustomed the process, they may not have engaged in as many movements. Therefore, we tracked the number of movements that occurred on each screen and compared them sequentially.

While it is valuable to understand what movements the participants made and who made them, the question we are most interested in answering is which movements are likely related to difficulty. To begin to answer this question, we calculated the average difficulty rating on screens where each movement occurred. We calculated this measure in two ways: probing questions and verbal reports. First, for each movement, we calculated the mean rating from the probing questions corresponding to ACS questions where the movement was present. If multiple movements occurred on a single question, the rating was included in the calculation for each movement.

The researcher noted, while remotely watching the participants complete the survey, that many participants did not report much differentiation between the ratings to the probing questions. Therefore, during the debriefing, the researcher asked if there were any questions the participant thought were difficult. A second comparison of difficulty was thereby created which measured the frequency at which each movement was associated with a verbal report of difficulty. Specifically, we compiled all of the difficult questions and calculated the instances each movement occurred on those questions as a percent of the total number of times each movement occurred.

In addition to comparing the two measures separately, we can also combine them to create an index of difficulty. Specifically, instead of a categorical scale of difficulty, we created a binary index variable equal to '1' if the question had a verbal report of difficulty or the question was rated as 'Neither easy nor difficult,' 'Somewhat difficult,' or 'Very difficult.' Otherwise, the index variable was equal to '0.' Using this new variable, we calculated the number of times participants engaged in each of the movements on a difficult screen, that frequency as a percent of all movements on difficult screens, and that frequency as a percent of the number of the total number of difficult questions. This analysis can tell us which movements are occurring on difficult screens and also on how many of the screens they occur. This gives us a broader understanding of which movements might be associated with difficulty than just looking at the probing questions and verbal reports separately.

It is also possible that it is not a single movement that is indicative of difficulty, but multiple movements that occur on the same screen. For example, it is possible that Hover by itself is not related to difficulty, but if Hover and Marker occur on the same question, the combination of the two movements might be. Therefore, we compared the average rating to the probing questions when there were no movements to the ratings when there were movements. We expect ratings to increase with the number of movements, which would provide further support that the movements are related to difficulty.

2.3.3 Focus of Attention

While examining frequencies and ratings from the probing questions provides some hypotheses as to what the different mouse movements mean, knowing the participants' focus of attention may help explain what they were thinking about while exhibiting the behaviors. This, in turn, could help inform whether different types of difficulty are associated with different movements. To determine the focus of attention, eye movements captured by Tobii were examined while the participants were engaging in the different mouse movements. Unfortunately, in this study the quality of the eye tracking was only usable for half of the respondents (n=15) because participants either leaned too far forward or to the side while completing the survey, which was out of range to capture their movements. Areas of interest were used to classify the eye movements. Specifically of interest was whether the participant was focused on the question text, the response options, or the 'Previous'/'Next' buttons. Focus of attention was measured by identifying the different areas of interest the participant fixated on while engaging in each mouse movement. For example, if a participant hovered over the 'Next' button for three seconds, each area the eye fixated on in that same time frame was captured. Using this information, we calculated the percent of the time participants' eyes focused on each area of interest for each movement.

2.3.4 Response Time

The primary measure of respondent difficulty used in past research is response latency. In computer assisted telephone interviews, IVR, and Web surveys, various researchers have shown that respondent difficulty is associated with longer response times (Ehlen *et al.*, 2005; Heerwegh, 2003; Redline *et al.,* 2009; Conrad *et al.,* 2007). It is expected that if mouse movements are associated with difficulty and difficulty is related to longer response times, response times will be longer for questions on which movements are displayed.

Therefore, median response times were calculated for each question when no movements were used and were compared to median response times for each question in which participants engaged in any single movement and multiple movements. All of the probing questions were calculated together instead of separated by question because the occurrence of movements was rare on these screens. Additionally, the probing questions can be used as a benchmark for the additional time spent just due to the act of making a movement because none of the participants should have had difficulty answering these questions.

2.3.5 Debriefing Explanations

Response times have already helped researchers identify difficult questions. Additionally, these times are easy to calculate and implement. Therefore, in order to provide more information than we already have, mouse movements need to be able to provide additional information; identify the type of difficulty respondents are having. Therefore, after completing the survey, the participants were not only asked what questions were difficult for them to answer, but also what it was about the question that was difficult. We coded their responses into seven different categories:

- Looking for a specific response option
- Don’t know the answer to the question
- Difficulty mapping situation to response options
- Comprehension (understanding question/response options or what to include/exclude)
- Re-reading or double checking
- Thinking
- Unsure of best response option

Using these descriptions, we first tallied the number of times each movement occurred on a question associated with each of the seven explanations. Then, to calculate the percent of explanations associated with each movement, we divided the tallied number by the total number of explanations provided within each movement. For each movement, we focused on the explanation that received the highest percent of reports. This measure will help inform whether a specific movements is linked to a specific explanation, if many different movements are linked to specific explanations, or if there is no relationship at all.

Considering participants could engage in multiple movements and provide multiple explanations on a single question, we needed to account for this in our coding. In cases where there were multiple movements linked to a single explanation, the explanation was applied to each movement. In cases where participants provided multiple explanations, we used the explanation that was provided first. However, participants only provided multiple explanations in 3.4 percent of cases (4 instances out of 119 total explanations). Although this coding is not clean in that it isolates explanations and movements, it is the best we could do with the information provided and it does start to give us a sense of why participants engaged in the different movements.

2.4 Results

We first examine the reliability of the coders overall and by mouse movement. We then describe the typical responding behavior of the participants, how frequently they engaged in each movement, and which movements were associated with higher difficulty ratings. We also determine whether participants looked at the response options, question text, or both while engaging in each movement and whether response times increased when movements were present. Finally, using the information obtained in the debriefings, we look at the most common explanation given when participants engaged in each movement.

2.4.1 Reliability Results

The reliability between coders was calculated overall, as well as by movement (Table 2.3). The overall agreement between the coders, across all movements and questions, was a kappa of 0.54, which is considered moderate (Landis and Koch, 1977).

Table 2.3 Coder Reliability by Mouse Movement

Movement	Kappa	Strength of Agreement
Horizontal reading	0.79	Substantial
Response-to-Next	0.78	Substantial
Marker	0.70	Substantial
Hover	0.66	Substantial
Vertical reading	0.54	Moderate
Response-to-Response	0.44	Moderate
Response-to-Question	0.29	Fair
Response-to-Space	0.26	Fair
Intent	0.12	Poor

Intent was the only variable with poor reliability. We decided to exclude this movement from further analysis because it was not reliable enough from which to draw meaningful conclusions. Other than this movement, reliability was acceptably high and disagreements were easily resolved to produce a single dataset containing all the mouse movement observations.

2.4.2 Movements Observed Results

Of the predefined movements (Table 2.2), there were two movements which were not displayed by any of the 30 participants: Highlight and Click. Highlighting sections of text was seen in education literature to highlight important words or phrases. It could be that the question text for surveys is short enough that this type of behavior is not necessary. The hypothesis that respondents might make extraneous clicks on unknown or difficult words or phrases was not based on any prior research, so it is not surprising this movement was not observed. All of the other movements were observed multiple times by multiple respondents.

As the research suggests, the typical responding behavior observed for most participants was to place the mouse cursor in white space while reading the question and then move it towards the response options when they were ready to answer. Occasionally participants would engage in atypical behavior, such as those identified in Table 2.2. Since these actions are not part of their normal answering routine, it is possible there is an important meaning behind the change in behavior.

While most participants were considered Typical, 20 percent were identified as Readers. This is surprising because prior research suggests this is a rare occurrence. There did not seem to be any consistency in demographic characteristics across these participants. The only common factor was that they all had high levels of computer experience. Otherwise, their ages ranged from 20s to 50s and their education ranged from some college to a professional degree. One additional consideration was that poor eyesight might be related to reading along with the mouse. Only two of the six wore glasses, though it cannot be said how many wore contacts. From the data collected in this study, this type of behavior does not appear to be related to any particular demographic group of people, but rather may have to do with other factors such as attention to detail. However, since these participants behaved similarly for every question, including the probing questions, their behavior does not seem to have additional meaning nor be associated with difficulty; it is just an individual behavior pattern.

Now that we understand the general behavior the participants engaged in to answer the survey questions, we analyzed deviations from that behavior. Table 2.4 provides the percent of screens on which the different mouse movements were observed. The 'Percent of all Screens' column provides the percent of all the screens on which each movement occurred, whereas the rest of the table separates the ACS questions from the probing questions. In total, there were 1,204 screens observed; 603 were ACS questions (approximately 20 questions for each respondent) and 601 were probing questions.

The data demonstrate that the most prevalent behaviors were Marker, Horizontal reading, and Vertical reading. This is not surprising because these three behaviors are the best way for a person to keep track of where his or her focus of attention is.

None of the other actions were as prevalent, but this is not necessarily a bad result. It is clear from the probing questions and debriefings that participants did not experience difficulty on every screen. In fact, most participants only had trouble answering a handful of questions. Therefore, since one of our goals is to identify movements that are related to difficulty, the low overall prevalence rates provide differentiation between questions.

Table 2.4 Frequency and Percent of Screens on which Participant Engaged in each Mouse Movement

Movement	Percent of all screens	Frequency on ACS Screens	Percent of ACS screens	Frequency on probing screens	Percent of probing screens
Marker	12.73	147	24.38	8	1.33
Horizontal reading	12.23	116	19.24	33	5.49
Vertical reading	9.85	114	19.07	6	1.00
Hover	5.83	68	11.28	3	0.50
Response-to-Response	3.94	36	5.97	12	2.00
Response-to-Question	3.86	34	5.64	13	2.16
Response-to-Next	2.13	25	4.15	1	0.17
Response-to-Space	1.23	11	1.82	4	0.67

Before we can determine what the movements mean and which may be related to difficulty, it is important to see how many participants engaged in each movement in addition to how many total occurrences there were. To ensure the movements described were common across participants, Table 2.5 examines the number of participants that engaged in each movement at least one time. Additionally, the table shows the average number of screens on which participants engaged in each movement per participant and the maximum number of questions on which a single participant displayed each movement. Participants who engaged in a typical white space to response option to ‘Next’ button pattern were not included in this analysis; rather it focuses on atypical movements.

Table 2.5 Frequency of Participant Engagement in each Mouse Movement

Movement	Total Number of Participants	Percent of Participants	Average Number of Questions/ Participant	Maximum Number of Questions
Marker	30	100.00	5.17	15
Vertical reading	28	93.33	4.36	8
Response-to-Response	25	83.33	1.92	4
Hover	24	80.00	2.96	6
Response-to-Next	20	66.67	2.35	8
Response-to-Space	17	56.67	1.53	3
Horizontal reading	17	56.67	8.76	20
Response-to-Question	8	26.67	1.88	3

Table 2.5 shows that every participant engaged in at least one of the movements. For example, the first row quantifies the participants that used the mouse as a marker. In the ‘Total Number of Respondents’ column, we see all 30 participants (or 100 percent in the third column) used the mouse as a marker at least one time. The average participant engaged in this movement on 5.17 questions throughout the survey, and one participant used the mouse as a marker on 15 different questions.

Overall, more than half the participants engaged in every movement other than Response-to-Question. Additionally, the average number of questions per participant is highest for horizontal reading. This is not surprising because, as previously stated; six individuals were classified as Readers, which means they did this for almost every question. Similarly, Marker and Vertical reading also had higher averages per respondent, but almost every participant engaged in these movements. This means that some participants displayed these movements more often than others, which is a within person characteristic, but the movement was common enough across participants that between-participant effects can be measured. Therefore, there does not seem to be any cause for concern that these movements are only within-participant behaviors.

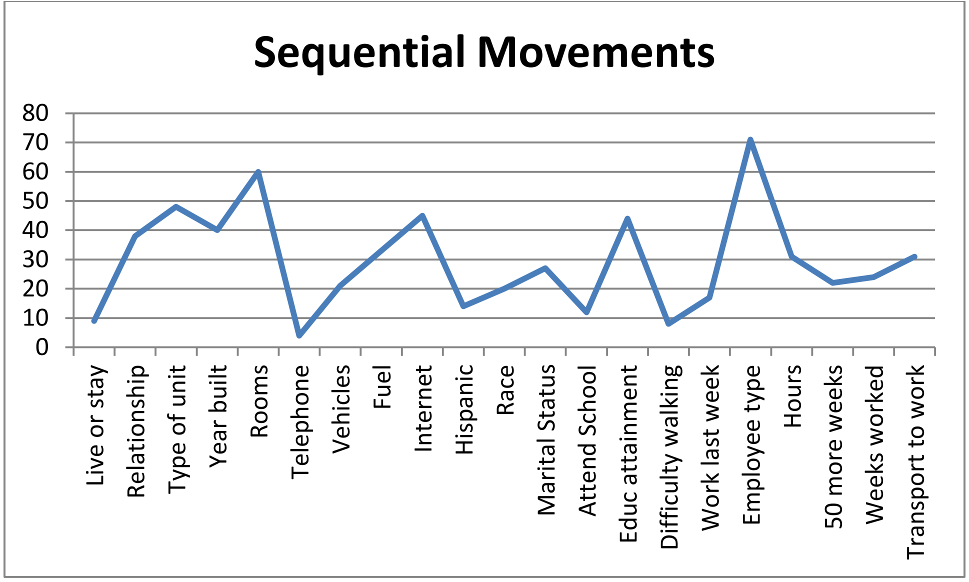
While many participants engaged in each of the movements, an additional question is whether these movements are related to age or education, as the response latency research suggests (Elen *et al.,* 2005). For most movements, we found minimal differentiation between age groups (18-34, 35-49, 50-64) and level of education (some high school, high school, some college, Associates degree, Bachelors degree, some graduate school, graduate degree) in how frequently they engaged in each movement. However, a few patterns did stand out, although their interpretation is unclear. From the 30-person participant pool for this study we saw:

- The oldest age group (50-64) read vertically less frequently than the other two age groups (3.3 instances per participant compared to 4.8 for 18-34 year olds)
- Participants with a graduate degree engaged in Response-to-Next more frequently than the other levels of education (4.5 instances per participant with a graduate degree compared to 2.4 instances per participant for individuals with some college or a Bachelors degree)
- The oldest age group (50-64) read horizontally less frequently than the other two age groups (4.6 instances per participant compared to 10.0 instances per participant for 18-34 year olds)
- Individuals with some graduate study or a graduate degree used the mouse as a marker more frequently than other levels of education (8.7 and 6.7 instances per participant respectively, compared to 5.7 for participants with an Associates degree, and less than 4.5 instances for the remainder of the education groups)

Arguments could be made that individuals with higher education may have more computer experience, or use the computer more frequently to read detailed information, so they are more likely to use the mouse as a marker. However, given the small size of each of the demographic groups it is not clear whether these findings represent actual relationships or are just driven by the sample. Therefore, it will be interesting to see whether these same relationships hold with a larger sample size.

Now that we know who is engaging in the different movements, we can start to understand what the movements might mean. One thought is that the participants might have been more likely to display these movements early in the survey, as a way of becoming accustomed to the format of the instrument. Figure 2.1 provides the total number of movements across all participants sequentially.

Figure 2.1 Number of Mouse Movements Displayed on each Question



We can see from Figure 2.1 that participants did not engage in more movements at the beginning of the survey than they did at the end. In fact, there does not appear to be any pattern between location in the survey and the number of mouse movements displayed. Rather, it seems the number of movements is mostly related to characteristics of the questions themselves. The questions with the fewest number of movements, 'Telephone' and 'Difficulty walking,' are short questions with only two response options (Yes/No). On the other hand, the questions with the most movements, 'Rooms' and

‘Employee type,’ are long questions with multiple clauses and special instructions. Therefore, it is likely that the movements are related to some feature of the question.

Another hypothesis is that the movements are related to difficulty, either due to question complexity or comprehension. Table 2.6 provides the average difficulty rating from the probing questions (one was ‘very easy’ and five was ‘very difficult’) on questions where each movement occurred.

Table 2.6 Average Probing Question Rating by Mouse Movement

Movement	**Mean Rating**	**St dev**
Response-to-Question	2.91	1.51
Response-to-Next	2.09	1.19
Response-to-Response	2.06	1.39
Marker	1.84	1.18
Hover	1.78	1.06
Response-to-Space	1.72	0.98
Vertical reading	1.45	0.86
Horizontal reading	1.45	0.91

The mouse movements associated with the higher mean ratings are the different types of regressive movements, other than Response-to-Space. This is an encouraging finding because similar results were seen with eye regressions and confusion in Redline *et al.*’s 2009 study. Similarly, the lower ratings were associated with Vertical and Horizontal reading, which appear to be behaviors more typical of some individuals.

Table 2.6 also shows that there was not a lot of differentiation between difficulty ratings. The reason for this is twofold; participants answered based on their own situations, so it is possible participants did not think any of the questions were difficult. On the other hand, they may not have interpreted the probing questions as the researcher intended. For example, a question like ‘When was your house built?’ may not be considered a difficult question, because it is easy to understand. However, answering it can be very difficult if the person does not know the answer. Therefore, ratings could vary dramatically depending on how the participant interpreted the probing question: the actual difficulty of the question or the difficulty answering the question. Further, the rating process and examples were not provided before the survey started to help participants get a sense of the range of questions they would experience, so there may have been a relative effect of comparing future responses to the first provided.

To account for the limited differentiation between ratings and to help uncover subconscious actions, the debriefing reports were used as an additional measure of difficulty. Table 2.7 provides a count of how frequently each movement occurred on screens the participant said were difficult. In addition to the frequency, the final column of the table provides the percent of times each movement occurred on a screen the participant said was difficult during the debriefing as compared to all instances of that movement.

Table 2.7 Average Occurrence of Verbal Reports of Difficulty for each Mouse Movement

	Movements Associated with a Verbal Report of 'Difficult'	
Movement	**Frequency**	**Percent of Movement**
Response-to-Question	4	36.36
Response-to-Space	7	28.00
Response-to-Next	8	23.53
Response-to-Response	8	22.22
Marker	31	21.09
Vertical reading	22	19.30
Hover	11	16.18
Horizontal reading	14	12.07

We see from the table that of all the instances where participants engaged in Response-to-Question, they indicated the question was difficult 36 percent of the time. The results here are consistent with the findings from Table 2.6. The main outlier is

Response-to-Space. Otherwise, the top half and bottom half of the lists are consistent, with the different regressive movements at the top of the list and the other movements towards the bottom. The discrepancy between the two tables concerning Response-to-Space could be because there were only 26 instances of this movement.

In addition to looking at the difficulty ratings and verbal reports separately, they were also combined to further assess which movements might be associated with difficulty. Table 2.8 summarizes how frequently each movement occurred on all difficult questions (from the ratings or debriefings). The 'Movements on Difficult Screens as a Percent of all Movements' column is the number of times each movement was associated with a difficult screen as a proportion of the total number of occurrences of each movement. The 'Movements on Difficult screens as a Percent of all Difficult Screens' column represents the number of times each movement was associated with a difficult screen as a proportion of the total number of difficult screens (n=115).

This table can be analyzed several different ways. Looking at the 'Movements on Difficult Screens as a Percent of all Movements' column, 64 percent of the Response-to-Question occurrences were on questions participants thought were difficult. On the other hand, this event only happened on six percent of all the questions participants considered difficult. To the opposite effect, looking at the first row, only 34 percent of the occurrences of Marker took place on difficult questions, but the movement occurred on almost 43 percent of the questions that were difficult. One reason for these discrepancies could be that person-level differences are not being considered at this point. Additionally, different movements may be associated with different types of difficulty. For example, deciding whether to include a closet as a room might be different than figuring out which type of fuel an apartment uses. The closet example requires respondents to re-read the question, instructions, inclusions, and exclusions, whereas the fuel example requires them to think about whether they have a fuel pump outside or other information not provided in the details of the question. If participants engaged in different movements for different types of difficulty, then it stands to reason that no single movement would stand out.

Table 2.8 Mouse Movements Observed on all Difficult Questions

Movement	Frequency of Movement on Difficult Screens	Movements on Difficult Screens as a Percent of all Movements	Movements on Difficult screens as a Percent of all Difficult Screens
Marker	50	34.01	43.48
Vertical	25	21.93	21.74
Hover	19	27.94	16.52
Horizontal	18	15.52	15.65
Response-to-Next	18	52.94	15.65
Response-to-Response	13	36.11	11.30
Response-to-Space	8	32.00	6.96
Response-to-Question	7	63.64	6.09

Finally, it may be that it is not a single movement that is associated with difficulty, but rather multiple movements in combination. Table 2.9 provides the average difficulty rating across different numbers of movements on a question

Table 2.9 shows that as the number of movements on a screen increases, so does the difficulty rating. With the exception of the last row, the results suggest that it is likely the number of movements is related to difficulty. Without a model we cannot determine which combination of movements is most related to difficulty, but this hypothesis will be further tested in Chapter 3 using interactions.

Table 2.9 Average Probing Question Rating by Number of Mouse Movements

Total movements	Difficulty Rating	St dev	N
0	1.31	0.74	269
1	1.36	0.73	193
2	1.79	1.18	92
3	1.89	1.22	38
4	2.36	1.28	14
5	1.00	.	1

2.4.3 Focus of Attention Results

It is clear from this analysis that the participants regularly engaged in a variety of behaviors with the mouse while taking the survey and some of these behaviors may be related to difficulty. However, it is not clear where their focus of attention was while exhibiting in these behaviors. Table 2.10 shows where participants were looking while making each movement. The columns describe the part of the screen on which participants' gaze was fixated. The 'Percent Focused on Both' column represents people who looked at both the question text and the response options. The 'Unclear' column means the gaze plot was located between the question text and the response options, so it was not clear where they were looking. Additionally, Horizontal and Vertical reading were not included in this table because the participant's focus was the same as the location of the mouse.

To understand where participants looked while hovering, we see that in 42.86 percent of cases, they fixated on the question text, in 23.81 percent of cases they fixated on the response options, and in 33.33 percent of cases they focused on both the question text and response options while hovering. Therefore, it becomes clear that participants looked over all the areas of interest while hovering. From the wide range of focus, we can hypothesize that the purpose of hovering might be to double check an answer or to reread the question and/or response options to ensure they are selecting the most appropriate response.

Table 2.10 Eye Focus while Engaging in each Mouse Movement

Movement	Percent Focused on Question Text	Percent Focused on Response options	Percent Focused on Both	Percent Unclear
Hover	42.86	23.81	33.33	0.00
Marker	23.64	72.73	0.00	3.64
Response-to-Response	0.00	90.48	0.00	9.52
Response-to-Question	60.00	0.00	40.00	0.00
Response-to-Next	5.88	64.71	29.41	0.00
Response-to-Space	14.29	42.86	14.29	28.57

Unlike hovers, where participants looked at all the different areas of interest, the majority of Marker instances focused on the response options. This follows from past research, which has suggested people mark an option they like and then scan the rest to see if there is one that is more appropriate (Rodden *et al*., 2008; Guo and Agichtein, 2008). It is also logical that respondents engaging in Response-to-Response, which measures movement between response options, would have almost all their focus on the response options. This movement most likely suggests uncertainty between the two response options.

Response-to-Question, on the other hand, measures participants that moved from the response option to the question text and back again. This type of movement seems to suggest more uncertainty, which can be seen from the focus of attention. Participants focused on either the question text or the question text and response options. If participants searched through the response options and did not find the response they were looking for, they may have returned to the question text to make sure they understood what it was asking. This movement occurred often for the ‘Employee type’ question, especially for unemployed participants. The question asked what type of business the participant worked for last week. The last line in the question states that if the person did not work last week they should answer for the last job they held. However, many participants did not see this language the first time through, so once they did not see ‘unemployed’ in the response options, they returned to the question to gather more information.

At first glance, Response-to-Next also seems to be a check. Similar to Hover, participants answered the question, but then glanced back at the response options, or the response options and question text, to make sure they understood the question and answered it accordingly.

Finally, focus of attention for Response-to-Space seems to be quite spread out across the different areas of interest. This movement starts on the response options, moves away into white space and then returns. We hypothesize that participants thought they had an answer, then realized it may not be the most appropriate, so they stepped back to re-evaluate and then answer the question.

These interpretations and hypotheses, based on participants’ focus of attention, if correct, support the prior findings that Vertical reading, Horizontal reading, and Marker are not related to difficulty and are just how participants kept track of where they were on a page. Similarly, the different regressive movements, which were related to higher difficulty ratings, do seem to suggest a reevaluation of the question, response options, or both. It is still unclear what hovering means. In some cases, it seems to relate to difficulty, but not in the majority of instances. More data, an interaction between hovering and another movement, or differentiating between hovering over the question text and the ‘Next’ button might help uncover when this movement is indicative of difficulty. However, taking all this information together, the data provide a basis for hypotheses to test in the next chapter and a reasonable assumption that there are movements that are related to difficulty answering survey questions.

2.4.4 Response Time Results

Further support that these movements are related to difficulty can be found by looking at the relationship between the presence of movements and response latencies, since we know latencies are related to difficulty. Table 2.11 shows the median response time (in seconds), by question, for questions with no movements, questions with one movement, and questions with multiple movements. The table is ordered by the questions participants rated as the most difficult to answer. For this sample of participants, response times increased when movements were present for every question in the instrument. Additionally, the more movements in which a participant engaged resulted in even longer response times, although the sample sizes become quite small. Since these findings are in line with the results from the response latency research, Table 2.11 may offer more support that these movements are related to difficulty. However, it also takes more time to make more movements. Therefore, we cannot blindly make the leap that since mouse movements are related to longer response times and longer response times are related to difficulty, that mouse movements must be related to difficulty.

Table 2.11 Median Response Time for Questions with and without a Mouse Movement

	No Movement			One Movement			Multiple Movements		
Question	N	Median Response Time	St d	N	Median Response Time	St d	N	Median Response Time	St d
Year built	14	10.73	6.53	13	10.44	6.05	2	21.62	7.66
Employee type	7	19.23	8.75	11	24.94	4.20	12	34.77	11.02
Rooms	9	23.15	8.85	10	32.91	7.99	9	42.52	23.35
Weeks worked	8	14.58	3.44	6	17.15	11.90	3	21.62	7.66
Transport	13	7. 81	3.07	13	9.89	6.04	3	16.17	5.48
Internet	16	17.17	5.54	8	16.83	10.37	6	38.83	12.75
Fuel	15	11.35	5.08	9	16.09	6.23	6	16.96	6.72
Fifty or more weeks	16	8.54	2.53	12	14.48	5.53	2	19.62	3.06
Hours	18	11.52	4.20	9	12.60	5.35	3	27.47	18.17
Relationship	7	12.50	2.71	10	13.28	4.97	7	21.81	17.63
Vehicles	22	7.66	3.03	6	9.74	5.85	2	27.86	2.04
Live or stay	25	9.07	2.81	5	17.29	9.44	-	-	-
Type of unit	9	13.00	3.81	14	15.33	4.87	7	17.48	6.73
Educational Attainment	13	14.90	7.73	11	12.33	6.01	6	21.01	10.22
Facilities	27	4.80	1.57	3	5.10	0.76	-	-	-
Marital	16	6.49	2.22	12	8.00	3.83	2	13.35	5.74
Race	22	4.52	1.88	6	13.02	5.44	2	22.14	7.10
Hispanic	22	5.39	1.65	7	6.71	2.45	1	8.10	-
Work last week	18	7.79	2.29	12	11.29	3.95	-	-	-
Diff walking	24	5.12	2.06	6	5.68	3.27	-	-	-
Attend school	22	5.94	1.97	6	6.96	1.20	2	9.98	6.15
Probing Questions	570	3.68	1.82	42	5.71	3.49	-	-	-

To help differentiate between longer response times as a result of more movements as compared to difficulty, it may be helpful to see if response latencies are longer for questions that respondents rated as more difficult as compared to easy questions. Keeping in mind that there was not much differentiation between difficulty ratings for questions, eight of the top 10 questions rated as most difficult also had the largest difference in response times when a movement was present as compared to no movement. The exceptions were ‘Year built’ and ‘Transport to work.’ With ‘Year built,’ many people simply did not know the answer to this question. So while they rated it as difficult, they did not spend additional time trying to answer. This also suggests that just because a movement is present does not mean response times are drastically increased. It is not as obvious why ‘Transport to work’ had a smaller than expected difference in response times. Most participants who struggled with this question took two forms of transportation to work every day (drove to the metro and then took the train to the office). Therefore, it is possible they answered based on their first mode of transportation or the one on which they spent the majority of time.

While there were two cases where longer differences were expected, there was also one case where participants rated a question as very easy, yet there was a large difference between response times. ‘Race’ was ranked the 4th easiest of all the questions, but it had the second longest difference in response time. This is most likely because some participants read through all of the response options just to make sure they were not missing anything or just to see what races were listed, which would likely result in either vertical reading or using the mouse as a marker.

Since these data suggest there is a relationship between difficulty and response times, we can again try to tie specific movements to difficulty by looking at which movements were associated with the longest differences in response times. The five questions with the largest difference in response times between questions with no movements and with one movement are: ‘Rooms,’ ‘Race,’ ‘Employee type,’ and whether the respondent worked fifty or more weeks (‘Fifty or more weeks’). Table 2.12 displays how often each mouse movement occurred on each of these questions.

Table 2.12 Frequency of Mouse Movements for Questions with Long Differences in Response Time

Question	Difference between response time when any movement and no movements	Movement							
		Horizontal reading	Vertical reading	Hover	Marker	Response-to-Response	Response-to-Question	Response-to-Next	Response-to-Space
Rooms	14.77 sec	12	4	9	14	4	2	1	6
		10.34%	3.51%	13.24%	9.52%	11.11%	18.18%	2.94%	24.00%
Race	9.87 sec	2	7	2	3	0	0	1	2
		1.72%	6.14%	2.94%	2.04%	0.00%	0.00%	2.94%	8.00%
Employee Type	9.40 sec	10	14	22	7	2	3	2	3
		8.62%	12.28%	32.35%	4.76%	5.56%	27.27%	5.88%	12.00%
Internet	7.45 sec	7	5	2	10	3	1	6	3
		6.03%	4.39%	2.94%	6.80%	8.33%	9.09%	17.65%	12.00%
Fifty or more weeks	6.19 sec	6	0	3	8	0	0	2	0
		5.17%	0.00%	4.41%	5.44%	0.00%	0.00%	5.88%	0.00%

For each question, the first row provides the frequency of each movement while the second row provides the frequency as a percentage of all instances of each movement.

A first glance at Table 2.12 suggests participants engaged in a wide variety of movements on all of these questions. Looking at the first row for ‘Rooms’, we see that participants displayed each movement at least one time, with a maximum of 14 times for Marker. Additionally, the second row for ‘Rooms’ shows that even though there were only two instances of Response-to-Question, this accounted for 18 percent of all the instances of that regressive movement. Both of these measures can tell us something about the meaning of each movement. The raw number tells us which movements may be associated with the more difficult questions while the percent can link each movement to specific question characteristics or allow us to hypothesize what the movements might mean.

Focusing on the frequencies and starting with ‘Rooms’, the most common movement was Marker, followed by Horizontal reading (14 and 12 instances, respectively). We know this question is complex and difficult for respondents due to the inclusions and exclusions (Schober and Conrad, 1997). In the debriefings from this study, this question received the third highest difficulty rating of all the questions. Taking this information together, it appears that participants had a starting guess at an answer and used the mouse as a marker on that response option while they thought about the rooms in their home and considered the inclusions/exclusions. As part of this consideration, they likely read the inclusions/exclusions carefully, so they followed along with the mouse while reading. This leads us to believe using the mouse as a marker is related to thinking or rereading and Horizontal reading is related to paying attention to the details of the question.

‘Race’ is a very different question than ‘Rooms’ in that it is a simply worded, short question with many response options. Not surprisingly, the most common movement on this screen was Vertical reading (seven instances). The most realistic interpretation is that some participants read through all of the response options instead of clicking the first suitable option. Since there are many response options, it takes time to read through them all, which is the reason for the large difference in response times. However, due to the low prevalence of other movements and the very low difficulty rating, it does not appear that respondents had trouble answering this question or that Vertical reading is related to difficulty.

Similar to ‘Rooms,’ ‘Employee type’ is a very complex question. Additionally, the response options are long and detailed. The most common movement on this question was Hover (22 instances). As discussed in the debriefings and supported by the difficulty ratings (employee type had the second highest average difficulty rating), many participants did not read the question carefully enough to be able to answer it accurately. Therefore, these participants likely hovered over the question text or ‘Next’ button while they went back and forth between the response options and question text to try to find the information that would allow them to answer the question. While Hover does not appear to always be related to difficulty, in this case we have strong support that it does. Separating hovering over the question text and ‘Next’ button may help with this differentiation.

‘Internet’ was problematic for respondents partially because it was a new format compared to the other questions on the survey. Additionally, several participants used open networks in their building or neighborhoods and they did not know what type of service the owner used, resulting ‘Internet’ in being rated the 6th most difficult question. It is for these reasons there is a wide variety of prevalent movements on this question. However, the most frequent movement was Marker (10 instances). Similar to ‘Rooms,’ it is likely participants gravitated to a particular response option and kept the cursor on that option while reading the others to ensure they had the correct response.

Finally, ‘Fifty or more weeks’ was more of a math problem for participants, We did not see as many movements as we did for the other questions, but the most common movement was Marker (eight instances). Again, participants likely went with their initial reaction to the question and then used the mouse as a marker on that response option while they did mental math and/or read the inclusions and exclusions in the question text. Similar to ‘Rooms,’ Horizontal reading was the second most common movement (six instances), providing further support as to what participants do while using the mouse as a marker.

The percentages provided in Table 2.12 allow us to verify hypotheses regarding what question characteristics may be related to each movement. Specifically, we believe Horizontal reading occurs on long questions with a lot of detail. From the ‘Horizontal’ column, we see this movement occurred most frequently on ‘Rooms’ and ‘Employee type’ (10.34 percent and 8.62 percent, respectively), which are the longest and most complex of the five questions listed. We also believed that participants would be more likely to read vertically if there are more response options. However, we actually saw a higher percent of participants reading vertically on ‘Employee type’ (12.28 percent), suggesting it might not be the number of response options, but a combination of the number and the complexity of the content. From the analyses in Section 2.4.2, we believe the regressive movements are associated with more difficult questions. Of the questions listed in Table 2.12, participants rated ‘Rooms’ and ‘Employee type’ as the most difficult. We saw the highest percent of Response-to-Response and Response-to-Space occurred on ‘Rooms’ while the highest percent of Response-to-Question occurred on ‘Employee type.’

Although this analysis does not provide any concrete relationships between the movements and reasons participants made them, it does help us further understand why participants engaged in different movements and what they were thinking about.

2.4.5 Debriefing Explanation Results

While we know which movements participants engaged in and where their attention was focused we still do not know what they were actually thinking. Through the debriefings, we received a total of 119 explanations of what participants were thinking about on questions throughout the instrument. Every participant provided information on at least one question. Table 2.13 provides the most common explanation participants provided in the debriefing on screens in which they engaged in each movement.

Consistent with our prior hypotheses, Vertical reading and Marker do not appear to be related to confusion, but rather looking for a response option. Additionally, the regressive movements appear to be related to difficulty, which we also saw in the prior analyses. Specifically, these movements were all exhibited when participants did not know the answer, when they had difficulty mapping their situation to the response options, or when they needed to double check or re-read information. Finally, Hover looks to be related to both browsing response options and difficulty, which explains the ambiguous results seen in the prior analyses.

Table 2.13 Most Common Explanation Provided on Questions where Mouse Movements Occurred

Movement	**Most Common Explanation**	**Percent of Explanations Associated with each Movement**
Horizontal reading[6]	Thinking and Double checking/Re-reading	33.33 (each)
Vertical reading	Looking for a specific option	27.59
Hover	Looking for a specific option and Don't know	24.00 (each)
Marker	Looking for a specific option	19.29
Response-to-Response	Don't know	42.86
Response-to-Space	Don't know	42.86
Response-to-Next	Double checking/Re-reading and Don't know and Difficulty mapping	25.00 (each)
Response-to-Question	Double checking/Re-reading	25.00

Before reading too much into the results from Table 2.13, it should be noted that participants engaged in multiple movements that were associated with just one explanation, so there is not a one-to-one relationship between movement and reason. Further, not every participant provided commentary for his/her actions. Therefore, there are not a lot data from which to draw conclusions, less than 10 instances in some cases. Additionally, most participants only provided explanations for questions they thought were difficult or the researcher thought they were having difficulty answering. Therefore,

[6] Readers were not included in the analysis for this movement because we know what they are thinking about when reading horizontally.

it is not necessarily an accurate sampling of all the times a participant might engage in these behaviors, especially for movements such as Marker or Hover.

2.5 Conclusions

The goal of this study was to identify a set of mouse movements used by survey respondents while completing Web surveys. We successfully identified six different movements: reading horizontally, reading vertically, using the mouse as a marker, hovering, regressing between different response options, and regressing between the response options and ‘Next’ button.

In addition to understanding the movements participants made, it is more valuable to determine what these different actions meant and to understand why participants engaged in them. By examining which movements were associated with higher difficulty ratings and where participants’ focus of attention was while engaging in the different movements, we were able to pinpoint a subset of the movements that appeared to be related to difficulty answering a question: Response-to-Response and Response-to-Next. Other movements, such as Horizontal reading, Vertical reading, and Marker seemed to involve keeping track of the participant’s focus of attention. We were unable to generate a hypothesis regarding Hover.

Due to the exploratory nature of this analysis, the sample size was only 30 participants. While this sample provided enough information to form hypotheses about what these different movements mean and which are related to difficulty, they cannot be statistically tested.   Therefore, it is still unclear which individual movements are related to difficulty, whether a combination of movements is predictive of difficulty, and if a model can capture the relationship between the movements and difficulty in a way that can be used in real-time to assess whether respondents need assistance answering a question.

The next chapter in this dissertation attempts to address these outstanding issues. With a larger sample size, we can statistically test the hypotheses proposed in this study. We can also generate a hierarchical model, which will account for individual differences, to understand which combination of movements suggests that a participant needs assistance. Additionally, participants answered each question in the subsequent study based on scenarios instead of autobiographically. This allowed us to vary the type of difficulty to test whether different movements are more predictive of one type of difficulty than another. This analysis will provide additional information that can help researchers provide tailored help to respondents.

Chapter 3: Identifying Which Movements are Related to Difficulty and Developing a Model to Predict when a Respondent is having Trouble Answering a Question

3.1 Background

The study discussed in this chapter is directly related to the study discussed in Chapter 2 of this dissertation. In the first study, we identified a set of mouse movements that the participants commonly engaged while completing a Web survey (Table 3.1).

Table 3.1 Mouse Movements Identified in Chapter 2

Definition	**Description**
Horizontal reading	Mouse follows along with the eye to read along from left to right
Hover	Mouse stays over the 'Next' button or the question text for more than 2 seconds
Marker	Mouse stays over a radio button or response option text for more than 2 seconds
Response-to-Response	Mouse moves back and forth between two options.
Response-to-Question	Mouse moves from response options to question and back again one or more times
Response-to-Next	Mouse moves from response options to 'Next' button and back again one or more times
Response-to-Space	Mouse moves from response options to white space and back again one or more times
Vertical reading	Mouse is used to follow the eye from top to bottom or bottom to top

In addition to looking into which movements participants used, we were also interested in which may be related to difficulty. We used ratings from the probing questions, information from the debriefings, and information on participants' focus of attention while engaging in the different mouse movements (from the Tobii recordings) to generate several hypotheses regarding which movements are related to respondent difficulty answering a question. Specifically, we hypothesized:

- Instances of Horizontal reading, Vertical reading, and Marker **are not** associated with an increase in difficulty
- Instances of hovering with the mouse **may be** associated with an increase in difficulty

- Instances of regressive movements **are** associated with an increase in difficulty

Unfortunately, due to a restricted sample size in the first study, we were unable to test these hypotheses.

The goal of the current study, therefore, is to statistically test each hypothesis and determine which movements are associated with difficulty. Using this information, we will then attempt to generate a hierarchical random effects model capable of predicting when a respondent is having difficulty answering a question. This model can then be used to provide real-time assistance to Web respondents who are having trouble answering questions.

In addition to determining which movements are associated with difficulty in general, it would be beneficial to be able to determine what type of difficulty people are facing. In a lab setting, eye tracking technology can help inform researchers as to what the respondent is struggling with, as was discussed in Section 1.2.3. However, in production the researcher is blind to this information. By testing whether specific types of difficulty are associated with different mouse movements, it may become possible to not only predict when a respondent needs help, but also what type of help they need.

Currently, response latencies are used to predict when respondents are having trouble answering a question, with longer than average response times suggesting a problem (Schober and Conrad, 1997; Conrad *et al.*, 2006). However, using response times to predict difficulty has several shortcomings. First of all, for first time surveys, median response times are not known for each question. Additionally, in a production setting, respondents could be taking the survey at home, at their place of employment, at a library, or many other places. All of these locations have many distractions which could lead the respondent to step away from the computer or temporarily switch to a different task such as checking email or tending to a child. If a researcher is using response times to measure difficulty, the clock continues to progress while respondents are engaged in other activities which have nothing to do with difficulty or the survey task at all. Therefore, many of these respondents would be provided help unnecessarily. Using mouse movements helps ensure, at the very least, the respondent is at their computer and has the survey activated and they may have further benefits as well.

3.2 Study Design

This study consisted of 100 participants answering 20 questions from the ACS. Participants answered each question based on a pre-written scenario and then rated the difficulty of each question after answering it. While participants completed the survey, we recorded the sessions using Tobii eye tracking hardware and software. Using these recordings, we later coded the mouse movements used in the analysis. This section discusses the survey, the scenarios used, the participants, and the procedures used to collect the data.

3.2.1 Web Survey

The data for this study consisted of the same 20 questions taken from the Web version of the ACS that were used in Chapter 2 along with the probing questions following each ACS question (Appendix F)[7]. As in the previous study, help and clarifying information were not available to participants. In addition to the 20 ACS and probing questions, there was also a series of demographic questions, including gender, age, race, education, and computer experience. We included the probing questions to ensure participants actually found some of the questions to be difficult. Finally,

[7] The roster question is not used in the second study because the use of scenarios eliminated the need to define other people in the household for purposes of the Relationship question.

participants were debriefed after they completed the survey to determine why they engaged in certain movements and why they rated questions that were intended to be easy as difficult[8]. The debriefing protocol from the first study was modified for this study and can be found in Appendix G.

3.2.2 Scenarios

As opposed to the first study where respondents answered the questions autobiographically, in this study they answered based on pre-written scenarios. For 17 of the 20 questions, there were two scenarios that accompanied each question. For the other three questions, we used one scenario but had two versions of the question which varied the question text and format. We will use "scenario" to describe both of these situations, as they will be discussed together.

For each question, the scenarios were designed such that one that involved a simple cognitive mapping so that answering the question based on the information in the scenario was straightforward, whereas the other involved a complex cognitive mapping, which made the question more difficult to answer and may have required assistance to answer correctly. In addition to simply making some scenarios more difficult than others, we also varied the type of difficulty so we could determine whether specific types of difficulty were related to specific mouse movements. The specific types of difficulty we manipulated were: imperfect fit (difficulty mapping one's personal experience to the response categories), misalignment (commonly used words are used in a different way), and technical or unfamiliar terms.

Because there were two scenarios for each question in this study, it was important to randomize which scenario participants received. Additionally, we needed balance both

[8] This information was used to rewrite some of the scenarios used in the next chapter.

within and between respondents. Specifically, each participant received 10 randomly selected straightforward scenarios and 10 randomly selected complex scenarios. This randomization was conducted such that, for each question, 50 participants received the straightforward version and 50 received the complex version. This balance allowed us to compare both across scenario types and within each question.

Randomizing while maintaining this balance required several steps. First, for each participant Id, we assigned 10 questions to the straightforward version of the scenario and 10 to the complex version so across all participant Ids there were 50 straightforward versions and 50 complex versions of the scenario. Next, for each Id, we selected two questions using a random number generator built into the Perl programming language. We then swapped the versions of these two questions. For example, if question two was assigned the straightforward version of scenario and question 12 was assigned the complex version of the scenario, then after the swap question two was assigned the complex version and question 12 was assigned the straightforward version.

Next, using the same random number generator, we selected another participant Id. This Id had to be assigned to the same versions of the two selected questions as the first participant Id that was selected (after the swap). The versions for the second participant Id were then swapped so they no longer matched the versions for the first Id. This ensured that we maintained the balance of 50 participants receiving the straightforward version and 50 receiving the complex version of each scenario. We repeated these steps for all of the remaining Ids and then repeated the entire process 10 times for all users to ensure the results were completely random and shuffled.

To guarantee there was sufficient differentiation between the difficulty of the straightforward and complex scenarios, all of the items were pretested. There were three rounds of pretesting, with approximately 20 participants in each round (22 first round, 22 second round, and 20 third round). Table 3.2 provides the demographic characteristics of the pretest population[9].

Table 3.2 Demographic Characteristics of Pretesting Population

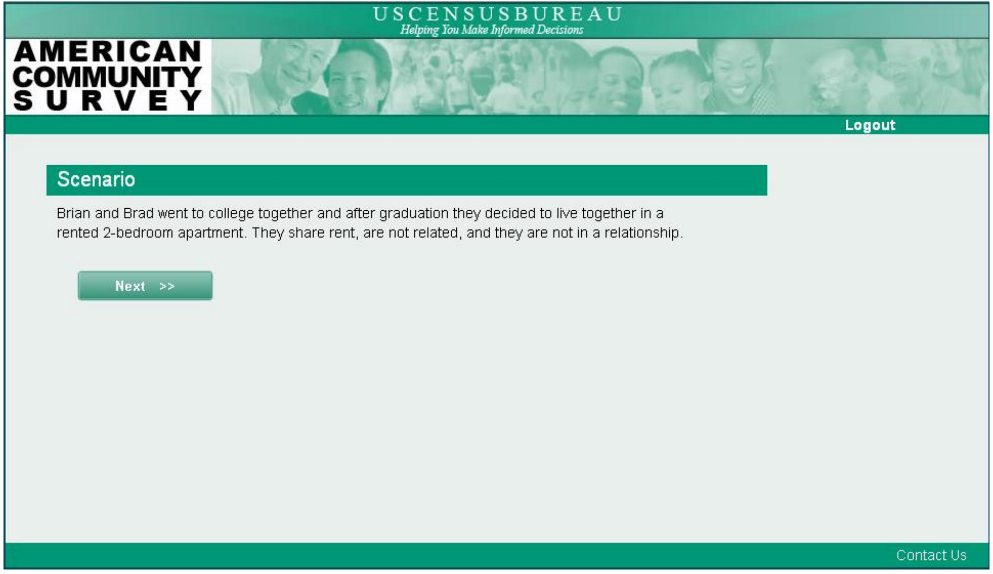
Gender	**Number**	**Percent**
Male	32	50.0%
Female	32	50.0%
Age		
18-34	53	82.8%
35-49	5	7.8%
50-64	5	7.8%
65+	1	1.6%
Education		
Some High School	1	1.6%
High School Graduate or equivalent	3	4.7%
Some college	9	14.1%
Associate's Degree	2	3.1%
4-year college Degree	20	31.3%
Some Graduate school	3	4.7%
Post Graduate Degree	13	20.3%
Race		
Black	3	4.7%
White	50	78.1%
Other	11	17.2%

The pretest participants came from a convenience sample in the Washington, DC metro area. Each participant received a stack of 20 pieces of paper. Each sheet of paper had a scenario (Figure 3.1), followed by an ACS question and response options, which was followed by the same probing question used in Chapter 2. Participants received one of two stacks of paper. Each stack consisted of 10 straightforward scenarios and 10

[9] Not every participant provided their educational attainment.

complex scenarios, which had been randomly selected for each stack. In other words, if the first 10 scenarios were straightforward in stack 1, then those 10 scenarios were complex in stack 2.

Figure 3.1. Screenshot – Example of a Scenario


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
Scenario
Next >>
Contact Us

Brian and Brad went to college together and after graduation they decided to live together in a rented 2-bedroom apartment. They share rent, are not related, and they are not in a relationship.

In the first round of pretesting, participants answered and rated all of the questions. The researcher then tested which of the scenarios had significant differences between the ratings. For scenarios that did not reach a significant level of difference, the scenarios were re-written and provided to the second set of participants. The same strategy was followed until there was sufficient difference in the difficulty for all of the scenarios. Appendix H provides the results of the significance testing[10] and Appendix F provides the final question/scenario combination for each question.

3.2.3 Data

The same Tobii eye tracking hardware and software were used to capture eye and mouse movement data as were used in the first study (Section 2.2.3). Half way through

[10] Significance testing was not done for "Educational Attainment" because the heading/no heading manipulation does not change difficulty in the conventional sense. Additionally, no testing was done for "Telephone" because of an oversight.

the data collection we upgraded to the Tobii X-60 hardware. The original hardware was built into a monitor, which limited the flexibility in terms of monitor size and ability to adjust the viewer to participants with varying height, posture, and visibility. The Tobii X-60, on the other hand, is a separate component from the monitor, which allowed us to use a larger monitor and also have more flexibility in positioning the viewer. This made the experience more comfortable for participants and ideally reducing the number of cases with poor eye movement capture that was experienced in the first study, although we did not end up using the eye tracking data in the analysis[11]. Further, the change in hardware did not affect the coding of mouse movements because the recordings from both devices used the same resolution and the same software version.

3.2.4 Participants

The target number of participants for this study was 100 individuals[12] from Washington, DC metropolitan area. Participants were recruited using flyers posted around the University of Maryland campus, postings to the University of Maryland's list serves, and text advertisements on Craigslist (Appendix I). We wanted to capture a diverse group of participants because we did not know whether certain groups moved the mouse differently than other groups. As mentioned in Chapter 2, Schober and Conrad (1997) found that older respondents had longer response times than younger respondents. Similarly, we thought younger respondents may use the mouse more than older respondents because they grew up using computers with a mouse as a pointing device. Additionally, it is possible that people with higher education use computers more than others. Therefore, the following criteria were used as a recruitment guide (Table 3.3):

[11] We had intended to use the new hardware from the beginning of the study, but there was an acquisition issue, so we started without it since we did not know how long the situation would take to resolve.

[12] This sample size allowed us to detect differences of 0.03 with a power of 0.80.

Table 3.3. Demographic Recruitment Criteria

Gender	**Percent of Respondents**
Male	50
Female	50
Age	
18-34	30
35-49	30
50-64	30
65+	10
Education	
High School Graduate or equivalent	10
Some college	30
4-year college Degree	40
Graduate work	20

This range in age and education should help us determine whether different groups of people consistently behave differently than other groups. Although we saw some small differences in mouse use between different age and education groups in Section 2.4.2, the small sample size may have uncovered false relationships or masked true relationships. If these relationships do in fact exist, a broad range of participants will improve the modeling of movements because different models can be targeted to different groups.

After one month of recruiting in March 2012, the final set of participants consisted of 100 individuals. These participants ranged in age from 18 to 66 and their education ranged from some high school to a post graduate degree. Table 3.4 provides a breakdown of the demographic characteristics of the participants. Each participant answered a series of screener questions to determine eligibility based on age, education, and computer experience (Appendix J). Additionally, any person that participated in the first study or had participated in three or more research studies in the year preceding this study was not eligible. Finally, we decided early in the data collection process that due to the cognitive burden of the task (reading and remembering scenarios and applying them to the question), only individuals who completed high school would be included in the analysis.

Table 3.4 Demographic Characteristics of the Participants

Gender	**Percentage of Respondents**
Male	40
Female	60
Age	
18-29	54
30-39	20
40-49	14
50-59	9
60+	3
Education	
Some High School	2
High School Graduate or equivalent	9
Some college	36
Associate's degree	4
4-year college degree	19
Some graduate school	14
Post graduate degree	16
Race	
Black	43
White	37
Asian	17
Other	3

We were unable to reach the proposed recruiting criteria because the nature of recruitment could not target specific groups. By restricting the sample to people at the University of Maryland and people who use Craigslist, we were limited in our ability to recruit a wide range of individuals. However, it is not expected that this will limit our findings, especially since there is still a range of ages and levels of education.

As with the first study, all of the participants came to the JPSM office at the University of Maryland to complete the survey and were reimbursed $30 for time and travel.

3.2.5 Procedure

Data Collection Procedures

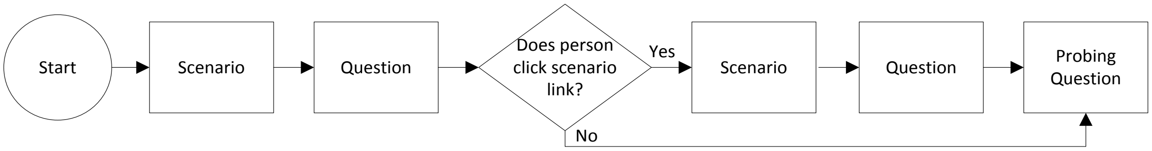
Eligible participants came into the JPSM office and sat in front of the Tobii eye tracking monitor. Each participant listened to an introduction and background to the study read by the researcher (Appendix G). This introduction was very similar to the one used in the study discussed in Chapter 2, except it explained that participants would be answering based on scenarios instead of their own experience. They were then asked to read and sign a consent form and answer a brief questionnaire covering demographic information and computer and Internet use (Appendix E). Finally, the participants had their eyes calibrated for eye tracking.

To minimize interference between looking at the scenario and looking at the question, before the researcher logged participants into the survey, she stressed that participants read the scenario very slowly and carefully before moving on to the question. If it was absolutely necessary, the participant could refer to the scenario using a 'Scenario' link that appeared on each question page, but they were told to avoid doing this if at all possible[13]. Additionally, the researcher provided instructions on how to rate the difficulty of each question. Specifically, participants were told to base the rating on their own experience answering the question, not based on how the person in the scenario

[13] This was done because if participants saw the scenario after seeing the question, they may do most of their thinking while looking at the scenario, which would limit their mouse movements. For this same reason, we also did not allow cheat sheets.

or the "average person" might rate the difficulty. Figure 3.2 provides the flow that participants followed when answering each question.

Figure 3.2. Flow Chart Describing Participants' Progression through each Question



After providing instructions on how to navigate through the instrument, the researcher left the room while the participant completed the survey. From a different computer, the researcher was able to see the participants' screen while they completed the survey[14]. The researcher noted when respondents engaged in any of the specified mouse movements on questions related to straightforward mappings or when they rated a straightforward mapping as difficult because we did not expect as many movements on these screens.

After participants finished the survey, the researcher debriefed them on the survey experience. Participants had access to all the questions in a PowerPoint document to facilitate recall and they were asked if there were any questions they thought were difficult and what made them difficult. Finally, a retrospective think aloud was used to help explain why participants engaged in certain behaviors on straightforward scenarios.

Mouse Movement Measurement

The mouse movements defined in the first study were used as a basis for this study. Since no participants in the first study clicked on words or highlighted phrases,

[14] The viewer used to monitor the survey from another room was disabled for the final six participants. Therefore, the researcher could not note specific behaviors to discuss in the debriefing.

these movements were not considered. Additionally, while analyzing the data from the first study, the researcher identified seemingly distinct behaviors that had previously been considered a single movement. Specifically, Horizontal reading and Hover were counted as single movements in the first study, but depending on what the participant was reading or hovering over, the movements could potentially have very different meanings. Therefore, the Horizontal reading movement was split into horizontally reading the question text (Horizontal reading – Question) and horizontally reading the response options (Horizontal reading – Response). Similarly, instead of tracking hovers in general, we tracked hovers over the question text (Hover – Question) and the 'Next' button (Hover – Next) separately. The other movements were the same as in the prior study and are described in Table 3.5 along with coder instructions.

Table 3.5 Description of Mouse Movements and Coding Rules

Definition	**Description**	**Additional Instructions**
Horizontal Reading - Question	Mouse follows along with the eye to read along with text from left to right across question text	Only flag once if occurs
Horizontal Reading - Response	Mouse follows along with the eye to read along with text from left to right across response option text	Only flag once if occurs, indicate which response option/options
Vertical Reading	Mouse is used to follow the eye from top to bottom or bottom to top	Flag each instance
Hover - Question	Mouse stays over the question text for more than 2 seconds	Flag each instance, Record duration
Hover - Next	Mouse stays over the 'Next' button for more than 2 seconds	Flag each instance, Record duration
Marker	Mouse stays over a radio button or response option text for more than 2 seconds	Indicate which option, Flag each instance, Record duration
Response-to-Response	Mouse moves back and forth between two response options.	Indicate which options, Flag each instance
Response-to-Question[15]	Mouse moves from response option to question and back again	Indicate which option, Flag each instance
Response-to-Next	Mouse moves from response option to 'Next' button and back again	Indicate which option, Flag each instance
Scenario	Respondent clicks 'Scenario' link	Only flag once if occurs

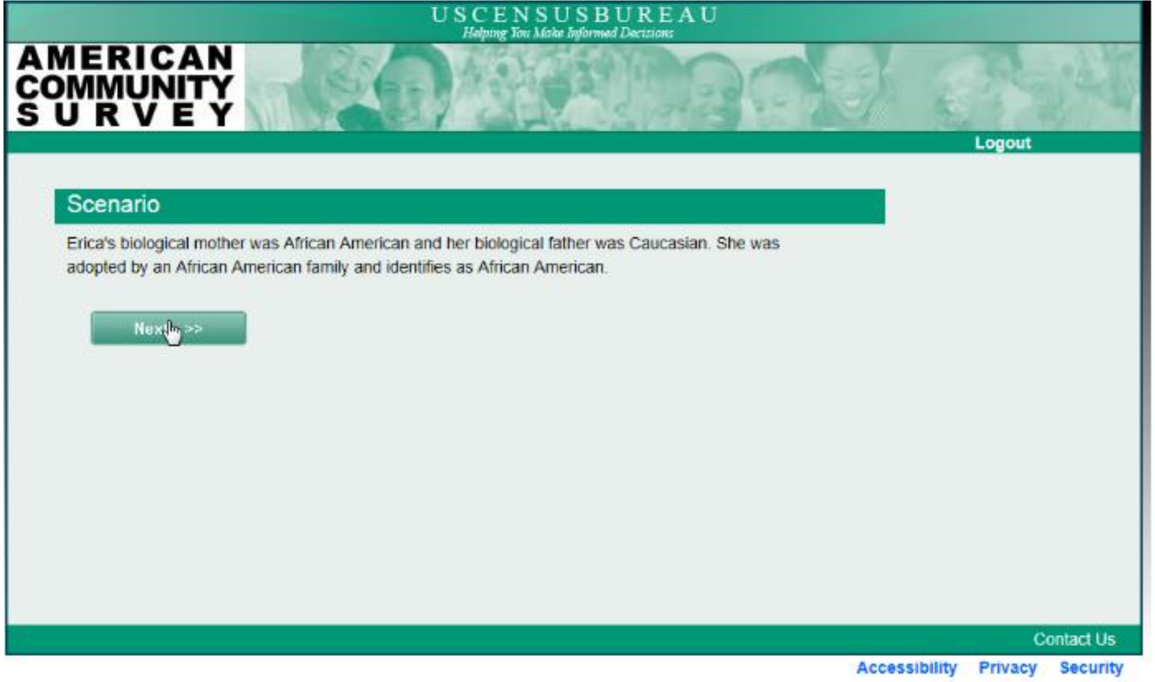
In the first study, we only considered whether or not a movement occurred on a question. However, multiple movements might have more meaning. Therefore, we also

[15] This movement was excluded from the analysis in Chapter 2 due to low reliability, possibly resulting from a low frequency of occurrence. As this study uses a larger sample size, we decided to code this movement to see if it is more reliable.

tracked how often each movement occurred on each question. The ‘Additional Information’ column in Table 3.5 specifies what data were collected for each movement.

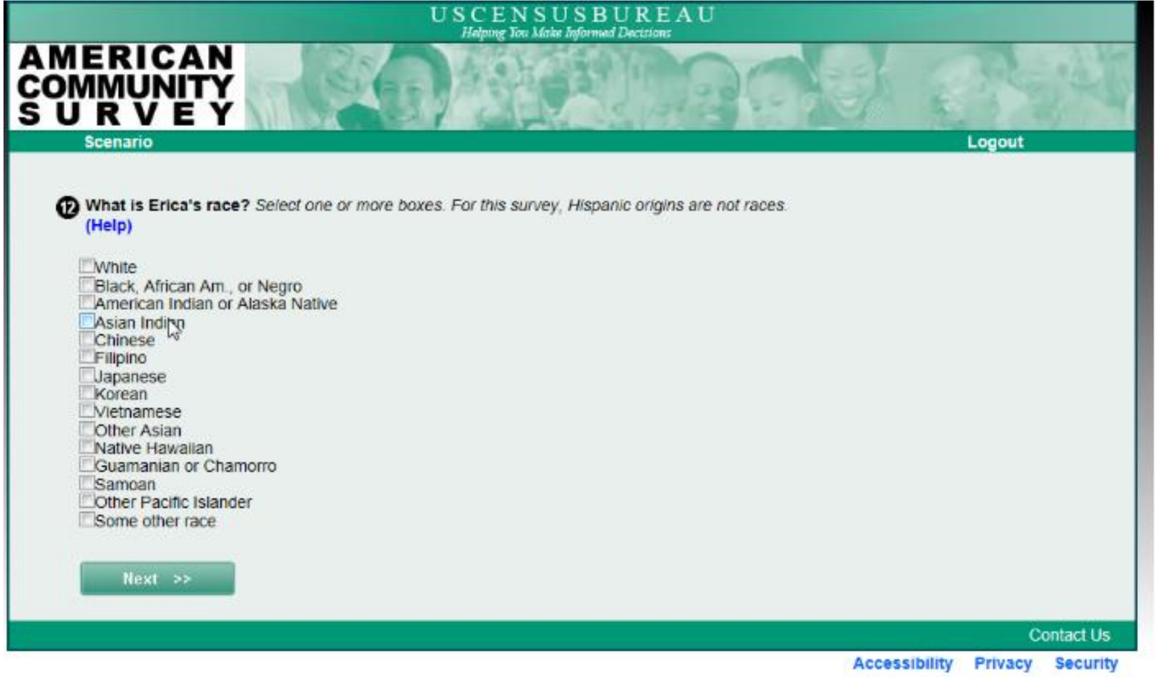
The movements described in Table 3.5 were only considered movements if the participant engaged with the mouse to create the movement. Specifically, if the participant clicked ‘Next’ and when the page loaded the cursor was on a response option, this did not count as Marker. For example, Figures 3.3 and 3.4 provide before and after screenshots of this phenomena. The participant clicked ‘Next’ on the ‘Race’ scenario. When the screen loads, the cursor is highlighting the “Asian Indian” response option. Assuming the participant leaves the mouse in this location for two or more seconds, this would not be coded as Marker. Rather, the participant would need to engage the mouse and deliberately place it on a response option, question text, or ‘Next’ button in order for it to count as a movement. Because the movements were defined this way, almost all Hover – Next movements occurred after a response had been selected.

Figure 3.3 Screenshot of Cursor Location before Selecting ‘Next’ on Scenario Screen


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
Scenario
Next >>
Contact Us
Accessibility Privacy Security

Erica's biological mother was African American and her biological father was Caucasian. She was adopted by an African American family and identifies as African American.

Figure 3.4 Screenshot of Cursor Location after Selecting ‘Next’ on Scenario



⓬ **What is Erica's race?** *Select one or more boxes. For this survey, Hispanic origins are not races.*
(Help)

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native
□ Asian Indian
□ Chinese
□ Filipino
□ Japanese
□ Korean
□ Vietnamese
□ Other Asian
□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander
□ Some other race

Finally, we assigned each question to one of three types of difficulty (imperfect fit, misalignment, and technical or unfamiliar terms) to see whether specific movements were more related to one type of difficulty than another. These types of difficulty were originally defined by Tourangeau and his colleagues in their 2006 paper. Imperfect fit is a lexical processing issue and occurs when respondents have difficulty mapping their experience to the answer categories. For the manipulations in this study, we considered imperfect fit to be when the participant could answer with multiple response options and he or she is not sure which applies more appropriately. Misalignment is a referential processing issue and occurs when common words are used differently than the respondent is accustomed to. For the manipulations in this study, we considered misalignment to be when the correct answer depends on how certain words are defined. Finally, questions also can contain terms the respondent is unfamiliar with or does not know the definition to. To manipulate this type of difficulty, we included uncommon words in the scenarios and questions. Table 3.6 provides the type of difficulty for each question, as well as the manipulation that caused the difficulty.

For each type of difficulty, we outline which questions manipulated that type of difficulty and what word or phrase was intended to be difficult for participants. For example, the first question under ‘Imperfect Fit’ is whether the individual in the scenario attended school in the past three months. The answer to this question depends upon whether continuing education is included in the definition of school. Similarly, under ‘Misalignment,’ the participant needs to know how the researcher defines “built” in order to say in what year the building was first built. To some people the building would need to be complete before it is considered built, whereas other people may think it is when construction starts. Therefore, Table 3.6 aims to explain the type of manipulation we induced for each question and the specific issue the participant faced when answering the question.

There were three questions for which we did not assign a type of difficulty or include in this analysis. The ‘Rooms’ question was intended to invoke misalignment, but participants had such difficulty remembering the floor plan that the intended manipulation was ineffective. Similarly, the question asking participants whether the person in the scenario had worked 50 or more weeks was supposed to focus on how many weeks there are in a year. However, participants mostly struggled with the mental arithmetic and not the intended manipulation. Finally, ‘Educational attainment’ manipulated the format of the question, so it did not fall into any of the categories and was therefore not used.

Table 3.6 Type of Difficulty and Manipulation for each Question

Imperfect Fit	
Question	**Mapping Ambiguity – Is it included in the definition?**
Attend School	Continuing education
Hours worked	Uncompensated hours
Work last week	Summer vacation
Vehicles	Recreational vehicle
Transport to work	Multiple locations
Type of unit	Storage unit
Misalignment	
Question	**Referential Ambiguity - What definition applies?**
Year built	Built
Difficulty walking	Difficult
Race	Genetic versus self-identification
Hispanic	Genetic versus self-identification
Telephone	Household
Live or stay	Live/stay
Technical/Unfamiliar Terms	
Question	**What is the definition?**
Fuel	Geothermal heat
Employee type	Federal Government employee/Independent Contractor
Marital status	Married/registered domestic partnership
Relationship	Roomer/Boarder
Internet	Subscribe
Not using	
Question	**Reason**
Rooms	Difficulty remembering scenario
Weeks worked	Mental arithmetic
Educational attainment	Differences in formatting, not comprehension

We expected participants to engage in more regressive movements for questions with mapping issues and more hovers and using the mouse as a marker for questions with technical or unfamiliar terms. The questions with difficult mappings had two possible answers depending on how the participant thought about the question. Therefore, we expected participants faced with imperfect fit to engage in Response-to-Response while they considered which of the two possible mappings is correct. Similarly, we expected participants experiencing misalignment to engage in Response-to-Question as they considered alternate definitions. Alternatively, when participants did not know what a word meant, we did not expect to see regressive movements because they were not deciding between options. Rather, we expected them to either hover on the unfamiliar word, or use the mouse as a marker while they decide the best option.

Information on the type of difficulty participants experienced when exhibiting specific movements could allow survey administrators to provide more tailored help that would address a respondent's specific question or issue instead of general help text that may or may not address their specific problem.

3.3 Analysis

We first describe how we compared the strength of agreement between the two coders and developed a final set of mouse movements to use throughout the analysis. We then discuss how we conducted the descriptive analysis, including how we measured the relationship between the mouse movements and difficulty and how we related different the different types of difficulty to specific movements. Finally, we explain how we developed a model that can be used to predict when respondents are having trouble answering questions.

3.3.1 Reliability

In order to obtain a final set of mouse movements, two undergraduate University of Maryland students were hired to independently code each of the Tobii recordings to determine when participants engaged in each of the pre-specified movements (Table 3.5 in Section 3.2.5). Both coders went through a training which defined each movement of interest (using Table 3.5), showed examples of each movement from past recordings, and involved two training sessions where they coded a recording and then the researcher went through their final codes with them to make corrections or identify movements they missed. The coders were also instructed to provide notes within Tobii to help the researcher understand what they coded.

After the students finished coding all of the recordings, the researcher compared their results. Using the notes and timestamps, the researcher was able to be certain two codes were actually referring to the same event. Kappa values were calculated, using the same methods that were discussed in Section 2.3.1 on page 38, to get a sense of the reliability between coders. Landis and Koch (1997) kappa standards were used to determine the strength of reliability between the coders: ≤0 = poor, .01-.20 = slight, .21-.40 = fair, .41-.60 = moderate, .61-.80 = substantial, and .81-1.00 = almost perfect.

Once the reliability was calculated, the researcher resolved all disagreements. When making a final decision, the researcher was blind to whether the question was associated with a straightforward or complex scenario. Additionally, the final codes were determined as if the instrument had been programmed to track mouse movements. For example, if a participant used their mouse as a marker on a response option, but the cursor was not in what would be defined as the “area of interest” had the tracking been automated, it was not counted. This resulted in a final set of movements for analysis.

3.3.2 Descriptive Statistics Analysis

The first thing we needed to consider was whether participants regularly engaged in the 11 specific mouse movements in which we were interested (Table 3.5 on page 83). In other words, did multiple participants engage in each movement and did they do so multiple times throughout the survey? To measure this, we calculated the total number of occurrences of each movement, using all the screens (ACS questions and probing questions) and also just using the ACS questions. However, it was not enough to say a particular movement occurred over 100 times if only one or two participants engaged in the movement. This type of situation would not help in predicting when a typical respondent is having difficulty. Therefore, we also calculated the percent of participants that engaged in each movement as well as the average number of instances per participants for each movement.

While it is interesting to know how the participants behaved while completing the survey, of primary interest is whether these movements are related to difficulty answering a question. To determine whether participants engaged in more movements when answering a difficult question compared to an easy question, we summed all instances of all of the different movements and compared the average number of movements across each level of difficulty in the probing questions.

To determine whether the individual movements were related to difficulty, we looked at the percent of instances where each movement occurred for participants who rated the question as difficult. For example, if there were 500 instances of using the mouse as a marker and 100 of those instances had a difficulty rating of 'Neither easy nor difficult,' 'Somewhat difficult,' or 'Very difficult,' then the percent of instances related to difficulty would be 20 percent. This percent was calculated for each movement to help determine which movements were most related to difficulty. Using the Marker example, if the average rating for using the mouse as a marker is more difficult for questions associated with a complex scenario than questions associated with a straightforward scenario, but only 20 percent of the instances were rated difficult, then this movement would likely not be helpful in predicting difficulty.

This analysis was conducted twice; once using all the movements and the second time restricting to multiple movements by the same participant on the same question. For example, if a participant used the mouse as a marker two or more times on ‘Relationship’ they were included in this analysis. The basis for this differentiation is that we hypothesize some movements may not be related to difficulty if they only occur once, but this relationship may change for multiple occurrences. Again using the Marker example, of the 500 instances, suppose there were 60 screens on which participants used the mouse as a marker multiple times. Of these, 40 screens were rated difficult. In this case, 67 percent of the instances had a high difficulty rating, which is much more meaningful than the 20 percent overall.

The use of scenarios allowed us to compare the average number of occurrences of each movement between straightforward and complex scenarios. This analysis will help identify which movements may be related to difficulty and which are likely more typical of how people behave while answering questions on the Internet.

Ideally, relating mouse movements to difficulty will provide additional information than using response latencies[16]. We attempted to show the added benefit of mouse movements over response times by making two comparisons simultaneously: the average rating when each movement was and was not present and the average response time when a movement was and was not present. This analysis will help inform whether an increase in response time is necessarily related to an increase in difficulty, or if mouse movements can help identify cases where difficulty exists but response times do not reflect this.

[16] We conducted a supplemental analysis to determine whether response times were longer when a movement was present compared to no movements. This analysis was not included in the Results section, but can be found in Appendix K.

In addition to learning about the different mouse movements the participants made while completing the survey, we also wanted to know if different types of people made different movements, or if the movements were common across demographic characteristics. Specifically, we considered race (White, Black, Asian, Other), age (18-25, 26-35, 36-45, 46-55, 55+), gender, and education (High School, Some College, College Degree, Some Graduate, Graduate Degree, Some Post Graduate, Post Graduate Degree). We compared the mean number of movements within each demographic group to see if different types of people behave differently with the mouse. This information will help determine which demographic variables should be controlled for when creating a model to predict difficulty or if different models should be created for different populations.

3.3.3 Relationship between Mouse Movements and Types of Difficulty

There are many types of difficulty a respondent can experience while completing an Internet survey, but we focused on imperfect fit, misalignment, and technical terms because they are common in government surveys and they are easy to manipulate. The first part of the analysis is similar to the general mouse movement analysis: comparing the frequency of each mouse movement across the types of difficulty and a control group to determine whether some movements were more common for one type of difficulty than another. The control group consisted of questions associated with the straightforward scenarios, since no difficulty was manipulated.

In addition to overall frequencies, which included all questions with all levels of difficulty ratings, the frequencies for each type of difficulty were also calculated for just the instances in which the participant rated the question difficulty as ‘Neither difficult nor easy,' 'Somewhat difficult,' or 'Very difficult.' Finally, we looked at the percent of times each movement, within each type of difficulty, was associated with a difficult rating to determine whether the movement would help in predicting the type of difficulty an individual is experiencing.

3.3.4 Models

The descriptive analysis suggested which movements were likely related to difficulty. However, this information cannot necessarily be directly translated into a model. It is possible there is additional information, such as interactions between movements or the frequency of a single movement, which is essential to fitting a good model, but could not be uncovered through the initial analysis.

To begin developing a model, we needed to decide what type of model most accurately fit the data. Since the dependent variable is binary and not linear, a generalized linear model was used. However, one assumption of this model is that the data are uncorrelated. In the case of mouse movements, the observations have dependence because some participants engage in certain movements more frequently than others. For example, nine of the respondents read horizontally for every question. The meaning of this movement is very different for someone who uses this movement as a general behavior and someone who uses it once or twice across the entire survey. Therefore, it is important to account for this nesting of movements within participant. This can be accomplished by including a random effect in the linear predictor, which incorporates the correlations between the movements of each participant.

Although the overall goal of this analysis was to fit a model that can accurately predict when respondents are having trouble answering a survey question due to confusion, the first model we investigated was quite simple. The purpose of this was to understand each variable and determine which were important to the model. Therefore, for the first model, we treated every movement as a binary variable. The dependent variable, difficulty, was coded as ‘0’ if the probing question was rated as ‘Very easy’ or ‘Somewhat easy’ and was otherwise coded as ‘1.’[17] The independent variables, representing the 11 different mouse movements, were coded ‘0’ if the movement did not occur and ‘1’ if the movement occurred at least one time on a question.

While this model tells us which variables are important predictors of difficulty, it does not utilize all of the information that we have. Specifically, the coders tracked each time each movement occurred on each question, essentially creating counts. However, the model will not be properly specified if we treat all of the variables as counts because Poisson distributions assume that the counts follow a linear trend. Considering few participants engaged in each movement more than one time on a question, it is unlikely the counts will be linear. Therefore, they were transformed into categorical variables. The categorical variables were coded as follows:

- 0 – No movement
- 1 – One instance
- 2 – More than one instance

In addition to distribution issues, there is also a problem with some counts having very few observations. If a cell’s count contained fewer than eight observations, they remained binary (either the movement occurred or it did not occur) in the model. Even in cases of cells containing more than eight observations, if the model appeared mis-specified, the binary indicators were used. The original random effects model fit was:

[17] The models discussed in this section were also run where the dependent variable was coded ‘1’ if the probing question was rated ‘Somewhat difficult’ or ‘Very difficult’. This analysis can be found in Appendix L.

$$y = \beta_0 + \beta_n\mathbf{X}_n + \gamma$$

where y is whether the question was difficult or not, β_0 is the intercept, $\beta\mathbf{X}_n$ where β_n are the regression coefficient parameters and $\mathbf{X}_n$ is equal to each of the mouse movements identified in Table 3.5 on page 83, and γ is the subject-specific random effect.

Using subject-specific random effects ensured we accounted for variation in in how individual participants use the mouse as well as between the specific movements. To model data with this unusual structure, SAS® created a procedure, PROC GLIMMIX, in its 9.2 release. This procedure was specifically designed to perform estimation and statistical inference for generalized linear mixed models which incorporate random effects (Dai *et al.*, 2006). All of the models fit in this chapter use the GLIMMIX procedure.

Before settling on a final model, there were several diagnostics to check in addition to the goodness of fit statistics provided in the GLIMMIX output. The first diagnostic procedure was to check for correlations between the independent variables. To do this, we ran the model multiple times, each time removing one of the independent variables. If the parameter estimates or standard errors changed notably due to the removal of one of the variables, there was evidence that it was correlated with another variable and one needed to be removed. The second diagnostic procedure was to ensure there were not any influential participants that were driving the results. Therefore, the model was repeatedly run, each time removing one of the 100 participants. Again, if there were drastic changes in the parameter estimates or standard errors, this suggested that the results of the regression were driven by one participant and not representative of the data as a whole.

The models described in this section help us pinpoint which movements will help predict respondent difficulty. However, in addition to understanding what these movements mean, we want to know whether they provide more information than response latencies. Therefore, after we determined which movements were related to difficulty, we added response time to the model to see if the movements offer any additional information that response times miss. This would provide support that modeling mouse movements is worth the additional effort and, moving forward, should be used in predicting difficulty.

Mouse movements are even more meaningful as compared to response times if specific movements are associated with specific types of difficulty. This relationship was tested by developing a separate model for each mouse movement:

$$y_{\text{movement}} = \beta_0 + \beta_n\mathbf{X}_n + \gamma + £$$

where y was whether or not each movement occurred, β_0 was the intercept, β_n were the regression coefficient parameters, $\mathbf{X}_n$ was equal to each of the types of difficulty, γ was the subject-specific random effect, and £ was a dummy variable for whether the participant thought the question was difficult. Each movement acted as an independent variable and the treatment (type of difficulty) was the predictor variable. The treatment variable had four levels – control (where the version was 'straightforward'), misalignment, imperfect fit, and technical. Using the 'Contrast' statement in Proc GLIMMIX, we were able to compare whether each movement was significantly more associated with one type of difficulty than another. Additionally, a dummy variable for difficulty, as provided by the participant in the probing questions, was added to the model to control for instances where the version of the scenario was ‘complex’ but the participant did not actually have difficulty answering the question.

3.4 Results

We began by comparing the overall strength of agreement between the two coders and also for each mouse movement. Next, using descriptive analyses, we determined which movements were related to difficulty and attempted to identify individual movements that were related to the specific types of difficulty. Finally, we generated a model capable of predicting when a survey respondent is having trouble answering a question.

3.4.1 Reliability Results

The overall reliability between coders, across all movements and questions was 0.49, which is considered moderate by Landis and Koch’s standards. Out of a total of 3,388 coded movements, the two coders agreed on 1,756 and 1,632 needed to be reconciled by the researcher. The overall reliability in this study is comparable to the first study; both are considered moderate.

In addition to overall reliability, we examined the reliability across the different movements. Table 3.7 provides the kappa value for each movement and its associated strength of agreement.

Considering the coders needed to make somewhat subjective judgments while watching undeliminated video, these reliabilities are quite high. Additionally, the reliabilities match what we would expect based on the distinctiveness of the different movements. In the case of Hover – Next, Hover – Question, and Marker, movements that were right around the 2-second threshold were often coded by one interviewer and not the other. Similarly, in the case of Response-to-Response and Response-to-Space, whether the cursor was on a particular response option often depended on how the student defined the area of interest. For example, does the cursor need to be completely on the response option or can it be just a few pixels outside? Errors like this will not be an issue in practice, however, because the movements, their definitions, and explicit areas of interest will be programmed, thus eliminating the subjective component.

Table 3.7 Reliability and Strength of Agreement by Mouse Movement

Movement	**Kappa**	**Strength of Agreement**
Horizontal reading – Question	0.63	Substantial
Response-to-Next	0.58	Moderate
Horizontal reading – Response	0.57	Moderate
Marker	0.53	Moderate
Hover – Next	0.53	Moderate
Vertical reading	0.50	Moderate
Response-to-Question	0.49	Moderate
Response-to-Response	0.39	Fair
Response-to-Space	0.35	Fair
Hover – Question	0.34	Fair

Reliability for all of the movements was acceptable and most discrepancies were easily reconciled.

3.4.2 Descriptive Statistics Analysis Results

Every participant engaged in at least one movement of interest while completing the survey. Using the mouse as a marker was the most common movement (99 percent of participants) while Response-to-Question and Horizontal reading – Response options were the least common (24 percent and 22 percent of participants, respectively). Table 3.8 shows each movement's frequency and how many participants engaged in each movement across all screens (ACS and probing questions) while Table 3.9 shows only the movements that occurred on ACS questions. The tables show very similar results, with the exception: the order of Vertical reading and Horizontal reading – Question and the order of Response-to-Response and Response-to-Space. This discrepancy arose because a larger percent of Response-to-Space and Vertical reading occurred on probing questions. Participants often read through the list while deciding which to select, so this pattern is not surprising.

Table 3.8 Frequency of Participants Engaging in each Mouse Movement (ACS and Probing Questions)

Movement	Number of Instances	Percent of Respondents[18]	Instances/ Respondent
Marker	838	99	8.46
Hover – Next	305	77	3.96
Vertical reading	256	76	3.37
Horizontal reading – Question	249	49	5.08
Response-to-Response	238	81	2.94
Response-to-Space	214	74	2.89
Response-to-Next	180	66	2.73
Horizontal reading – Response	51	22	2.32
Hover – Question	48	33	1.45
Response-to-Question	39	24	1.63

Table 3.9 Frequency of Participants Engaging in each Mouse Movement on ACS Questions

Movement	Number of Instances	Percent of Respondents	Instances/ Respondent
Marker	773	99	7.81
Hover – Next	282	75	3.76
Horizontal reading – Question	241	49	4.92
Vertical reading	239	75	3.19
Response-to-Response	145	63	2.30
Response-to-Space	195	72	2.71
Response-to-Next	112	52	2.15
Horizontal reading – Response	51	22	2.32
Hover – Question	47	33	1.42
Response-to-Question	39	24	1.63

[18] Because there were 100 participants for this study, the percent of respondents also reflects the actual number of respondents.

Roughly 11.5 percent of all movements occurred on the probing questions. The probing questions were simple questions that the participant expected and saw repeatedly. The majority of these movements were the different regressive movements (other than Response-to-Question) and Marker. In general, we see the majority of movements occurring on the experimental ACS questions. This coincides with our hypothesis that the movements are related to difficulty because the ACS questions are more complicated and less familiar than the probing questions.

A major concern with this research was that a small proportion of participants would engage in these movements and even if there were a lot of observations, they may all be from the same participant. However, Tables 3.8 and 3.9 show that not only do all participants engage in these movements, but the observations are spread out across many participants. In fact, over 20 percent of participants engaged in the least frequent movement. Therefore, there is no cause for concern that these movements are not applicable to a greater population of Web survey respondents and the sample sizes across movement and participant are large enough for a meaningful analysis.

Now that we know what movements participants made, we can determine which were related to difficulty. First, we compared the straightforward and complex scenarios. As predicted, participants averaged 1.2 movements on questions associated with complex scenarios and only 0.7 movements for questions associated with straightforward scenarios $t(1739) = 8.26$, $p < .001$.

However, it is useful to know which specific movements participants made on the complex scenarios compared to the straightforward scenarios so we can begin to tease out which may be related to difficulty. Table 3.10 shows the mean number of instances of each movement when the scenario was complex and the mean number when the scenario was straightforward. The 'Difference' column reflects the mean for complex scenarios minus the mean for straightforward scenarios and the 't-test' column shows the results of the hypothesis test that the means are equal.

The results found in Table 3.10 are somewhat consistent with the hypotheses outlined in Section 3.1. As predicted, the instances of both types of horizontal reading are not significantly different across treatment. However, we did not expect there to be a significant difference for Vertical reading, since this movement is generally a tool to sort through longer lists.

Table 3.10 Comparison of the Mean Number of Movements for Questions Associated with Straightforward and Complex Scenarios

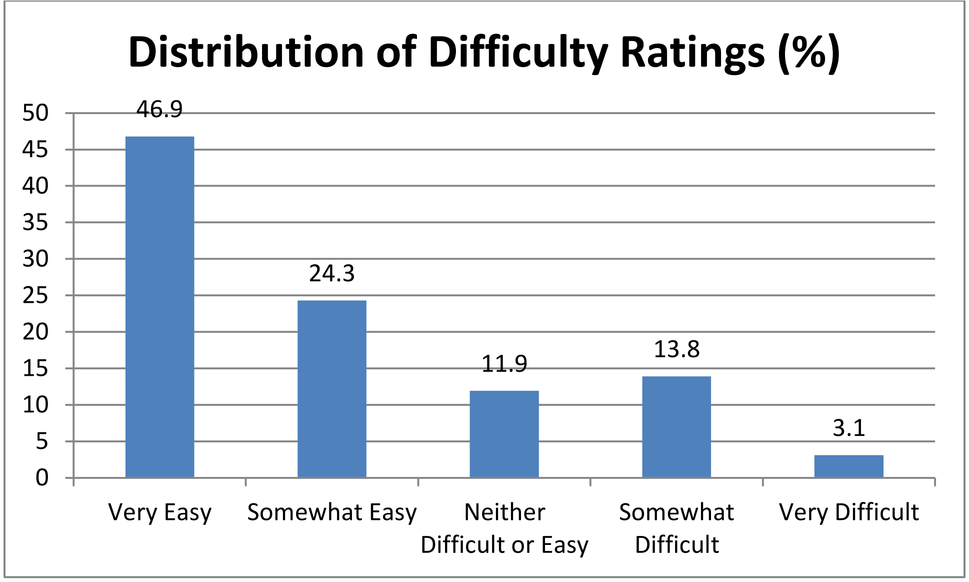
Movement	**Mean - Straightforward (se)**	**Mean – Complex (se)**	**Difference (se)**	**t-test**
Marker	0.27 (0.017)	0.50 (0.026)	0.24 (0.030)	7.64*
Response-to-Space	0.07 (0.009)	0.12 (0.012)	0.05 (0.015)	3.53*
Response-to-Response	0.05 (0.008)	0.10 (0.010)	0.05 (0.013)	3.77*
Hover – Next	0.12 (0.011)	0.16 (0.014)	0.05 (0.017)	2.79*
Response-to-Next	0.04 (0.006)	0.07 (0.010)	0.04 (0.012)	3.22*
Horizontal reading – Response	0.01 (0.003)	0.02 (0.006)	0.01 (0.006)	1.86
Response-to-Question	0.02 (0.004)	0.02 (0.006)	0.01 (0.007)	1.16
Horizontal reading – Question	0.06 (0.008)	0.07 (0.009)	0.01 (0.012)	0.57
Hover – Question	0.02 (0.005)	0.02 (0.005)	0.00 (0.007)	0.16
Vertical reading	0.99 (0.010)	0.14 (0.012)	-0.85 (0.016)	2.55*

*Significant at the alpha = 0.05 level

We expected the occurrence regressive movements to be significantly different between treatments, which they were, other than Response-to-Question. Finally, participants engaged in Hover – Next and Marker significantly more on questions associated with complex scenarios. From the last study we hypothesized Marker was not related to difficulty. However, in this study we included the total number of instances on a single question instead of a binary indicator, which could be driving the difference in results.

A more direct way to measure difficulty was using the probing questions that followed each ACS question. Figure 3.5 provides the distribution of the probing questions as rated by the participants. Ideally, there would be a more even distribution across the difficulty levels instead of clumping at 'Very Easy' and 'Somewhat Easy.' However, since these ratings are subjective, it is difficult to know how participants rated the probing questions as well as the consistency both within and between participants. The entire scale was used so we still have a range of observations to work with.

Figure 3.5 Distribution of Difficulty Ratings



We next compared the average number of movements for each difficulty rating and found the average number of total movements increased as the difficulty rating increased (rated as more difficult) (Table 3.11).

Overall, there was a significant increase in the number of movements as participants reported more difficulty answering questions $F(9,2027) = 28.01$, $p < .001$[19]. At a high level, these results suggest that there is a relationship between difficulty and mouse movements. However, as with the comparison of questions associated with straightforward and complex scenarios, we also wanted to know whether the existence of movements in general was where the relationship lies or if specific movements were related to difficulty.

Table 3.11 Average Number of Mouse Movements Exhibited at each Level of Difficulty

Difficulty Rating	**Average Number of Movements**
Very easy	0.63
Somewhat easy	0.91
Neither easy nor difficult	1.09
Somewhat difficult	1.82
Very difficult	1.65

For each movement, Table 3.12 provides the average difficulty rating when each specific movement was not present and when it was present along with the difference and test statistic. To understand the table, we start by looking at the first row. The 'Rating – no movement' column provides the average rating to the probing questions from all screens on which Response-to-Next did not occur. Similarly, the 'Rating – movement' column provides the same measure on all screens on which that movement did occur. However, it is common for multiple movements to occur on a single screen. Therefore, the first row does not focus on screens where Response-to-Next occurred or did not occur in the absence of all other movements. Rather, other movements are occurring in both

[19] A Newman Keuls test showed that the difference between 'Somewhat easy' and 'Neither easy nor difficult' as well as the difference between 'Somewhat difficult' and 'Very difficult' were not statistically significant.

the ‘Rating – no movement’ and the ‘Rating – movement’ columns. This introduces noise into the measurement, but because there is a comparable mix of noise in both columns, we do not believe it affects the relationships.

Table 3.12 Average Rating Provided for each Mouse Movement

Movement	Rating - no movement	Rating - movement	Difference	t-test	p-value
Response-to-Next	1.96	3.12	1.16	9.20	<.0001
Hover – Question	2.00	3.02	1.03	5.69	<.0001
Response-to- Response	1.97	2.77	0.80	7.05	<.0001
Marker	1.81	2.53	0.73	12.02	<.0001
Horizontal Reading– Response	2.01	2.55	0.54	2.54	0.0146
Response-to-Space	1.97	2.51	0.53	5.15	<.0001
Response-to-Question	2.01	2.44	0.43	1.75	0.0888
Hover – Next	1.98	2.28	0.30	3.78	0.0002
Horizontal Reading– Question	2.01	2.14	0.13	1.20	0.2287
Vertical	2.01	2.13	0.13	1.51	0.1323

Table 3.12 and the results from Table 3.11 confirm our hypotheses and suggest that not all of these movements are related to difficulty. For example, Vertical reading is more a personal difference in how some participants read through the response options.

The final factor to consider when discussing the various mouse movements and what they mean was whether all types of people engaged in the movements equally or whether the movements varied by age and education. We originally planned to measure differences in computer and Internet experience as well, but there was not much variation as all participants had high levels of experience.

While the movements participants engaged in did vary by the level of education they had received, we could not find an obvious pattern where movements were always used most frequently by a particular group. Further, we did not find the same relationships identified in Section 2.4.2.

On the other hand, there was more of a pattern when it came to age, which was also similar to what we saw in Study 2. Younger participants, ages 18-35, moved the mouse significantly more frequently than the other age groups across almost all of the different movements (Horizontal – Question, Hover – Next, Vertical reading, Response-to-Response, and Response-to-Next). Appendix M provides the significance tests across all demographic characteristics and movements. Even where the differences were not significant, the number of movements made by the younger age groups, especially 26-35, were nominally higher for all movements other than Response-to-Space. This difference could occur because younger people have been using computers for most of their lives, whereas older people likely started using computers later. While their level of experience could be equivalent, the way different age groups interact with the computer could be related to when they began using it.

3.4.3 Relationship between Movements and Types of Difficulty Results

To determine whether specific movements are associated with a specific type of difficulty, we looked at the percent of movements associated with each type of difficulty (Table 3.13).

Table 3.13 Distribution of Percent of Mouse Movements across Type of Difficulty[20]

Movement	**% Technical**	**% Imperfect**	**% Misalignment**
Marker	40.56	23.22	36.22
Horizontal reading – Question	39.09	35.45	25.45
Hover – Next	40.43	21.99	37.59
Hover – Question	52.63	42.11	5.26
Response-to-Response	37.04	14.81	48.15
Response-to-Next	33.87	25.81	40.32
Response-to-Space	40.95	19.05	40.00

[20] We only included movements we thought were related to difficulty in this analysis.

The first thing to notice from this table is the imperfect fit column. For all movements other than Horizontal reading – Question and Hover – Question, the percent of the movements associated with imperfect fit is much lower than the other two types of difficulty. This suggests we may not be able to find a good predictor of imperfect fit since most movements appear not to be associated with this type of difficulty.

There are only a few cells in this table that really stand out as potential predictors of type of difficulty: Hover – Question in the '% Technical' column and Response-to-Response in the '% Misalignment' column. These movements occur much more frequently in these cases than they do for the other movements, suggesting they may be important variables in predicting each type of difficulty. Creating a model that can test the relationship between the different movements and types of difficulty, as well as measure the predictive power, will be more helpful in determining which movements are related to the different types of difficulty and will be discussed in the next section.

3.4.4 Model Results

The first model fit used only dummy variables, representing whether or not each movement occurred on a question for a participant. This model appears to fit fairly well, as its Chi-sq/df statistic is close to one (0.84). Looking at the parameter estimates, we see Vertical reading $F(1,1927) = 0.17$, $p = n.s.$, Response-to-Question $F(1,1927) = 1.48$, $p = n.s.$, and Horizontal reading – Response $F(1,1927) = 0.95$, $p = n.s.$ were not significantly predictive of difficulty. After removing these variables from the analysis, we reran the model. To make the analysis more clear, we focus on odds ratios instead of the actual parameter estimates. Table 3.14 provides information on how much more likely it is that someone who engaged in each of the movements had difficulty answering a question as compared to someone who did not make the movement.

Table 3.14 shows that participants who hovered over the question text with the mouse were 6.75 times more likely to have difficulty answering a question than participants who did not. It appears that Hover – Question, Response-to-Next, and Marker had the most predictive power, but all of the variables contributed to predicting difficulty. We originally hypothesized that Horizontal reading was not related to difficulty. While Horizontal reading – Response is not, Horizontal reading – Question is. This is likely because we were able to control for individual behaviors through the random effects, which we were not able to do in the previous analyses. By controlling for clustering within participant, we were able to see that this movement was actually related to difficulty as long as the participants were not Readers.

Table 3.14 Fit Statistics and Odds of each Mouse Movement being Predictive of Difficulty (Dummy Model)

Fit Statistics	
Participants	100
Observations	2037
-2 Res Log Pseudo-Likelihood	9806.58
Generalized Chi-Square	1705.54
Generalized Chi-Square/DF	0.84

Movement	**Odds Ratio Estimate**	**Confidence Interval**
Marker	3.31	2.58 - 4.23
Hover – Question	6.62	3.08 - 14.25
Response-to-Next	5.39	3.22 - 9.02
Response-to-Response	2.47	1.60 - 3.84
Response-to-Space	2.23	1.50 - 3.35
Hover – Next	1.49	1.05 - 2.10
Horizontal reading – Question	1.60	1.06 - 2.42

This model gives us a good starting point in our analysis. However, we tracked how often each movement occurred which, if included, may strengthen the relationship between the different movements and difficulty. Unfortunately, for most variables, there were very few instances where the counts were greater than one. Specifically, Response-to-Response, Hover – Question, and Horizontal reading – Question had fewer than eight observations of counts greater than one, so these variables continued to be treated as dummy variables. For the other variables, while there were a sufficient number of observations where a movement occurred more than one time on a question, as the counts went above two, the incidence rate was very low and the distributions were very skewed. Therefore, we decided to use categorical variables instead of counts.

The first attempt at fitting a model with these new specifications included the following variables:

- Hover – Next (3 Categories)
- Hover – Question (Binary)
- Marker (3 Categories)
- Horizontal reading – Question (Binary)
- Response-to-Response (Binary)
- Response-to-Space (3 Categories)
- Response-to-Next (3 Categories)

For the categorical variables, the three categories corresponded to ‘Zero,’ ‘One,’ and ‘Multiple’ movements. This model did not meet the convergence criteria, likely due to the small cell sizes of Response-to-Next and Response-to-Space. Therefore, these variables were converted back to a dummy variables and we re-fit the model.

The new model met convergence criteria and offered an overall good fit with a Chi-sq/df of 0.82. All of the variables were significant predictors of difficulty except the multiple level of Hover – Next. This estimate had a large standard error, likely due to the small cell size. Therefore, this variable was also converted back to a dummy variable, resulting in Marker being the only remaining categorical variable.

After checking the model diagnostics, we found a strong correlation between Hover – Next and Response-to-Next. When Response-to-Next was removed from the model, the parameter estimate for Hover – Next became insignificant and the standard error more than doubled. However, the same effect did not occur to Response-to-Next when Hover – Next was removed. Therefore, Hover – Next was removed from the model and Response-to-Next was retained.

We also checked this final model for influential participants. Specifically, we wanted to determine whether a single participant was driving the results. There was no single respondent that drastically altered the results from the model; removing respondents did not lead any of the parameter estimates to change more than 0.10 and almost all changes were less than 0.05. Additionally, all standard errors remained similar to the complete model as did the p-values.   The movements that were most susceptible to influence were Response-to-Response, Horizontally reading – Question, and Hover – Question. Interestingly, in cases where the parameter estimates changed more than 0.05 for Hover – Question, the relationship was actually strengthened. So, by keeping these participants in the model, we may be underestimating the actual effect of this movement. To help identify outliers, Appendix N provides scatterplots of the differences between the final parameter estimate for each movement and the parameter estimate after removing each participant. The differences did not seem large enough to require removing these participants, so all participants were used for the final analysis.

Horizontal reading – Question was also affected by a handful of participants, but not in the same direction as Hover – Question. There were six participants which, when removed, resulted in Horizontal reading – Question becoming insignificant. The actual changes in the parameter estimates were less than 0.05, but this movement was one of the weaker predictors. Even after removing these cases, the movement was still marginally significant with p-values around 0.0503. Therefore, we retained them in the model, but kept this result in mind when we drew conclusions about how to use this information moving forward (Section 3.5).

The final step in generating a final model was to check for interactions. We used a saturated model to check every potential interaction; including all interactions in a single model and testing each interaction independently. However, none of these interactions were significant. The output from the final model can be found in Table 3.15. We can see that all of the variables are significant predictors of difficulty. Additionally, looking at the odds ratios allows us to more easily understand the impact of each movement.

As we saw with the binary model, Hover – Question is the movement most likely to suggest difficulty. Additionally, the relative effect of using the mouse as a marker changed after converting it to a categorical variable. When Marker was a dummy variable, it increased the likelihood of rating a question as difficult by 3.31 times. However, it is significantly more predictive when the movement is engaged in multiple times on a question. This result was expected because many participants used the mouse as a marker on straightforward and complex questions alike just to keep their place while they browsed other options. Therefore, the additional movements signify more indecision or confusion as opposed to being a placeholder.

Table 3.15 Results from Final Logistic Regression Model using Mouse Movements to Predict Difficulty

Fit Statistics	
Participants	100
Observations	2037
-2 Res Log Pseudo-Likelihood	9783.08
Generalized Chi-Square	1717.99
Generalized Chi-Square/DF	0.85

	Parameter		
Variable	**Estimate**	**Std Err**	**p-value**
Intercept	-1.81	0.15	<.0001
Marker-Multiple	1.71	0.23	<.0001
Marker-One	1.00	0.14	<.0001
Marker-Zero	.	.	.
Response-to-Response	0.87	0.23	0.0001
Response-to-Next	1.73	0.26	<.0001
Response-to-Space	0.79	0.21	0.0002
Hover – Question	1.96	0.41	<.0001
Horizontal Reading – Question	0.47	0.21	0.0266

	Odds Ratio	Confidence
Movement	**Estimate**	**Interval**
Marker - Multiple vs Zero	5.52	3.52-8.64
Marker - One vs Zero	2.71	2.07-3.54
Response-to-Response	2.38	1.52-3.72
Response-to-Next	5.66	3.38-9.48
Response-to-Space	2.20	1.45 – 3.55
Hover – Question	7.10	3.20-15.87
Horizontal Reading – Question	1.60	1.06-2.42

While the model above represents the final set of mouse movements that are predictive of difficulty answering a question, we need to determine whether this model provides any information that cannot be obtained simply by using response latencies. Therefore, we added a response time dummy variable (0 = less than median response time, 1 = greater than or equal to median response time) to the model. Table 3.16

compares the odds ratios of the original model to the model with the response time variable included.

Table 3.16 Comparison of Odds Ratios with and without Response Time

	Movements Only		Movements and Time	
Movement	**At least one movement vs none**	**Multiple movements vs none**	**At least one movement vs none**	**Multiple movements vs none**
Hover – Question	7.10		5.72	
Response-to-Next	5.66		4.39	
Response time	N/A		3.32	
Marker	2.71	5.52	2.01	3.54
Response-to-Response	2.38		2.27	
Response-to-Space	2.20		1.86	
Horizontal Reading – Question	1.60		1.59	

As expected, response time was a significant predictor of difficulty. So much so, that it explained some of the effect of all of the other variables. This is not surprising since each of these movements takes time to complete, especially Hover – Question, Hover – Next, and Marker, which require at least two seconds. However, all of the movements are still significant after adding time to the model. This shows that mouse movements to add value to the response time only model currently in use.

From the exploratory analysis, we saw that younger participants engaged in movements more frequently than older participants, suggesting that age might be an additional predictor. Using the categories described in Section 3.3.2, and the results from that analysis, age was divided into two groups: 35 and younger and older than 35. This variable was a significant predictor of difficulty $F(1,1914) = 7.00$, $p = .008$, and its interaction with Marker was also significant $F(2,1912) = 3.20$, $p = .041$. However, its usefulness in practice is questionable. The purpose of this analysis was to design a model that could be used in real time to predict when respondents are having difficulty. In most cases, respondent age is not known at the beginning of the survey, and there are some surveys where demographic questions are asked at the end, so age is not known for the majority of the survey questions. Therefore, while age may be a significant predictor, it is not practical to include in the final model.

The previous analyses showed that mouse movements provide more information than using response times alone to predict respondent difficulty. It would be even more helpful if tracking the movements could provide insight into the type of difficulty a respondent is having. This would allow researchers the opportunity to provide tailored help to those who need it. However, it does not appear that mouse movements are related to the types of difficulty manipulated in this study: misalignment, imperfect fit, and technical terms.

Focusing on the movements we thought may be related to a specific type of difficulty from the exploratory analysis, Response-to Response and Hover – Question, we did see that Response-to-Response was associated with misalignment. However, it was not significantly more predictive than technical difficulty $F(1,1597) = 0.75$, $p = n.s.$ Similarly, Hover – Question was related to technical difficulty, but not significantly more than imperfect fit $F(1,1597) = 0.0$, $p = n.s.$ Because we were looking for only one type of difficulty that was associated with each movement for the goal of tailoring help, if multiple types of difficulty are associated with a movement, it will not improve tailoring, and is therefore not a significant finding for this analysis.

While there were no significant findings in the type of difficulty for this study, it is likely because the difference between the types of difficulty, especially misalignment and imperfect fit, were too subtle to result in different mouse movements. Therefore, it is still possible that mouse movements can predict other types of difficulty not studied here.

3.5 Conclusions

In this chapter we tested the hypotheses generated in Chapter 2. We identified six movements that were related to respondent difficulty: Marker, Hover – Question, Horizontal reading – Question, Response-to-Response, Response-to-Space, and Response-to-Next. From the preliminary analysis we laid out the following hypotheses:

- Instances of Horizontal reading, Vertical reading, and using the mouse as a marker **are not** associated with an increase in difficulty
- Instances of hovering with the mouse **may be** associated with an increase in difficulty
- Instances of regressive movements **are** associated with an increase in difficulty

Most of our hypotheses were confirmed. Vertical reading was not significant to the model. Marker was a significant predictor of difficulty, but creating a categorical variable instead of a binary variable increased its predictive power. Horizontal – Question was a significant predictor although we predicted it would not be. This was likely because we were able to control for the people who read every question fully while following with their mouse. Additionally, this movement is influenced the most by individual participants and it is the weakest predictor. The regressive movements, other than Response-to-Question, were significant predictors, as expected. Response-to-Question was likely not a significant predictor due to its low incidence rate. In the first study we did not distinguish between hovering over the ‘Next’ button and the question text. This was clearly an important distinction to make because Hover – Question was a significant predictor while Hover – Next was not.

We were originally unsure of the meaning of hovering and did not think that using the mouse as a marker was related to difficulty because these were the most common movements and were often used to double check answers. The finding that Marker was more predictive when there were two or more instances on a question helps support this hypothesis. On the other hand, Hover – Next was only significant when Response-to-Next was in the model, suggesting the two are related and hovering on its own is not significant, again clarifying the originally ambiguous hypothesis.

In addition to identifying a set of movements that are indicative of difficulty, we also showed that mouse movements provide more information than just using response times to predict when respondents are experiencing difficulty. To begin with, the movements guarantee the respondent is engaged in the survey task and not away from the computer or checking their email in a separate window. Further, having multiple indicators provides more precision in the prediction that a person is having difficulty with a question. This will likely result in fewer people receiving help when they do not need it.

Taking all of these findings together, we propose modeling the following movements to maximize the likelihood of identifying respondents who are having difficulty: Hover – Question, Response-to-Next, Marker after the first occurrence, Response-to-Response, and time (if it is available). Although Response-to-Space and Horizontal reading – Question were significant predictors, we do not suggest including them because people who engaged in these movements were less than twice as likely to experience difficulty. Since the ultimate goal is to provide help to respondents, we do not want to drastically over-estimate who is having difficulty or there may be a negative effect, such as survey abandonment.

Despite the benefits of using mouse movements, they were not able to predict the specific type of difficulty respondents experienced. One possible reason is that the movements simply do not have that power. Alternatively, there may not have been enough differentiation between the types of difficulty manipulated in this experiment or the movements are more capable of predicting other types of difficulty that we did not look at. In the first case, the scenarios may not have actually induced the exact type of difficulty intended. Secondly, there are many types of difficulty respondents can experience (Tourangeau *et al.,* 2006; Tourangeau *et al.,* 2000) and the movements may be better at identifying these types of difficulty as compared to the three we manipulated. Additional testing could have participants specify what they found difficult about different questions and their responses could be grouped into several difficulty categories. This would ensure the respondents actually experienced the difficulty and would test a broader range of types of difficulty.

Chapter 4: Experimenting with Mode of Help Administration to Optimize Accuracy and Satisfaction

4.1 Background

Chapters 2 and 3 of this dissertation proposed a new way to identify Web survey respondents that are having difficulty answering a question. These new indicators, mouse movements, can be used to test the quality of survey questions and determine whether specific questions are difficult for respondents to answer. Once survey administrators identify difficult questions, improvements can be made to the questions through survey revisions or in real time.

However, even well-written, clear questions can be difficult for some respondents and these same difficulty indicators can be used to identify these respondents as well. If survey administrators know a respondent is having trouble answering a question, they can intervene and provide them with help in real time. The goal of providing this help is to increase accuracy. To increase accuracy, however, it is important that respondents pay attention to the help provided. It is also important that the help is not frustrating or annoying or the respondent may abandon the survey.

Across survey modes, there is strong evidence that providing unsolicited help in surveys does increase accuracy (Schober and Conrad, 1997; Ehlen *et al.*, 2005; Conrad *et al.*, 2007). Specifically, on questions associated with complicated mappings, Schober and Conrad (1997) saw accuracy increase from 28 percent when no help was provided to 87 percent when the interview was conversational in nature and the respondent could obtain help. Similarly, Ehlen and his colleagues (2005) showed that for complicated mappings, accuracy increased from 20 percent to 64 percent when help was provided. Finally, in Conrad and his colleagues' experiment (2007), respondents answered 24 percent of questions associated with complicated mappings correctly when no help was provided, but answered 48 percent correctly when provided unsolicited help.

These studies show that in some cases respondents need help answering survey questions, even if they do not realize it, and providing them with assistance can have a large impact on accuracy. In Web surveys to date, the help provided to respondents has taken the form of a standard text box (Lind *et al.*, 2001; Conrad *et al.*, 2006; Conrad *et al.*, 2007). While this is the standard design, text boxes do not take advantage of the technology available in Web surveys. It is possible that we can use the versatility of the Web to provide help in another mode that could further increase accuracy.

The basis for this hypothesis comes from the education literature. Kalyuga (1999) explained that a combination of modes can help people process data because they can focus on two things at the same time. Specifically, when students were shown a graphic with audio describing the graphic content, they understood the information better than when a text explanation accompanied the graphic. This is because processing too much information in a single sensory mode can result in overload. Human brains are wired to multi-task, so processing information from multiple sensory modes simultaneously is a natural ability (Baddeley and Hitch, 1974; Penny, 1989; Kalyuga, 1999; Fang *et al*., 2006).

Although a survey context is different than educational learning, in the education example, the students were using two modes to process two related pieces of information at the same time. Similarly, in a Web survey, the help content is information that clarifies the question. It is possible that respondents will be able to process help from another sensory mode more easily than from the same mode. For example, if the help is in text format, respondents can only look at the question/response options or the help text at one point in time. However, if the help is auditory, they can peruse the response options or re-read the question while listening to the help content. Therefore, we believe that providing audio help will increase accuracy compared to a standard text box because the respondents will have an easier time processing the information.

In addition to trying to increase accuracy through multi-modal processing, making the interaction with the computer more like interacting with another person could increase respondents' attention, which in turn could also increase accuracy. Specifically, in education, students taught by animated agents rated their learning experience as more entertaining than students who were not taught by agents (Barlow *et al*., 1997). Additionally, through his research on multimodal interfaces, Johnston (2007) expects that a more human interaction with a computer should make the user feel more engaged in the task. Finally, providing a computer interface with human attributes and the ability to interact with its users can also lead users to treat computers as social actors; thereby not offending them or hurting their feelings (Nass *et al.*, 1999). This could result in respondents paying closer attention to the information being provided because they do not want to be rude.

While users report enjoying using systems that are more human-like, it is not clear yet what the actual impact is on learning or survey responses. Specifically, Tourangeau and his colleagues (2003) experimented with using photos of the investigator and tailored feedback to create a more personal experience, but they did not see an impact of this manipulation on most survey responses or breakoffs (the only effect they found was on gender attitudes when the photo was of a female). On the other hand, there is some evidence that adding just a voice feature resulted in higher scores in a learning environment (Mayer *et al.*, 2003). However, this may be more a result of multi-modal presentation than the impact of the voice alone.

Even if creating a more conversational environment with the computer does not increase accuracy through increased attention, there may still be an added benefit. Specifically, computer users report preferring these types of human-like interactions with computers to standard machine-like interactions. For example, Barlow and his colleagues' education study (1997) found that animated agents had a positive effect on subjects' perception of their learning experience as well as making the experience more entertaining. Similarly, students rated lessons using human voices higher than lessons using computerized voices (Mayer *et al.*, 2003; Atkinson *et al.*, 2005) and Sproull (1996) found that users rated their experience as more satisfying when the computer had human-like attributes. In the survey literature, Conrad and his colleagues (2007) found that respondents liked having the ability to interact with the instrument to obtain help, although this was just in the form of requesting help.

Increasing satisfaction in providing help is important because there is some evidence that unsolicited help decreases respondent satisfaction. Conrad and his colleagues (2007) conducted an experiment where respondents either received no help, were able to request help, or had the system initiate help when respondents surpassed a predefined threshold of time to answer the question (there were two system-initiated help conditions; the first used the same threshold for all respondents while the second used one threshold for younger respondents and a longer threshold for older respondents). At the end of the survey, they asked respondents to rate their satisfaction with their experience. Respondents were no more satisfied with receiving system-initiated help than they were to not being able to receive any help at all. However, they rated satisfaction highest when they could request help. The researchers concluded that respondents liked having the ability to request help if they want it, but did not enjoy being provided with help when they did not request it. Therefore, it is possible that making the help environment more conversational could increase satisfaction with receiving unsolicited help.

The literature on providing help in Web surveys has mainly focused on how help should be provided: only if requested, always available, or if difficulty is detected (Conrad *et al.*, 2007; Lind *et al.*, 2001). In these studies, help was provided to respondents in a text box. It is possible that providing help in more personal and human-like modes could increase accuracy by creating more efficient information processing and generating more engagement. Additionally, respondents may enjoy receiving help if they feel like they are interacting with a human.

This study proposes two modes, in addition to standard text, in which help could be provided maximize respondent accuracy and satisfaction. We believe that providing respondents help via an audio recording with a human voice will allow them to process the help content and the question/response options simultaneously. We expect this ease in processing to result in higher accuracy for people who receive audio help compared to those that receive text help. Additionally, audio help also is more conversational than text help, so respondents may be more satisfied with it as compared to text. On the other hand, satisfaction could actually be lower with audio help because the information must be processed when the recording begins and at the speed of the recording, while text help can be processed at the respondent’s leisure and at the speed the respondent desires.

The other mode we propose that could increase accuracy and satisfaction is providing help via a chat, or instant message, between the respondent and an agent. Chatting is essentially an interactive conversation with the computer. In fact, the respondent may not even know whether they are chatting with a real person or a computer. Even if the computer is programmed to chat with the user and it is not a live agent at the other end, the chat can still be smart and responsive to needs as well as be personal and friendly. We believe that these human-like features will lead to increased satisfaction compared to text help.

Additionally, chats can address specific user issues instead of generating the same standardized response for everyone. This means that respondents only receive the information they want and not superfluous information that could confuse or annoy them more. Therefore, it is also possible that respondents receiving chat help will be more accurate. However, for this to happen, they have to ask the agent the correct question, which places the burden on the respondent.

The study described in this chapter aims to identify a mode in which to provide help that maximizes accuracy and satisfaction.

4.2 Study Design

Participants involved in this study answered 20 questions from the ACS based on pre-written scenarios designed to make some of the questions more difficult to answer. To help participants answer these questions, participants received help in one of three modes: text, audio, or chat. This section describes the questions, scenarios, and help used in this study, as well as the data collection procedures and demographic characteristics of the participants.

4.2.1 Web Survey

The data for this study consisted of the same 20 questions taken from the Web version of the ACS that were used in the first two studies along with the probing questions following each ACS question (Appendix N). In addition to the 20 ACS and probing questions, they also answered set of demographic and computer experience questions before beginning the survey. Finally, after completing the survey, participants answered between four and 12 questions regarding their satisfaction with the survey and the help they received. Participants received a different set of questions depending on whether they requested help, received model-initiated help, or did not receive any help (Appendix P).

4.2.2 Scenarios

This study asked participants to answer the survey questions based on prewritten scenarios for two reasons: to ensure participants experienced difficulty and so there was one correct answer for every question, thereby providing a measure of accuracy. The scenarios largely match those used in the second study described in Chapter 3. However, there were changes to 13 scenarios due to participants' ratings on the probing questions and comments we received during the debriefings in the second study. For example, the straightforward version of the Hispanic question was actually difficult for some participants, so it was altered for this study to make it less confusing. In other cases, such as with the Internet question, the help text did not address the manipulation in the scenario. In these cases, the scenarios were updated to reflect the help that we could provide. The scenarios that were updated are identified in Appendix O.

We randomized which scenario participants received so that, in total, each participant received 10 questions associated with straightforward scenarios and 10 questions associated with complex scenarios. In the end, after all of the sessions, for each question, 75 participants received the straightforward version and 75 received the complex version. The same randomization technique was used in this study as in the second and a description of it can be found in Section 3.2.2 on page 74. This balance allows us to compare both across scenario types and within each question and any order effects should be minimized.

Participants were exposed to the scenario before seeing the ACS question or help, if they received help. For this study, we wanted the ambiguities in the scenarios to become evident when participants read the question, thereby mimicking an autobiographical survey. For example, ACS Internet respondents sit at a computer in their homes, offices, or libraries and answer the questions about their own lives; we wanted to recreate this type of experience for the survey participants. Participants could reread the scenario by clicking the back navigation arrow on the toolbar or by clicking the “Scenario” link in the upper left had on the survey2. While a handful of participants used these tools to look at the scenarios multiple times, we did not explicitly direct participants to these options because we wanted to simulate an autobiographical setting as much as possible.

4.2.3 Help

Participants could receive help by requesting it using a link that followed the question text or they could be offered help automatically if they exceeded a predetermined time threshold. Each question had its own threshold, calculated from the median time it took participants in the second study to answer the complex version of each question. We used the median time instead of the mean to help account for outliers, or people who took an abnormally long or short time to answer the question. We felt using the median response time would tell us how long someone should take to answer each question under typical conditions and longer response times would likely be related to difficulty. If a participant used the help link, the timer was deactivated so he or she did not receive help twice.

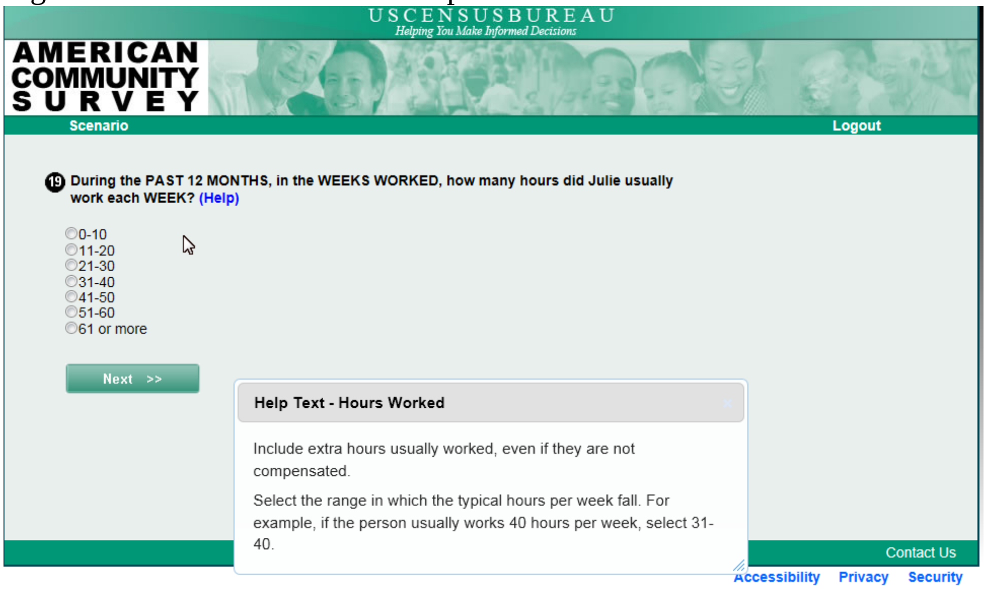
Most of the help content was taken from the help provided in the production version of the ACS. Some questions already had help associated with them, while others had probes from the CATI/CAPI interviewer manual that we converted to help content to address the specific manipulation in the complex scenario. Additionally, we made sure that the information provided in the help directly addressed the manipulation from the scenario so if the participant read the help, he or she would be able to answer the question accurately. Appendix O provides the help text associated with each question.

Participants received help in one of three modes: text, audio, or chat. We did not include a control group in this study (no help) because prior research has already shown the effect of providing help compared to no help (Schober and Conrad 1997; Lind *et al.*, 2001; Conrad *et al.,* 2006; Conrad *et al.,* 2007). For this study we wanted to focus on the effect of administration mode. Fifty participants received the help in each mode. We were not able to randomly assign participants to each mode because the chat condition was not ready at the time the study started. Therefore, data collection for that mode occurred independently and at a later date than the other two. Although, we were able to make the groups demographically similar by assigning 25 students and 25 Craigslist respondents to each mode as in the text and audio modes, there may be a threat to the internal validity of the study. Specifically, we cannot say with certainty whether differences in the chat mode are a result of the chat function or the sampled participants.

Text Help

The help in the text mode appeared as a standard pop up text box (Figure 4.1). Participants could change the size of the box, move the box, and close it. There were not any interactive features and the help content was provided in either sentences or bulleted lists depending on the question. Participants could move, resize or close the text box.

Figure 4.1 Screenshot of Text Help Box



19 During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did Julie usually work each WEEK? (Help)

○0-10
○11-20
○21-30
○31-40
○41-50
○51-60
○61 or more

Help Text - Hours Worked

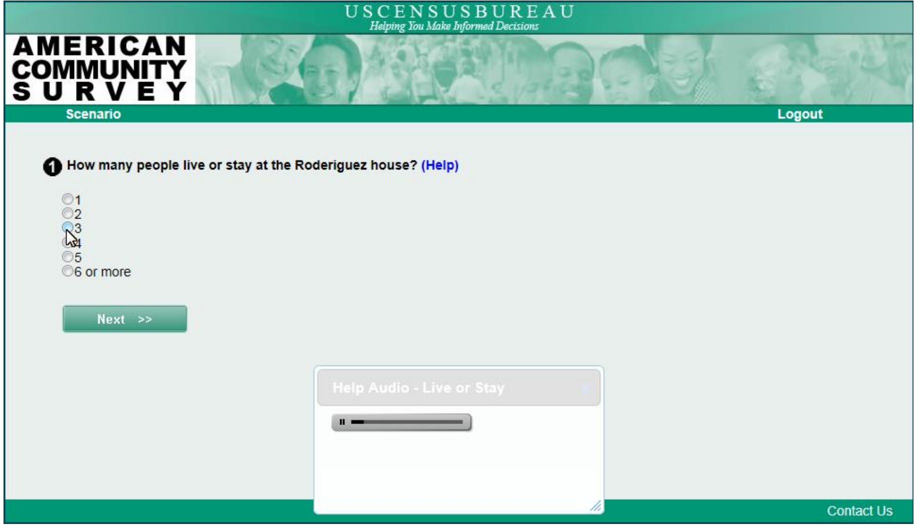
Include extra hours usually worked, even if they are not compensated.

Select the range in which the typical hours per week fall. For example, if the person usually works 40 hours per week, select 31-40.

Audio Help

In the audio mode, the help was provided to participants in a recording player (Figure 4.2). The player could be paused and restarted and participants could listen to specific parts of the help by clicking the progress bar of the player. Additionally, the player could be moved, resized, and closed. The content of the help was identical to that in the text condition; the researcher (female) read the text help aloud to generate the recordings for each question.

Figure 4.2 Screenshot of Audio Help Recording


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Help Audio - Live or Stay
Contact Us

1 How many people live or stay at the Roderiguez house? (Help)

○1
○2
○3
○4
○5
○6 or more

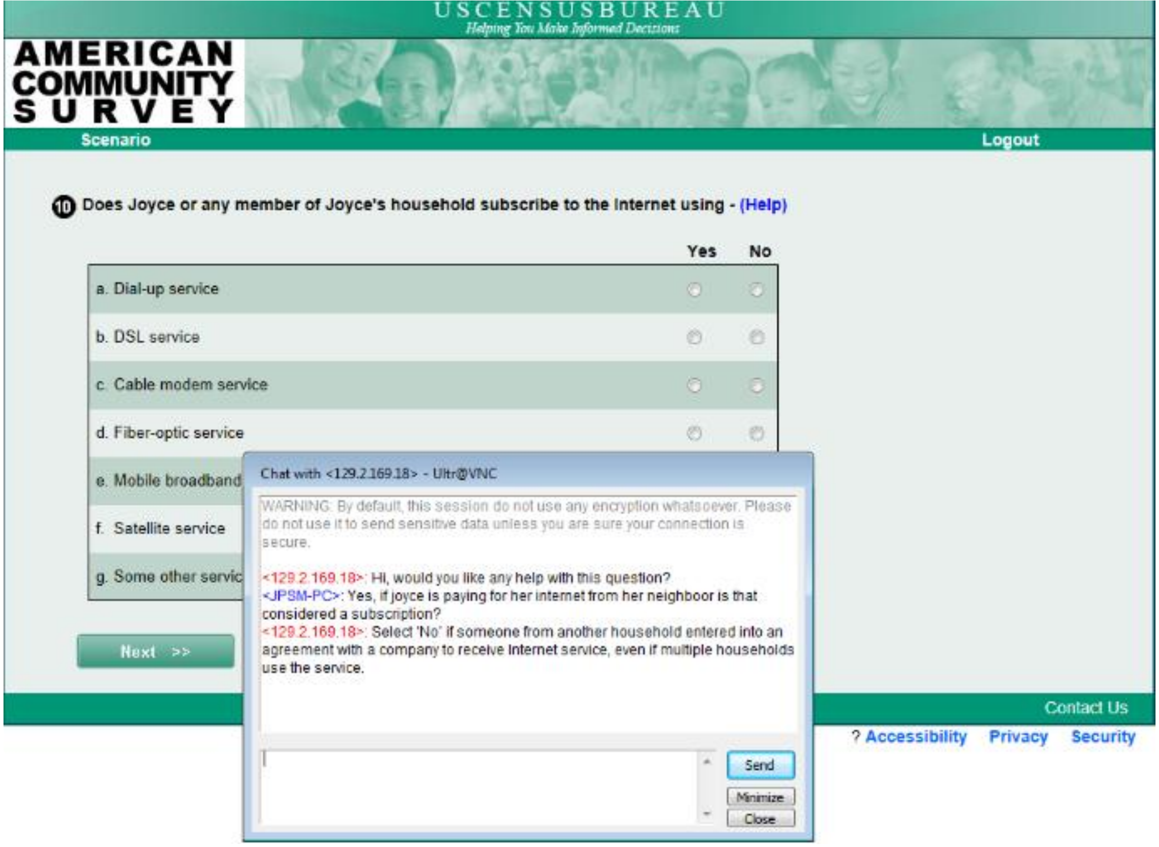
Chat Help

Due to cost issues at the Census Bureau, we were unable to activate the chat window as part of the survey instrument. Instead, we set up a remote connection between the participant's computer and the researcher's computer that included a chat function. When the median response time had elapsed, a small indicator appeared in the bottom corner of the screen. When the researcher saw the indicator, using the remote connection, she activated the chat window, typed the greeting, and sent the message. To account for the time required to complete these steps, the help initiation time was reduced by 1.5 seconds.

Another effect of the researcher needing to activate the chat window and it not being a part of the instrument was that the chat box did not pop up in the middle of the screen like in the other conditions. Rather, when the chat sent from the researcher's computer to the participant's, there was a tone and a tab flashed on the toolbar at the bottom of the screen. The participant then had to click the tab to see the chat. Participants were instructed to activate the window when they heard the tone.

Upon activating the chat, participants received a message asking if they would like any help (Figure 4.3). Although the chats were sent by the researcher, the participants did not know whether they were talking to a live agent or a computerize agent. We attempted to increase personalization by adding "Hi" before asking the participant if they needed help and responding with "you're welcome" if the participant thanked the agent.

Figure 4.3 Screenshot of Chat Help Box



⑩ Does Joyce or any member of Joyce's household subscribe to the Internet using - (Help)

	Yes	No
a. Dial-up service	○	○
b. DSL service	○	○
c. Cable modem service	○	○
d. Fiber-optic service	○	○
e. Mobile broadband		
f. Satellite service		
g. Some other servic		

Chat with <129.2.169.18> - Ultr@VNC

WARNING: By default, this session do not use any encryption whatsoever. Please do not use it to send sensitive data unless you are sure your connection is secure.

<129.2.169.18>: Hi, would you like any help with this question?
<JPSM-PC>: Yes, if joyce is paying for her internet from her neighboor is that considered a subscription?
<129.2.169.18>: Select 'No' if someone from another household entered into an agreement with a company to receive Internet service, even if multiple households use the service.

All help provided to participants in the chat condition used the same wording and information as was used in the other two conditions. However, depending on what question the participants asked, they may not have received all of the content. For example, the full help information for the vehicles question is:

Include in count:

- Cars, vans or SUVs if they are
  - Regularly kept at home AND
  - Used by household for nonbusiness purposes

Do NOT include in count:

- Recreational Vehicles
- Motorcycles

Participants in the text and audio modes receive all of this information. However, in the chat mode, if a participant asked, “Does an RV count as a vehicle?” the chat response was: “Do NOT include in count: Recreational Vehicles; Motorcycles.” Therefore, depending on what question the participant asked, they received the same wording, but not necessarily all of the information. In cases where the participant asked a question for which the response was not part of the script, the researcher responded with “I’m sorry, I don’t have the answer to that question.”

4.2.4 Participants

For this study, we recruited 150 individuals[21] from the Washington, DC metropolitan area in October of 2012. Participants were recruited using flyers posted around the University of Maryland, postings to the University of Maryland’s list serves, and print advertisements on Craigslist (Appendix I). Due to the larger sample size of this

[21] The sample size of 150 participants resulted from a power analysis, which allows us to detect differences in the percent of accurate responses of 0.25 with a power of 0.80.

study, half of the participants were recruited from the University of Maryland and half from Craigslist.

After responding to an advertisement, potential participants answered a set of screener questions (Appendix J). Anyone who participated in either of the first two studies described in this dissertation or had participated in three or more research studies in the past year was not eligible to participate. Further, only people that had completed high school were eligible. Table 4.1 provides the demographic characteristics of the study participants.

Table 4.1 Demographic Characteristics of the Participants

Gender	**Percent of All Participants**	**Percent of Text Participants**	**Percent of Audio Participants**	**Percent of Chat Participants**
Male	38.7	36.0	34.0	46.0
Female	61.3	64.0	66.0	54.0
Age				
18-29	64.0	62.0	64.0	66.0
30-39	13.3	14.0	10.0	16.0
40-49	14.0	10.0	18.0	14.0
50-59	4.0	8.0	4.0	0.0
60+	4.7	6.0	4.0	4.0
Education				
HS or equivalent	6.0	8.0	6.0	4.0
Some college	52.7	48.0	48.0	62.0
Associate's Degree	4.7	4.0	6.0	4.0
4-year college Degree	19.3	20.0	24.0	14.0
Some Graduate school	7.3	8.0	10.0	4.0
Post Graduate Degree	10.0	12.0	6.0	12.0
Race				
White	46.0	44.0	46.0	48.0
Black	38.7	32.0	42.0	42.0
Asian	9.3	22.0	2.0	4.0
Other	6.0	2.0	10.0	6.0

Our goal was to have the participant's demographic characteristics balance across modes. While there were some differences, we believe the balance is quite good considering we used a convenience sample and the chat participants were recruited later than the other two modes. Where there are larger differences (race and gender), we have no reason to believe that these characteristics will interact with satisfaction or accuracy.

Across all modes, we recruited more young participants than old. This is a result of half of the sample being drawn from a university. While not balanced, we were able to obtain a wide variety of age groups so limited testing can be done. In practice, we do not believe that a young sample limits our findings because research on the ACS has shown that younger people are more likely to use the Web to complete the survey (Tancreto *et al.*, 2012).

All of the participants came to the JPSM office at the University of Maryland to complete the survey and were reimbursed $30 for time and travel.

4.2.5 Data

The Tobii X-60 eye tracking hardware and Studio Enterprise software were used to capture eye movement data. Although eye movements were not the focus of this study, its functionality was still useful. Specifically, the Enterprise software generates a recording of each participant's session. The recording captures the participant's computer screen and everything that occurs on it for the duration of the Tobii session. In addition to general screen capture, the recordings can also show where mouse clicks occur and provide an eye gaze overlay. We used these recordings to code on which questions participants received help and which type of help they received (model-initiated or requested). Additionally, the recordings can also help us understand how participants used the help. For example, one of the potential benefits of audio help is that the participant could process the question and response options while simultaneously listening to the help. The eye tracking data allowed us to know if respondents actually behaved this way or if they just looked at the player.

4.2.6 Data Collection

To begin, participants came to the JPSM office on the University of Maryland campus. They were seated in front of a computer equipped with Tobii eye tracking hardware and software. After listening to a short description of the survey, read by the researcher, participants read and signed a consent form and answer a set of questions about their demographic information and computer experience (Appendix E). The researcher then calibrated the participant's eyes for the eye tracking and explained to them how to navigate through the instrument, read the scenarios, and answer the probing questions. For more information on this process, see Section 3.2.5 on page 81.

While the participant was taking the survey, the researcher watched his or her progress from another computer. She specifically tracked whether the participant received model-initiated help, requested help, or both. This was important because participants received different satisfaction questions depending on whether they received help and which type of help they received. Specifically, they answered a filter question asking whether they received no help, requested help, received model-initiated help, or received both model-initiated and requested help. If participants answered this question incorrectly, they received the wrong set of satisfaction questions. The researcher noted these cases and asked the correct set of questions during the debriefing. Additionally, during the debriefing, the researcher asked participants why they did not request help on questions they thought were challenging and what they thought of the help they did receive. Appendix Q provides the instructions the participants were given prior to starting the survey along with the debriefing questions.

4.3 Analysis

The primary purpose of providing help to respondents is so they answer questions more accurately. However, it is also important that the respondents are satisfied with the help they receive. Therefore, this section outlines the analyses we used to identify a mode of providing help that leads to high accuracy and satisfied participants.

4.3.1 Descriptive Statistics Analysis

For the help to have an impact on accuracy, we first needed to make sure participants actually received it. The first measure we looked at was the percent of participants that received help at least one time. Given participants could either request help or receive it through model-initiation, we also calculated the percent of participants that received each type of help at least one time, both overall and within each help mode (text, audio, chat).

How frequently a participant received help likely altered his or her accuracy across the instrument and overall perception of the help. Therefore, we also calculated the mean number of times, within each mode, that participants received each type of help (model-initiated and requested) in two different ways. First, we compared the mean number of questions on which participants received help across all participants. We examined the main effects of type of help received and the mode of help administration, as well as the interaction between the two main effects.

We also limited the analysis to only those participants that received each type of help. Among those participants, we compared the mean number of questions on which they received help using a one-way, between subjects, analysis of variance (ANOVA) to test the hypothesis that $\mu_{text} = \mu_{audio} = \mu_{chat}$. We did not expect there to be any differences in amount of model-initiated help between modes because, on average, participants should have taken the same amount of time to answer questions. However, we did expect differences between how many times help was requested across the different modes, and we believe these differences could be related to satisfaction.

4.3.2 Accuracy Analysis

The use of scenarios allowed us to know whether participants answered each question correctly. We first calculated overall accuracy throughout the instrument. Due to the experimental manipulation in this study, we expected participants to be more accurate on questions associated with straightforward scenarios. Additionally, because the questions associated with straightforward scenarios should be easier for participants, we expected them to receive more help on questions associated with complex scenarios. We briefly examined accuracy for the questions associated with straightforward scenarios, but focused this analysis on questions associated with complex scenarios and compared the percent of accurate responses when help was provided to the percent of accurate responses when help was not provided. Additionally, we believed participants would focus more on help they requested than help they received unsolicited. Therefore, we also compared accuracy between types of help. These analyses should help determine whether participants paid attention to the help and used it to answer the questions. We can also see whether they paid equal attention to the help when it was provided compared to requested.

Once we knew whether the help was useful to participants, we focused on which mode of help resulted in more accurate responses. Looking at each help mode separately, we compared the percent of accurate responses when help was received (either through the model or when requested) to when no help was received. This analysis will inform whether participants responded to the help similarly across help modes. Additionally, we examined whether the interaction of mode and whether help was received had an effect on accuracy.

The means described in this section were calculated across all observations, but the dependence within participant (as a result of each participant answering 20 questions) was accounted for using clustering[22]. In SAS®, PROC SURVEYMEANS and PROC SURVEYLOGISTIC were used to calculate means and test statistics while accounting for clustering. To test the main effects of receiving help and mode of help administration, as well as the interaction between the two variables, we used the following models:

Accuracy = β_0 + Mode

Accuracy = β_0 + Receive_help

Accuracy = β_0 + Mode + Receive_help + Mode*Receive_help

[22] We also looked into conducting this analysis using repeated measures and by calculating the mean accuracy for each participant and then taking the overall mean. We opted not to use a repeated measures analysis due to high missing data rates. We decided not to use participant means because we did not think the standard errors of the overall means properly reflected how frequently each participant received help or how reliable his/her accuracy was. Each of these analyses, however, generated the same results as the method used in this study.

where accuracy was a binary variable equal to '1' if the answer was correct and '0' otherwise, Mode was equal to 'text,' 'audio,' or 'chat,' and Receive_help was equal to '1' if the participant received help on a question and otherwise '0'. Taken together, these measures will help determine which help mode results in the highest accuracy.

While the procedures used in this analysis take the dependence into account in regard to the standard errors of the estimates, the means do not reflect that some participants are more or less accurate than others. To ensure that no single participant was driving the results, we systematically removed each participant from the dataset and reran the analyses.

4.3.3 Satisfaction Analysis

In a production environment, even if respondents provide more accurate responses after receiving help, if they consider it unwanted or an imposition, they may grow frustrated and abandon the survey. Therefore, it is also important to assess their satisfaction with the help they receive. Participants answered a series of questions designed to evaluate their satisfaction with the survey and the help they received (Appendix P). The questions asked participants about their overall experience with the survey, whether the help they received was helpful, whether it made the questions easier to answer, whether they read or listened to all of the information in the help, and how they reacted to receiving the model-initiated help. We probed reactions to the model-initiated help more than requested help because we expected this type of help would be more frustrating and annoying to participants.

First, we looked at each satisfaction question separately. For each question, we calculated the average score within each help mode and then compared the average scores across modes using a one-way ANOVA. This will inform whether participants found the help more useful, whether they paid more attention to it, or whether the model-initiated help was less frustrating in a specific mode compared to the other modes.

Additionally, there were two questions that were asked separately for participants that received model-initiated help and requested help: whether the help made questions easier to answer and how frequently they read/listened to all of the information. As we suspected that participants that requested help would be more accurate than participants that received model-initiated help, this analysis will help inform whether that is because they paid closer attention to the content of the help.

While each mode of help administration may have its benefits and drawbacks according to participants, what we were most interested in is which mode participants preferred overall. Therefore, we created two composite scores for each participant that summed all of the ratings for all of the satisfaction questions the participant answered. The first score only included satisfaction questions associated with the model-initiated help, whereas the second score included the satisfaction questions associated with requesting help as well. After the ratings were summed, we divided them by the number of satisfaction questions the participant answered. We then calculated a mean score for each help administration mode and compared the mean composite scores across mode using a one-way ANOVA. The first score tells us which mode participants preferred for model-initiated help, whereas the second score tells us which mode they preferred overall.

In addition to asking participants to rate their satisfaction, we also gave them the opportunity to provide us with their impressions of the model-initiated help through a write-in box. The comments may alert us to what specific aspects of the help participants find frustrating which can lead to future improvements that may increase satisfaction.

Finally, we asked participants a few questions regarding their feelings about help in general. Specifically, we asked if they prefer the help information to be available automatically (always on), if they prefer the system to detect when they need help, or if they prefer to request help. Additionally, we asked in which mode they prefer to receive help. We looked for the mode in which the highest proportion of participants reported a preference for model-initiated help and also the mode the highest percent reported preferring. To account for people’s tendency to like what they know, all participants that answered that they preferred the mode they were assigned to were asked what their second preference was. Analyzing both reports together should give a good indication if one mode stands out.

4.4 Results

We first examine a set of descriptive statistics, including how many participants received help, which type of help (requested or model-initiated), and how frequently. We then discuss overall participant accuracy, accuracy by mode, and by the type of help they received. Finally, we determine which mode of help administration participants preferred overall.

4.4.1 Descriptive Statistics Analysis Results

Overall, 90.7% of participants received help on at least one question during the survey. Roughly the same number of participants received help in each help mode (between 44 and 47 out of 50 participants). However, the type of help they received as well as how frequently varied across mode. As expected, more participants received model-initiated help than requested it; 88.7 percent received model-initiated help, whereas only 34.7 percent ever made a specific request. It is interesting that the requests for help were not consistent across mode. Specifically, 52.0 percent of participants in the text mode requested help for at least one question, whereas 32.0 percent requested help in the audio mode and only 20.0 percent in the chat mode, $\chi^2(2) = 11.54$, $p = 0.003$.

In addition to participants receiving different types of help across modes, they also received help on a different number of questions. Table 4.2 provides the mean number of questions on which participants received model-initiated help and requested help by administration mode. Additionally, it provides the overall mean number of questions on which participants received help by type of help received and help administration mode.

Table 4.2 Mean Number of Instances Help was Provided by Mode and Type of Help Across all Participants

Type of Help	**Text**	**Audio**	**Chat**	**Mean number of questions by type**
Model - Initiated	6.15	6.00	4.82[23]	5.65
Requested	2.10	0.91	0.42	1.15
Mean number of questions by mode	4.12	3.46	2.62	

Participants did not receive help on an equivalent number of questions across mode or type of help. On average, each participant requested help on significantly fewer questions than the model provided it $F(1) = 124.8$; $p < .0001$. Additionally, there was a significant main effect for the number of questions on which participants received help across administration mode $F(2) = 4.73$; $p = 0.10$ and this is largely a result of more participants requesting help in the text mode. However, we did not see a significant

[23] We expected the number of times a participant received model-initiated help to be consistent across modes because it was automated. However, because the researcher had to initiate the chat, she occasionally missed instances where help should have been provided. This happened on approximately 0.2 percent of questions

interaction between type of help received and the mode of administration *F*(2) = 0.57; *p* = *n.s.* It is also interesting to note that participants' help requests were directly related to how burdensome the help was to receive and process. For example, the text mode was the least burdensome for participants in that they could read it any time they wanted and did not have to put in any additional effort to obtain the information. Participants requested help on fewer questions in the audio mode, which required them to listen at the time the help initiated or pause and restart the recording. Finally, the chat mode was the most burdensome because the participant was responsible for generating a question and, likely as a result of this, they requested help on the fewest number of questions in this mode.

Table 4.2 showed that participants did not receive comparable amounts of help across mode or type of help. However, that analysis used all participants across all modes. It is also valuable to limit the universe to only participants that received each type of help to see if they used the help differently. Therefore, Table 4.3 provides the mean number of questions on which participants received each type of help, among participants who received each type of help. For example, the 'Requested' row only includes participants that requested help at least one time. Given these participants, we tested whether they received help on the same number of questions across modes.

Table 4.3 Mean Number of Instances Help was Provided by Mode and Type of Help Among Participants that Received each Type of Help

	Text		**Audio**		**Chat**		
Type of Help	**N**	**Mean (se)**	**N**	**Mean (se)**	**N**	**Mean (se)**	**F (df;** *p*-value)
Model - Initiated	44	6.70 (0.67)	43	6.56 (0.70)	47	5.13 (0.57)	1.84 (2, 130; 0.163)
Requested	26	4.15 (0.49)	16	2.69 (0.47)	10	2.10 (0.59)	3.99 (2, 49; <0.025)

We see that participants who received model-initiated help received it on about the same number of questions across mode. This is an expected result because mode did not have an impact on the average number of times participants received model-initiated help. On the other hand, we see that there were significant differences in the amount of help that participants requested between the different modes. Specifically, participants in the text mode requested help 1.5 times more often than in the audio mode almost twice as frequently than the chat mode. This difference may be related to how useful they found the help, which we will discuss in Section 4.4.3, or how burdensome it was. Because participants in the text mode requested help on more questions than participants in the other two modes, they also received more help overall. If receiving help is related to increased accuracy, which we discuss in Section 4.4.2, then we can expect participants in the text mode to be more accurate than the other participants. Additionally, if participants request help rather than having the model provide it, overall satisfaction may be higher.

4.4.2 Accuracy Analysis Results

Across all participants, regardless of mode and whether they received help, 72.2 percent of responses were accurate. Based on prior research, we expected accuracy to be higher when help was provided as compared to when it was not. Surprisingly, at first we found overall accuracy (across all modes) was significantly higher when no help was received (74.6 percent) as compared to when help was received (69.1 percent), $t(149) = 1.45$, $p = 0.019$.

However, 90.5 percent (210 out of 232) of instances where help was model-initiated in the chat mode resulted in the participant either immediately closing the window or declining to receive any help. Additionally, in the text and audio mode, approximately 26.7 percent did not actually receive any help. In the audio mode this consisted of pausing the recording before they actually received the information. On the other hand, in the text condition we were able to determine using the eye tracking cases where the participant never looked at the information in the help box. These cases were only accurate 55.4 percent of the time.

We recoded cases that did not actually "receive" any help and recalculated the accuracy. This resulted in more expected results, where participants correctly answered 74.5 percent of questions on which they received help compared to 71.6 percent of questions where they were either not offered help or declined it, $t(149) = 0.1503$, *n.s.* While this difference was not significant, it did trend in the expected direction. For the rest of the analyses discussed in this section, if a participant did not actually read or listen to the help provided, the case was not included in the received help group.

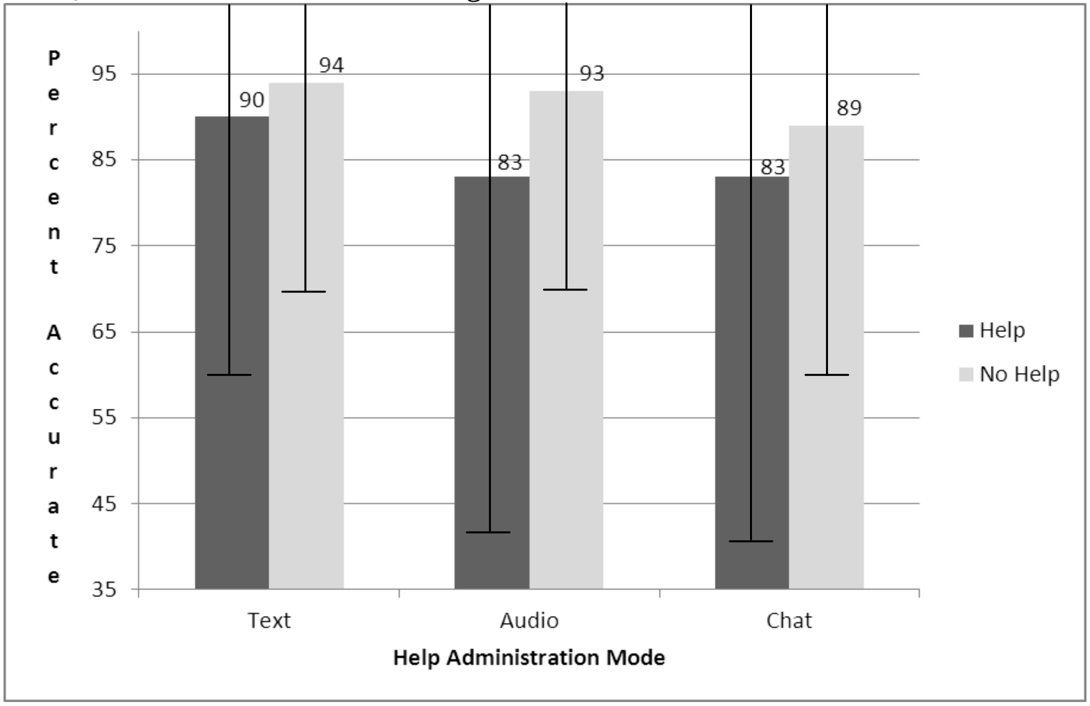
We also expected that participants would be more accurate if they requested help than if it was model-initiated because they would likely be more focused on it. As expected, we found that 81.1 percent of responses were accurate following a help request compared to 72.5 percent following model-initiated help, $\chi^2(1) = 6.44$, $p = 0.011$.

Although overall accuracy is important, we purposefully manipulated the difficulty of the questions because survey respondents generally answer straightforward questions accurately. In this study, participants answered 90.8 percent of questions associated with straightforward scenarios correctly compared to only 52.9 percent of questions associated with complex scenarios, $\chi^2(1) = 450.7$, $p < 0.001$. Additionally, participants received help on more questions associated with complex scenarios (69.3

percent) compared to questions associated with straightforward scenarios (30.0 percent), $\chi^2(1) = 131.11$, $p < 0.001$.

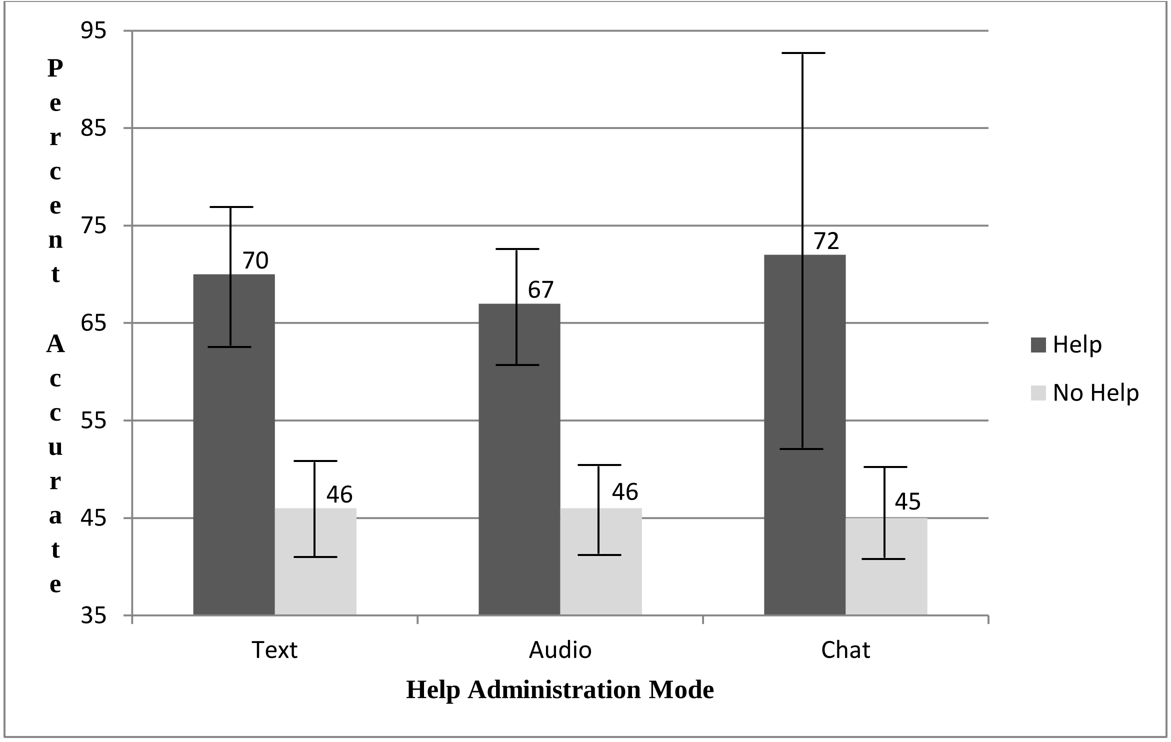
While 90.8 percent of participants answered questions associated with a straightforward scenario accurately, we wanted to see whether this accuracy varied by whether or not participants received help. Figure 4.4 displays the percent of correct answers on questions associated with straightforward scenarios, by mode and whether help was received. The upper bound of the error bars (95 percent confidence interval for the mean) is cut off from display because the scale is restricted to 100 percent. For this analysis, participants received help on 216 questions and did not receive help on 1,421 questions. The figure shows that there was no difference in accuracy when help was or was not provided between help administration mode nor overall. This is an expected result because the questions were intended to be easy enough for participants to answer correctly without receiving assistance.

Figure 4. Percent of Correct Answers, by Mode, when Help was and was not Received for Questions Associated with Straightforward Scenarios



As there are no differences in accuracy for questions associated with straightforward scenarios, we focused our analysis on the effect of help on questions associated with complex scenarios (Figure 4.5). For this analysis, participants received help on 487 questions and did not receive help on 1,160 questions.

Figure 4.5 Percent of Correct Answers, by Mode, when Help was and was not Received for Questions Associated with Complex Scenarios



For questions associated with complex scenarios, we see that participants did answer significantly more accurately across all modes when help was provided compared to when it was not[24]. However, the mode in which help was provided did not have an impact on accuracy. Additionally, there was not a significant interaction between the mode of help and whether help was received $\chi^2(2) = 0.755$, $p = 0.686$.

Despite our concern that a handful of participants could be driving these results, the diagnostic test did no show this to be an issue. Overall, the means did not change

[24] As a reminder, whether a participant received help was not an experimental manipulation in this study but was either model initiated or initiated by the participant, and therefore a function of each participant.

much as a result of removing any one participant from the analysis. Within treatment, removing participants did have a larger effect on the means (especially in the chat mode because so few participants received help), but these changes did not change the interpretation of any of the test statistics. Appendix R plots the difference between the overall means using all the data and after removing each participant.

4.4.3 Satisfaction Analysis Results

This chapter aimed to identify a mode of providing help that would maximize accuracy and satisfaction. As all three modes yielded similar levels of accuracy, we focused on which mode participants preferred. When asked about their overall experience with the survey, participants in all modes rated their experience as ‘Good’; providing ratings between 1.9 and 2.1 F(2,147)=1.17, *n.s* (where ‘1’ was excellent and ‘4’ was poor). However, when we asked specifically about the help they received, we did find some differences.

First, when participants who received model-initiated help were asked whether they found the help to be satisfying or frustrating (where ‘1’ was very frustrating and ‘5’ was very satisfying), they expressed the most frustration with the chat mode. Specifically, participants rated the chat condition 3.5 compared to 3.9 and 4.0 for the audio and text conditions, $F(2,128) = 3.42$, $p = 0.036$. However, as we previously mentioned, the majority of the help offerings via chat did not result in the participant actually receiving help. In the other modes, even if the participant did not want help, they generally received some information which may have helped them realize the information provided was useful, which in turn may have increased their satisfaction. To support this hypothesis, when we limited the population to participants in the chat mode that actually received help, the satisfaction score was 4.0, which is equivalent to the other two modes[25].

When asked whether the model-initiated help was helpful (where ‘1’ was very unhelpful and ‘5’ was very helpful), participants provided ratings between 4.4 for text to 3.7 for chat, with the audio rating falling in the middle. These ratings suggest that participants usually found the help helpful. They rated the help provided in the chat mode to be less helpful than in the text mode, $t(53) = 2.38$, $p = 0.021$, but there were no significant differences between the ratings for the audio mode compared to the other two modes.

We also asked participants whether the help made the questions easier to answer, both for model-initiated and requested help. Participants provided ratings between 4.4 and 4.2 for requested help ($F(2,42) = 0.14$, *n.s.*) and between 4.3 and 4.1 for model-initiated help ($F(2,87) = 0.58$, *n.s.*).

On the other hand, there were significant differences in how much of the model-initiated help participants received $F(2,87) = 4.56$, $p = 0.013$. For questions on which participants received model-initiated help, they were significantly less likely to listen to an entire audio recording than they were to read the full text or chat information (compared to text, $t(82) = 2.73$, $p = 0.008$; compared to chat $t(58) = 2.15$, 0.035). Specifically, participants in the audio mode indicated they sometimes listened to the full recording and sometimes did not, whereas participants in the other two modes indicated they usually read the complete help. We believe this is because once participants in the

[25] We also recalculated the satisfaction scores for the text and audio modes after limiting the universe to participants that actually received help. However, there were only two participants in the text mode and one participant in the audio mode that ignored the help for every question. Therefore, the original scores and recalculated scores were the same.

audio mode heard the information they needed to answer the question, they stopped listening. On the other hand, in the text mode it is easy and fast to skim the remaining information to see if there is anything relevant and in the chat mode, the help should only provide the information the participant asked for, so it is logical they would read the entire message.

Not surprisingly, participants read/listened to more of the help when they requested it than when it was model-initiated. Model-initiated ratings ranged from 2.9 to 3.9 whereas requested ratings ranged from 3.7 to 4.6 across modes. As with the model-initiated help, participants who requested help in the audio mode appear to listen to less of the help than the other conditions, but the difference between the audio and text modes did not meet statistical significance $F(2,42) = 1.60$, $p = n.s.$ These findings also help explain why accuracy was higher for questions on which respondents requested help as compared to when help was model-initiated.

While these individual analyses have helped identify some benefits and drawbacks of the different modes of providing help, a composite score analysis provides us with a final determination of which mode participants preferred overall. Tables 4.4 and 4.5 provide the composite scores by mode[26]. The first table provides a comparison of composite scores for the satisfaction questions associated with the model-initiated help whereas the second table includes the satisfaction questions associated with requested help as well. The first column shows the comparison being made. The score in the 'Estimate 1' column refers to the help mode listed first in the 'Comparison' column.

[26] We also compared the composite satisfaction scores by age group. Specifically, we first divided participants into those older than 45 and those 45 and younger and then those older than 29 and younger than 20. In both cases, the older age group was slightly less satisfied with the help as compared to the younger group, but the differences did not meet statistical significance.

Similarly, the score in the ‘Estimate 2’ column is the second help mode listed in the ‘Comparison’ column.

Table 4.4 Composite Satisfaction Score Comparison for Model-Initiated Help

Comparison	**Estimate 1 (se)**	**Estimate 2 (se)**	**Difference (se)**	**t**	**P**
Text vs Audio	4.12 (0.08)	3.84 (0.11)	0.28 (0.13)	2.12	0.037*
Audio vs Chat	3.84 (0.11)	3.70 (0.10)	0.14 (0.15)	0.92	0.361
Text vs Chat	4.12 (0.08)	3.70 (0.10)	0.42 (0.13)	3.17	0.002*

Table 4.5 Composite Satisfaction Score Comparison for all Help

Comparison	**Estimate 1 (se)**	**Estimate 2 (se)**	**Difference (se)**	**t**	**P**
Text vs Audio	4.26 (0.08)	3.91 (0.11)	0.35 (0.14)	2.54	0.0129*
Audio vs Chat	3.91 (0.11)	3.77 (0.11)	0.14 (0.16)	0.87	0.3848
Text vs Chat	4.26 (0.08)	3.77 (0.11)	0.49 (0.14)	3.52	0.0007*

Both tables show that scores for the text mode are significantly higher than both audio and chat. This suggests that, overall, participants are more satisfied with receiving help through a standard pop up text box. This finding is not what we expected based on the human-computer interaction and education literature. However, participant’s written and verbal comments regarding the help they received provide insight into their preferences.

The most common reason participants provided for being unsatisfied with the model-initiated help was that it was distracting. The help often initiated while participants were trying to read the question or response options and it made them lose their train of thought. While this happened in every mode, it seemed to be the most distracting for participants that received audio help. In the other two modes, the participant could ignore the help until they finished reading and processing the question. However, it is much more difficult to ignore a voice. Several participants indicated they would have liked the help more had it been better timed. They treated the survey as a test and they wanted to find the correct answer on their own. When the help jumped in before they had a chance to pick an answer, they were annoyed.

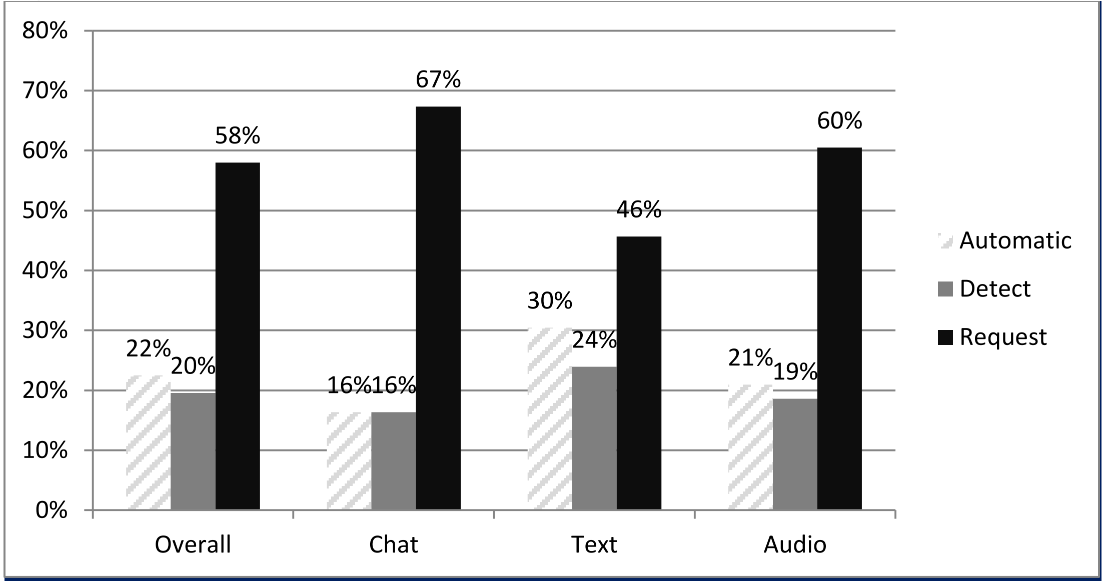
We also expected participants in the chat mode to enjoy the personal interaction and also prefer being able to ask specific questions instead of sifting through information they did not need. While participants that received one of the other two modes of help likely saw or heard some of the help information just because it was on the screen, very few participants in the chat condition received any help at all because they explicitly declined it before anything was presented. Therefore, they were not able to see the value of the help as the others were able to. Additionally, some participants reported difficulty formulating a question to ask the agent. They did not know exactly what to ask, so they did not ask anything. Finally, participants that received help via chat had to explicitly state that they needed help, which some said hurt their pride. On the other hand, in the other two conditions, the help was simply presented; they did not have to admit that they needed it or used it. For all these reasons, we believe that the negatives of the chat mode outweighed the positives for most participants.

In addition to asking participants how satisfied they were with different aspects of the survey, we also asked about their preferences when taking surveys in general. We found that the majority of participants (58.0 percent) would prefer to request help when needed as opposed to having the help content always available (22.4 percent) or having the system detect when they need help (19.6 percent)[27]. Figure 4.6 provides participant's reported preferences in how they want to receive help while taking surveys.

[27] Although having the help information automatically available was not a condition in this study, we included it in the options because it has been tested in prior research (Conrad *et al*, 2006).

We can see that in all cases that participants prefer to request help. However, the distribution is not as pronounced for the text mode as it is for the other two. We believe this is because participants in the text mode reported being more satisfied with the model-initiation experience and they found the help to be more helpful (although the difference between the audio and text modes did not meet statistical significance). Therefore, this likely made them see the value of having the help provided to them more than participants in the other modes realized.

Figure 4.6 Distribution of Help Initiation Preferences



We also asked the participants which mode of help they would prefer to receive in future surveys. The majority of participants who received help via text or chat reported preferring the mode that they received. Specifically, 76.1 percent of participants that received text help indicated they preferred text and 55.1 percent of the participants that received chat help indicated they preferred chat. On the other hand, only 44.5 percent of participants that received audio help indicated they preferred audio. This is likely because they believe the other two modes would be less distracting.

We also asked participants what their second preference was (for participants that selected the same mode as their treatment as their first preference). About 75 percent in both the audio and chat modes selected text. Of the participants in the text mode, 53 percent selected chat as their second choice and 41 percent selected audio. Therefore, we see participants who received help in the text mode do not have a strong preference between audio and chat, but participants in those modes strongly prefer text. It could be that text is the most familiar way of obtaining information on the Internet, or it is also possible that participants see fewer barriers in the text mode compared to the other two, such as needing speakers for the audio mode and needing to formulate questions for the chat mode.

4.5 Conclusions

In this study, we were most interested in finding a mode to provide model-initiated help that would maximize both accuracy and satisfaction. Past research has shown that providing model-initiated help increases accuracy, but can lead to frustration. Therefore, this study experimented with providing model-initiated help in different modes to see whether participants preferred one mode to another. Additionally, we felt strategies from the education literature could further increase accuracy.

We expected participants to be more accurate when they received audio or chat help. The audio help allowed participants to reread the question and response option while processing the help information, which we saw some did after reviewing the eye tracking data. On the other hand, participants that received help via chat had the ability to ask the agent exactly what they wanted to know instead of having to sift through irrelevant information and decide what applied. Contrary to our hypotheses, however, there was no difference in accuracy across all three help administration modes.

In the case of the audio mode, we believe some participants did not listen to the entire recording, either because it took too long, they thought they heard information that applied, or because it was too distracting. It is also possible that this task was not complex enough to require dual processing; it was just as effective for participants to read the help text and then reread the response options as it was to listen to the help and scan the response options at the same time.

For the chat administration, in order to receive a helpful response, participants needed to ask the right question. However, several participants reported having difficulty formulating a question to ask the agent. Additionally, participants may not have known exactly what they were confused about so having the opportunity to see or hear all of the help information may have clarified an issue they did not know existed. Therefore, relying on participants to ask the right questions is likely putting too much burden on them to have a positive impact on accuracy and can lead to unnecessary frustration.

We also expected participants to be more satisfied with the chat or audio help administration modes because they were more human-like interactions with the computer. However, participants preferred receiving help through a standard text box, possibly because it was less burdensome than the other modes. While the chat mode had the benefit of providing participants with a conversation and the exact information they needed to know, the drawbacks seem to outweigh the benefits[28]. Specifically, the few

[28] As we were not able to integrate the chat function into the survey program, participants needed to activate the chat window instead of it popping up in the middle of the screen automatically, as in the other modes. This may have made it easier to ignore or made it more frustrating because there was an extra step. No participants explicitly noted this as an issue, however.

people who actually received help via chat seemed to like it. However, it requires formulating a question and admitting needing help, both of which may be too difficult for participants to make this mode feasible[29]. Therefore, providing help via audio or text catches people who are unwilling to admit to needing help, but will pay attention to it if it is there.

There was an additional benefit to providing help via a text box that did not come out in the satisfaction questions. A higher percent of participants in the text mode that received model-initiated help ended up requesting help later in the survey. If participants did not like the imposition of model-initiated help, the fact that they saw the value of the help and proceeded to request it means their frustration was reduced because they did not receive information they did not want while their accuracy was increased because they continued receiving help.

In debriefings, several participants noted a major limitation to providing audio help: many people do not have speakers or they keep their sound off most of the time. Additionally, people completing the survey in libraries likely would not have sound activated. Therefore, this mode may not be feasible in a production environment.

Based on the research presented in this chapter, we believe providing help via a text box optimizes respondent accuracy and satisfaction. However, we do feel that there are some benefits to the chat mode that we think should be further explored. Because participants struggled with formulating a question, it would be interesting to see the impact of adding a chat feature to the text mode. For example, a respondent would receive help in the standard pop up text box and then there would be a link to chat at the

[29] We had also hypothesized that participants would be more polite in the chat mode because of how personal the interaction was. It was interesting to see that most respondents who declined help were very polite in doing so, saying “no thank you” or “no thanks” in most cases.

bottom of the box. If the help content did not fully clarify the respondent's question, he or she would have the opportunity to follow-up with an agent via chat. In this case, no one would be missing out on help that they need, but additional information would be available for those that need and want it that would be clear and to the point. However, this does bring up this issue of providing unstandardized help.

Chapter 5. Summary, Implications for Practice, and Future Research

5.1 Summary

It is well known that five people can interpret the same question in five different ways. To account for this, survey designers put each question through a rigorous series of tests and revisions until they are satisfied that respondents will understand the questions as intended. Unfortunately, even extensively tested questions can confuse respondents with atypical situations. This confusion can lead to reduced accuracy and measurement error in some cases. Researchers have shown that providing help to these respondents can increase accuracy, but respondents do not generally request help and often do not even realize they need it (Conrad *et al*., 2006; Schober *et al.,* 2003).

In interviewer-administered surveys, researchers have identified a set of indicators they can train interviewers to look for to identify respondents who are having difficulty answering a question. Specifically, the research suggests that these respondents often have disfluent speech patterns, such as using fillers, hedges, and pauses while answering survey questions (Smith and Clark, 1993; Brennan and Williams, 1995; Schober and Bloom, 2004). In cases where interviewers have been trained to notice these cues, they can re-read the question and definitions, probe respondents, or provide them with help to ensure they understand the question and how it applies to their situation. However, many surveys are moving away from interviewer-administration towards multi-modal or Web administration to try to cut costs. This raises the question of how to detect respondent difficulty in self-administered surveys.

In Web surveys, researchers have experimented with using long response latencies as an indicator that respondents are experiencing difficulty answering a question

(Conrad *et al.,* 2007; Ehlen *et al.,* 2005; Heerwegh 2003). While using this indicator to provide help has successfully increased accuracy (Conrad *et al.,* 2007; Ehlen *et al.,* 2005), there are several limitations. First, outside of a lab setting, long response times might not be related to difficulty but to respondent engagement. For example, respondents may be taking a long time to answer a question because they are not engaged in the survey task; they may be checking email, shopping in a different window, or not at their computer. Additionally, long response times cannot tell us what about the question is confusing to the respondent, so the best way to help them may not be clear. Baseline speeds also differ by respondent and this baseline is not known at the beginning of a survey (Fazio, 1990).

The first study in this dissertation (Chapter 2) aimed to determine whether there are indicators, similar to speech disfluencies, which suggest difficulty answering a question on a Web survey. We were interested in what general movements participants engaged in with their mouse while completing the survey and also which of those movements might be related to difficulty. Using Web browsing literature, education literature, and pretest observations, to identify 11 movements in which we thought survey respondents would engage. These movements included hovering over the question text or 'Next' button with the mouse, holding the mouse over a response option like a marker, following along with the mouse to read horizontally or vertically, and a set of regressive movements: moving the mouse from the response options to question text and back, from the response options to the 'Next' button and back, and back and forth between two response options.

The study used 30 participants and 20 questions from the ACS to conduct an exploratory analysis determine which of these movements participants used. Although we used a small sample size and did not conduct any significance testing, we found that for all but two of the movements we identified, at least one participant engaged in each movement. Additionally, while some participants moved the mouse more than others, almost all engaged in at least some of the movements multiple times throughout the survey.

In addition to knowing which movements participants engaged in, we also wanted to know which were related to difficulty. We could not conduct statistical testing due to our small sample size, so we used descriptive statistics to determine which movements might be related to difficulty. First, we asked the participants to rate the difficulty of every question they answered on a five-point scale. By examining the relationship between which movements occurred on questions that participants rated 'difficult,' which movements occurred on questions that participants explicitly said were difficult, and analyzing their focus of attention while engaging in the different movements, we reached the following conclusions: instances of regressive movements were likely related to difficulty, we were unsure whether instances of hovering were related to difficulty, and the rest of the movements were not related to difficulty.

In the second study of this dissertation (Chapter 3), we aimed to determine, through statistical testing and modeling, which movements were related to difficulty. The second study used a larger sample size of 100 participants. The focus of this study was on difficulty, so to ensure that participants experienced difficulty with at least some of the questions we asked them to answer the 20 ACS questions based on pre-written scenarios, which manipulated the level of difficulty, instead of answering based on their own experience. We found that, in general, participants in the second study behaved similarly to the participants in the first study. Specifically, every participant engaged in at least one of the movements while completing the survey and, overall, participants engaged in each movement multiple times. The replication across studies suggests that these movements are common in survey taking and what we can expect to see from a more general population.

To determine which of these movements were related to difficulty, we first conducted an exploratory analysis using the participant's difficulty ratings and the version of the scenario to find relationships between difficulty and specific movements. From this analysis we believed that horizontally reading the question text and vertical reading were not related to difficulty, using the mouse as a marker, regressive movements, and hovering over the 'Next' button were related to difficulty, and we were not sure hovering over the question text, or horizontally reading the response options were related to difficultly.

The analyses up to this point had not considered the relationships between the different movements or the nesting of movements within participant. Therefore, we used hierarchical logistical modeling to generate a model capable of predicting when a participant was having difficulty answering a survey question. We found that hovering over the question text, moving the mouse back and forth between response options, moving from the response options to the 'Next' button and back, and using the mouse as a marker two or more times on one question were the movements most predictive of difficulty.

In addition to identifying which movements were related to difficulty, we also tried to link specific movements to specific types of difficulty (imperfect fit, misalignment, and technical terms). We used logistic regression models to predict whether each type of movement was significantly related to the three types of difficulty, but we did not find any relationships. It is possible that the differences between these types of difficulty were too subtle to detect using mouse movements, but they may be an effective way to detect and differentiate between other types of difficulty.

The third study in this dissertation (Chapter 4) focused on providing respondents with help. In interviewer-administered surveys, once the interviewer recognizes the cues that a respondent is having trouble answering a question, they can intervene and provide assistance. In Web surveys, researchers have found that providing help increases accuracy, but can have a negative impact on respondent satisfaction with the survey (Conrad *et al.*, 2007).

In current research, help is provided via a standard text box. However, there may be other ways to provide help that still increase accuracy but do not decrease satisfaction. Specifically, providing help via an audio recording might increase accuracy because the psychology and education literature suggests people can more efficiently process inputs from multiple sensory modes than a single mode (Baddeley and Hitch, 1974; Penny, 1989). Therefore, we thought participants would be able to listen to the audio recording while reviewing the question and response options instead of having to focus on one or the other.

Another way to provide help is via an instant message, or chat. Some Web sites provide users the option to chat with an agent to get their questions answered. This feature allows respondents to ask the agent exactly what they need to know so they do not have to sift through superfluous information to find what they need. Therefore, we also thought providing help via a chat feature could have a positive effect on accuracy.

In addition to potentially increasing accuracy, both audio recordings (with a human voice) and chatting are human-like interactions with a computer and past study participants have indicated they prefer personal interactions with computers to more computer-like (automated) interactions (Johnston 2008; Lester *et al.*, 1997). From this research, we hypothesized that providing help via audio or chat would also increase participant satisfaction.

This study used 150 participants to assess accuracy and satisfaction with three different modes of help administration (text, audio, and chat). To measure accuracy and to vary the complexity of the questions, we asked participants to answer based on scenarios. To obtain help, participants could request help via a link or, if they exceed the median response time calculated from the second study, they would receive help automatically (model-initiated).

We used random-effect logistic regression models to determine whether there were differences in the percent of correct responses for participants that did and did not received help in each mode. Additionally, each participant answered a series of satisfaction questions after the survey and we used one-way ANOVA models to determine whether satisfaction varied by help administration mode.

We found participants were more accurate when they received help, across all modes. However, contrary to our hypotheses, they were no more accurate receiving help via audio or chat than text. Additionally, participants reported preferring to receive help via the standard text box compared to the other two modes. While the help triggered before the participant was ready to receive it in all modes, it seemed to be much more distracting and difficult to ignore in the audio mode. In the chat mode, participants reported having difficulty formulating a question to ask the agent and they also did not like admitting that they needed help. Therefore, the results from this study recommend providing help in a text box to maximize satisfaction and accuracy.

5.2 Limitations

Laboratory studies present several limitations. In the studies discussed in this dissertation, we were not able to obtain a representative sample due to the relatively small number of participants and recruiting challenges. While our sample was not representative, we did aim to recruit participants with varying age, gender, race, and education to try to minimize the effect of using a convenience sample. Despite this effort, generalizing these results to any greater population requires caution.

Another limitation was that students manually coded the mouse movements in the first two studies instead of using a computer program. From the reliability results, we know this observational coding resulted in both errors of omission and commission. Unfortunately, at the time these studies were conducted, we did not have access to technology that could track the mouse cursor coordinates and concurrently code each of the movements of interest.

All of the questions used in this dissertation came from the ACS. This survey only asks respondents factual questions about their household. It is possible that this limited the inferences we were able to draw from the mouse movements. Specifically, while we were unable to relate specific types of difficulty to specific movements, it could be because the ACS questions were limited in the type of difficulty respondents experience. Comparing movements made on the ACS to movements made on opinion or expenditure surveys, for example, may be able to uncover additional relationships.

Finally, all of the participants in this dissertation used a standard desktop computer with a keyboard and mouse. We used this set up because we believe the majority of households still use this technology. However, there are many individuals who use a laptop or mobile device instead of a desktop and this number is likely increasing. Some of these people may attach a mouse while others may use the track pad. Based on the research conducted in this dissertation, we have no way of knowing whether the results from these studies will translate to laptop users.

5.3 Future research

To address these limitations, we suggest some areas for future research. First, although we used convenience samples for all of our studies, we have no reason to believe differences in the sample would impact the results of the first two studies. However, conducting the survey in a real-world environment instead of a laboratory will lead to more normal behaviors, such as survey abandonment, which is important for the third study on providing help. While some participants reported being frustrated by model-initiated help, if this frustration does not lead to breakoffs, it may be worth providing anyway. Therefore, studying breakoffs as related to model-initiated help would be beneficial in determining whether this is a technique that can be used moving forward. Additionally, surveys of different lengths may be affected differently, so it is likely worth investigating this phenomenon across multiple surveys.

We also think it is important to confirm the results of the second study in this dissertation using programmed mouse movements instead of hand-coded movements. There should be areas of interest defined for each movement and a timer so there is no uncertainty or vagueness regarding whether a movement occurred or not. We do not expect the results to be drastically altered, but because there is inherent error in human coding it is worth verifying the results of this dissertation and also developing the programs to be able to track these movements across many surveys in real time.

One of the expected benefits of predicting difficulty using mouse movements was to be able to link specific movements to specific types of difficulty. However, we were not able to relate any movements to the types of difficulties we studied. We believe this was because there was not enough differentiation between the types of difficulty. If we were able to determine what type of difficulty respondents were experiencing, we would be able to provide them with tailored help, which could increase accuracy and reduce the burden of being exposed to information they do not need. Future researchers could experiment with different types of difficulty to determine whether relationships exist between types of difficulty and the various mouse movements.

Finally, it is also important to replicate the first two studies in this dissertation using a laptop and recruiting participants who predominantly use laptops. This will let us know whether our results can be extended to all computer users or if the findings are more limited. If desktop and laptop users do not behave similarly, there may be a different set of movements related to laptop use or this type of diagnostic tool may not be effective for these respondents.

5.4 Implications for Practice

The results from this dissertation can be used to improve current practices. Currently, response latencies are used to provide respondents with model-initiated help. However, this measure is limited in that it does not guarantee that the respondent is engaged in the survey task and respondents have different baseline response times that cannot be known before beginning most surveys. Using mouse movements to determine when respondents need help, on the other hand, does not have either of these limitations. We also discovered an additional benefit to using mouse movements to initiate help from the analysis in Chapter 4. These participants most commonly cited “distracting” as their primary reason for not liking the model-initiated help. Across all modes, the help initiated while participants were still reading the question text or response options and were not ready to process the information. However, they said that if it had been better timed, they would have liked it. Using mouse movements to initiate can help ensure the respondent is ready to receive help. Because the movements are related to difficulty, they do not engage in them until they are experiencing it. Therefore, using mouse movements instead of latencies could be critical to increasing respondent satisfaction with model-initiated help.

Not all surveys are in a position to provide respondents with unsolicited help. However, using mouse movements to predict respondent difficulty can provide benefits to all types of surveys. Specifically, the tools can be used to pretest questionnaires. Generally, pretests consist of relatively small populations and can be expensive. These pretests are successful at identifying larger problems with survey questions, but more subtle issues can go undetected. Instead of, or in addition to, standard pretests, Web survey administrators can send their surveys (with embedded JavaScript language to track mouse coordinates) to a larger pretest population and determine which questions are problematic for respondents based on the movements in which they engage, all for no additional cost. This type of analysis should uncover more potential issues with survey questions that administrators can correct before the survey enters the field.

Similarly, researchers can also use mouse movements to assess survey questions after the survey period. They may be able to help explain unanticipated results if it becomes clear respondents did not understand the question. Additionally, the movements can be crossed with respondent characteristics to see if only certain types of people have difficulty on a question, which could impact data quality.

All of the studies in this dissertation attempted to provide survey researchers and administrators tools to identify and help respondents who experience difficulty answering questions. Whether these tools are used to provide help in real time or outside of the survey period, they offer the ability to gain a greater understanding of respondent behavior at a much larger level for minimal additional cost.

Appendix A. Advertisements for Study 1

Craigslist Print Same-sex Relationship Advertisement

The US Census Bureau and the Joint Program in Survey Methodology at the University of Maryland need your help in identifying potentially confusing questions on the American Community Survey. We are seeking participants who:

- Are 18 or older
- Have at least one year of Internet experience
- Can come to the University of Maryland (via metro or car)
- Answer 'Yes' to **at least one** of the following questions:
    - Are you in a same-sex legally married couple?
    - Are you in a same-sex co-habitating couple?
    - Are you part of a same-sex couple in a registered domestic partnership?

The study involves answering 20 questions from the American Community Survey and rating their difficulty. It will take about 45 minutes to complete and eligible participants will be compensated $30 for time and travel. The study will be conducted at LeFrak Hall at the University of Maryland – College Park.

If interested, please respond to this posting, email umd.survey@yahoo.com, or call 301-836-1347.

Craigslist Print Rare Population Advertisement

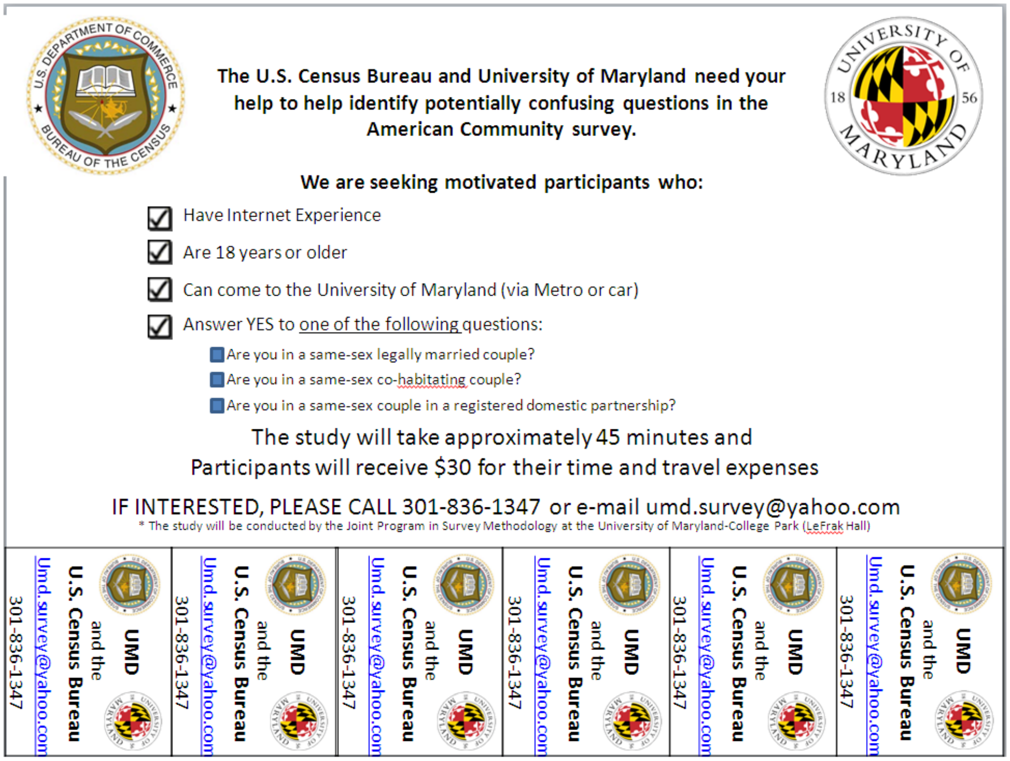
The US Census Bureau and the Joint Program in Survey Methodology at the University of Maryland need your help in identifying potentially confusing questions on the American Community Survey. We are seeking participants who:

- Are 18 or older
- Have at least one year of Internet experience
- Can come to the University of Maryland (via metro or car)
- Answer 'Yes' to **at least one** of the following questions:
    - Are you in a co-habitating opposite-sex relationship?
    - Have you attended a short course or continuing education class in the past 2 months?
    - Do you live in an apartment building with more than 5 units?
    - Do you have a joint custody arrangement where the dependent stays with the other parent most of the time?

The study involves answering 20 questions from the American Community Survey and rating their difficulty. It will take about 45 minutes to complete and eligible participants will be compensated $30 for time and travel. The study will be conducted at LeFrak Hall at the University of Maryland – College Park.

If interested, please respond to this posting, email umd.survey@yahoo.com, or call 301-836-1347.

Same-sex relationship flyer



The U.S. Census Bureau and University of Maryland need your help to help identify potentially confusing questions in the American Community survey.

We are seeking motivated participants who:

☑ Have Internet Experience

☑ Are 18 years or older

☑ Can come to the University of Maryland (via Metro or car)

☑ Answer YES to <u>one of the following</u> questions:

- Are you in a same-sex legally married couple?
- Are you in a same-sex co-habitating couple?
- Are you in a same-sex couple in a registered domestic partnership?

The study will take approximately 45 minutes and
Participants will receive $30 for their time and travel expenses

IF INTERESTED, PLEASE CALL 301-836-1347 or e-mail umd.survey@yahoo.com

* The study will be conducted by the Joint Program in Survey Methodology at the University of Maryland-College Park (LeFrak Hall)

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

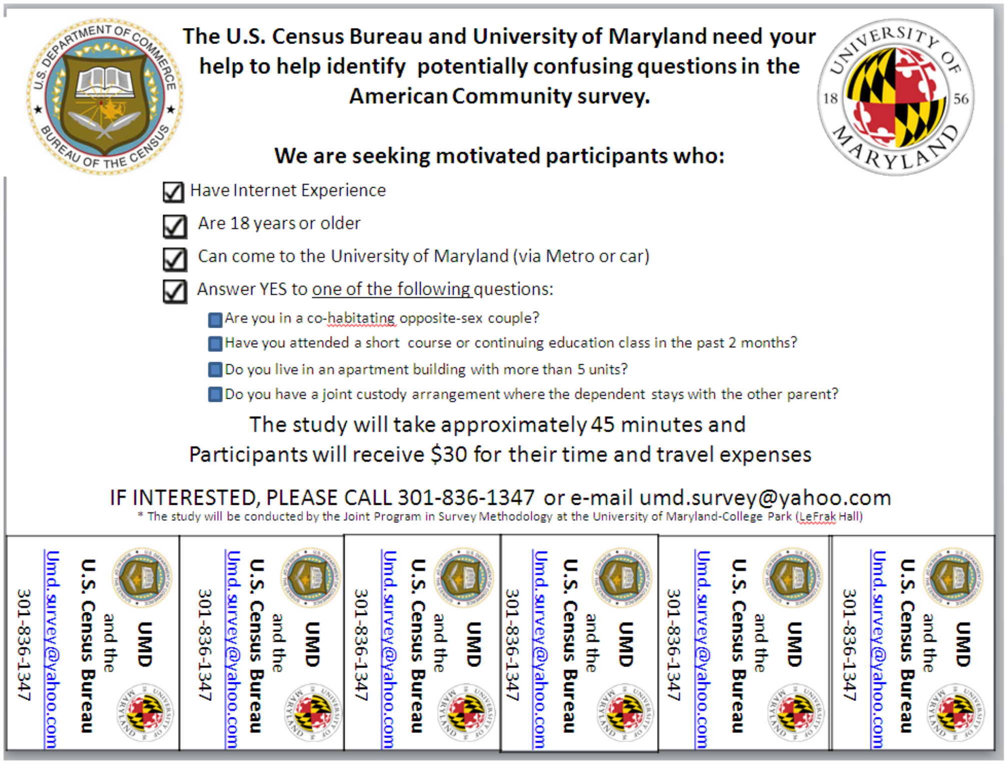
UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

Rare population flyer



The U.S. Census Bureau and University of Maryland need your help to help identify potentially confusing questions in the American Community survey.

We are seeking motivated participants who:

☑ Have Internet Experience

☑ Are 18 years or older

☑ Can come to the University of Maryland (via Metro or car)

☑ Answer YES to <u>one of the following</u> questions:

- Are you in a co-habitating opposite-sex couple?
- Have you attended a short course or continuing education class in the past 2 months?
- Do you live in an apartment building with more than 5 units?
- Do you have a joint custody arrangement where the dependent stays with the other parent?

The study will take approximately 45 minutes and
Participants will receive $30 for their time and travel expenses

IF INTERESTED, PLEASE CALL 301-836-1347 or e-mail umd.survey@yahoo.com

* The study will be conducted by the Joint Program in Survey Methodology at the University of Maryland-College Park (LeFrak Hall)

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

UMD and the **U.S. Census Bureau** Umd.survey@yahoo.com 301-836-1347

Appendix B. Screener for Study 1

Phone Screener

Thanks for your interest in participating in a Research Study with the U.S. Census Bureau and the University of Maryland. Our study will take about 45 minutes to complete and depending upon your eligibility you will receive a $30 cash reimbursement for traveling expenses to the University of Maryland, which is located in College Park, MD. In this study you will be answering a series of questions from the American Community Survey to help us identify any questions that might be confusing. Prior to scheduling you to participate in this study, I will need to gather some information from you to add you to our participant database. The information that you provide is confidential and you can decline to answer any questions if you do not want to provide the information at this time. We use this information to see which studies you will be most eligible for and to ensure that we have a diverse group of participants for each of our studies.

General

1. **Participant Number**

2. **Let's start with the spelling of your first and last name:**

   First name   Last name

3. **Are you male/female? (if necessary)**
   □ Male
   □ Female

4. **Have you participated in any research studies in the past year?**
   □ Yes
   How many? ____
   □ No

5. **Are you currently employed?** *If yes, go to 6. If no, skip to 7.*
   □ Yes
   □ No

6. **Are you employed by the Federal government?**
   □ Yes
   □ No

Demographics

7. **What is your current age?**

8. **I am going to read you a list of races. Please tell me which race, or races, with which you identify.**
Note: If reluctant, inform them that this does not impact their ability to participate in our studies-we gather this information to ensure that we have a diverse group of participants for each of our studies
□ White
□ Black or African American
□ Asian
□ American Indian or Alaska Native
□ Native Hawaiian or other Pacific Islander
□ Other

9. **What is your highest level of education COMPLETED?**
□ Never completed any school
□ Completed ninth grade or below
□ Some high school
□ Completed high school, or received a GED
□ Vocational training beyond high school
□ Some college
□ Completed college
□ Any graduate or professional education

10. **Are you currently married?** *If 'yes', skip to 12. If 'no', go to 11.*
□ Yes
□ No

11. **Are you currently in a registered domestic partnership or civil union?** *If 'yes', skip to 13. If 'no', go to 12.*
□ Yes
□ No

12. **Are you currently living with a partner of the same sex?** *If 'yes', skip to 14. If 'no', go to 13.*
□ Yes
□ No

13. **Are you currently living with a partner of the opposite sex?**
□ Yes
□ No

14. **Do you have a child in a joint custody arrangement?** *If 'no', skip to 16. If 'yes', go to 15.*
□ Yes
□ No

15. **On an average week, how much time does the child spend with you?** *Read options aloud*
□ Less than 50%
□ 50%
□ More than 50%

16. **Have you taken any continuing education or short courses in the past 2 months?**
□ Yes
□ No

17. **Do you currently live in an apartment building with 5 or more units?**
□ Yes
□ No

Computer/Internet Experience

18. **Do you have at least one year internet experience?**
□ Yes
□ No (*Inform them that most of our studies require at least one year internet experience and that we will keep them in mind for other studies that do not require internet experience.*) *SKIP to Contact Information*

19. **Name at least two things that you use the internet to do?**
a. ____________________
b.____________________

20. **Do you currently use the internet at least twice per week for searching for information, filling out forms, or shopping online?**
□ Yes
□ No

Contact Information

21. **Is there a daytime phone number where you can be reached? (Include preferred contact times if any)**
Home_______________
Mobile_______________
Work________________
E-mail____________________

22. **How did you hear about our study?**
□ Craigslist
□ Newspaper ad_____________
□ Flier posting____________
□ Referral
□ Other______________

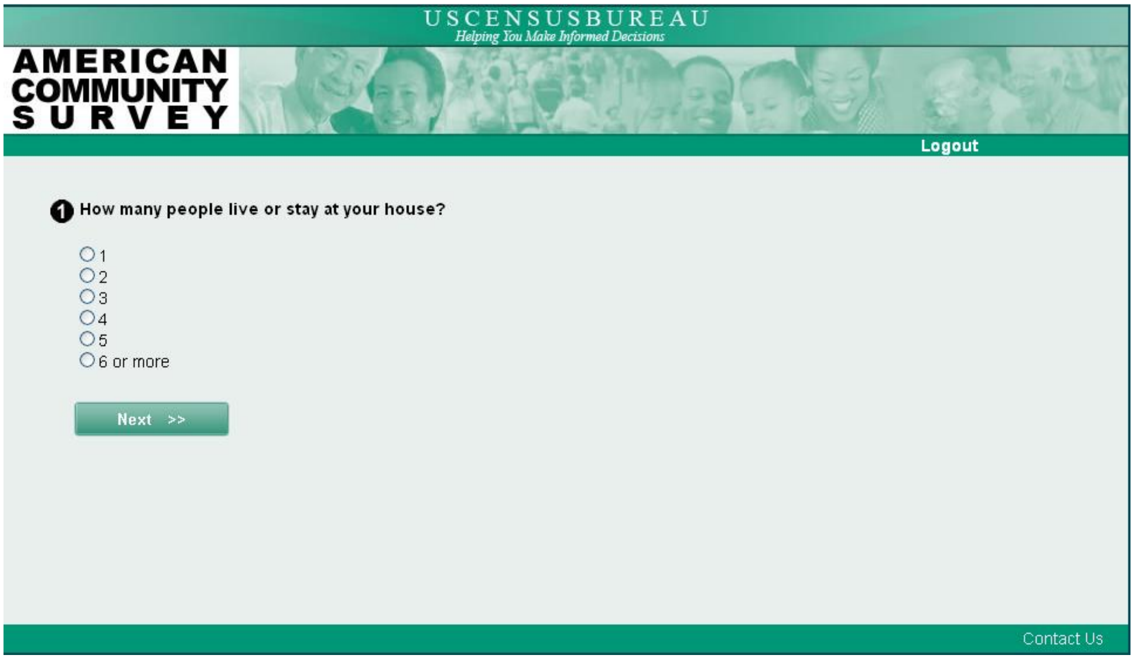
**If a caller is eligible but works for the Federal government, tell them they are eligible but they will not be able to receive the $30 reimbursement.

*If a caller is ineligible, read them the following text:

Unfortunately at this time you do not meet our eligibility requirements for this study. However, we will be conducting several additional studies over the next few months which you may be eligible for. Would it be okay if we keep your information on file and contact you for one of these studies?

Appendix C. ACS Questions Used in Study 1[30]

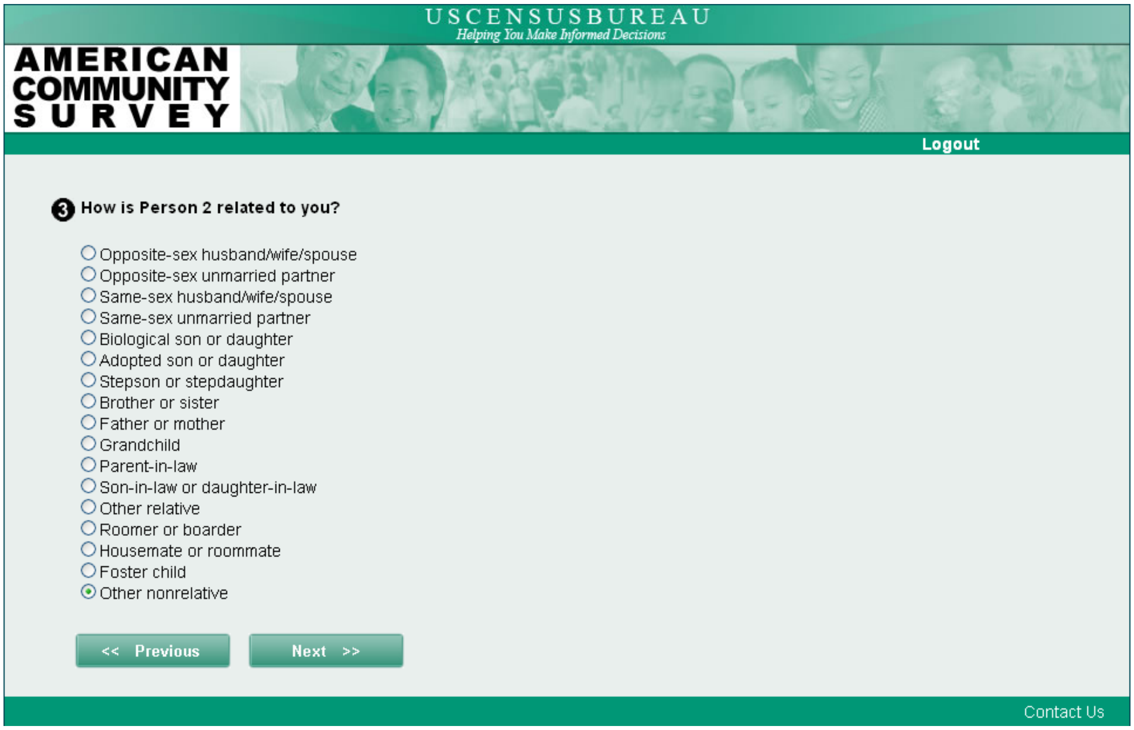
Live or stay


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
Next >>
Contact Us

1 How many people live or stay at your house?

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Relationship

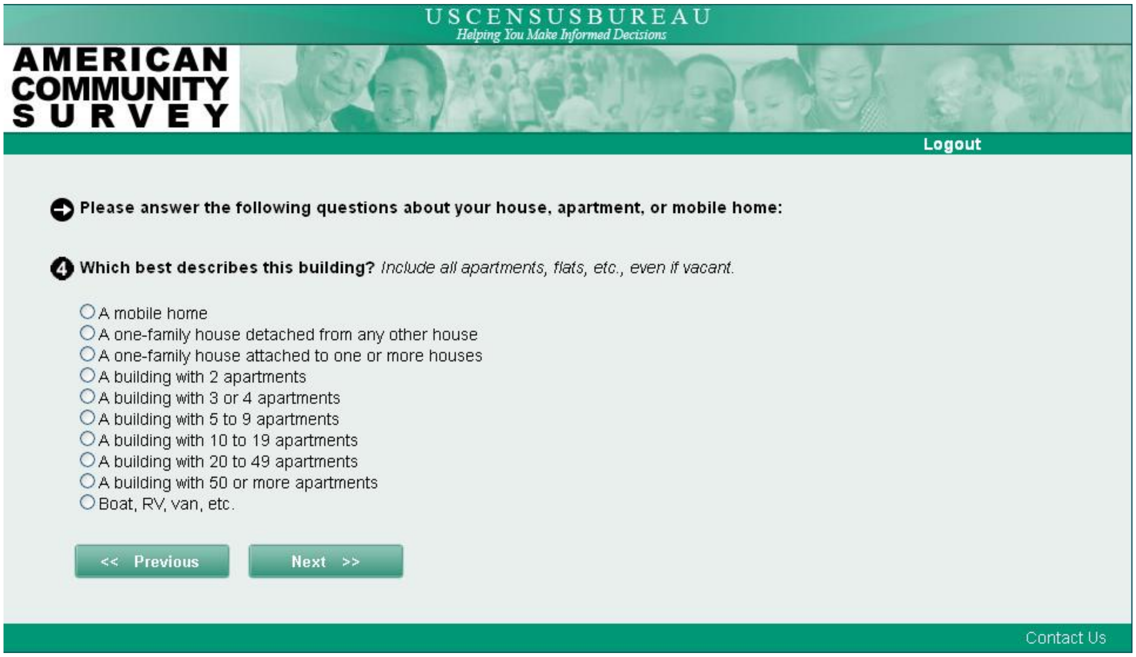

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

3 How is Person 2 related to you?

○ Opposite-sex husband/wife/spouse
○ Opposite-sex unmarried partner
○ Same-sex husband/wife/spouse
○ Same-sex unmarried partner
○ Biological son or daughter
○ Adopted son or daughter
○ Stepson or stepdaughter
○ Brother or sister
○ Father or mother
○ Grandchild
○ Parent-in-law
○ Son-in-law or daughter-in-law
○ Other relative
○ Roomer or boarder
○ Housemate or roommate
○ Foster child
● Other nonrelative

[30] Question 2 was intentionally omitted. It was a rostering question, which only purpose was to generate a name fill for question 3. No analysis was done on this question.

Type of unit

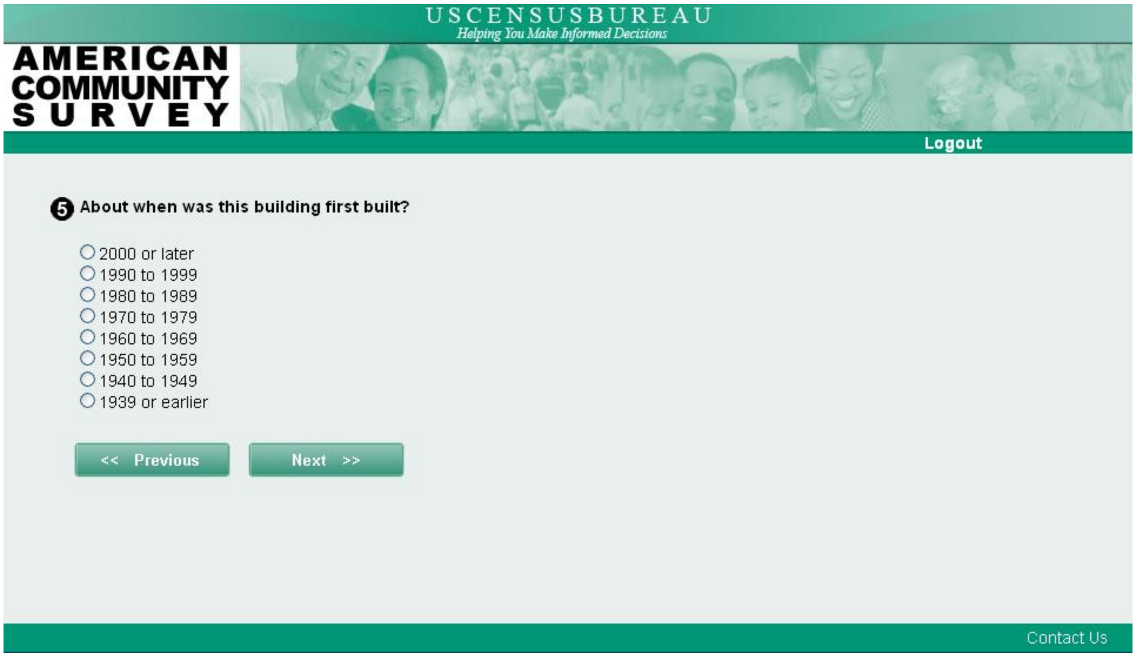

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

Please answer the following questions about your house, apartment, or mobile home:

4 Which best describes this building? *Include all apartments, flats, etc., even if vacant.*

○ A mobile home
○ A one-family house detached from any other house
○ A one-family house attached to one or more houses
○ A building with 2 apartments
○ A building with 3 or 4 apartments
○ A building with 5 to 9 apartments
○ A building with 10 to 19 apartments
○ A building with 20 to 49 apartments
○ A building with 50 or more apartments
○ Boat, RV, van, etc.

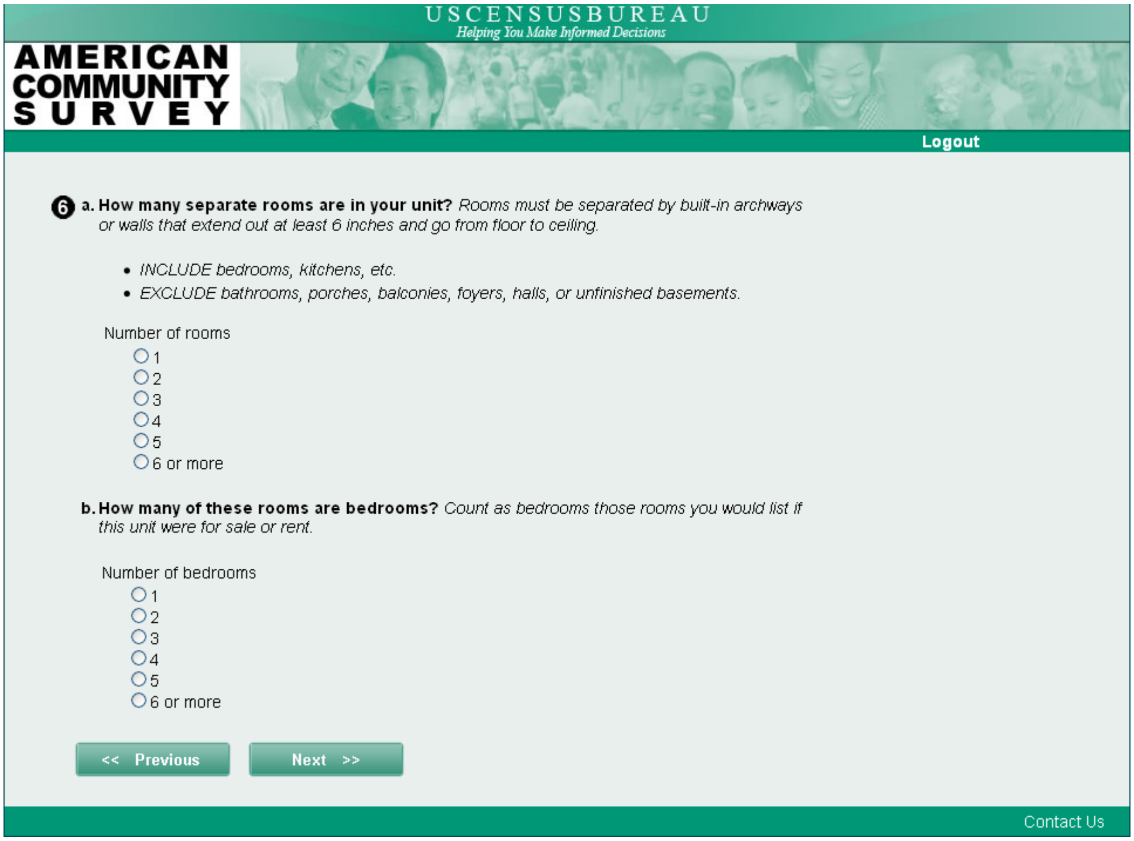
Year built


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

5 About when was this building first built?

○ 2000 or later
○ 1990 to 1999
○ 1980 to 1989
○ 1970 to 1979
○ 1960 to 1969
○ 1950 to 1959
○ 1940 to 1949
○ 1939 or earlier

Rooms



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

❻ a. **How many separate rooms are in your unit?** *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms

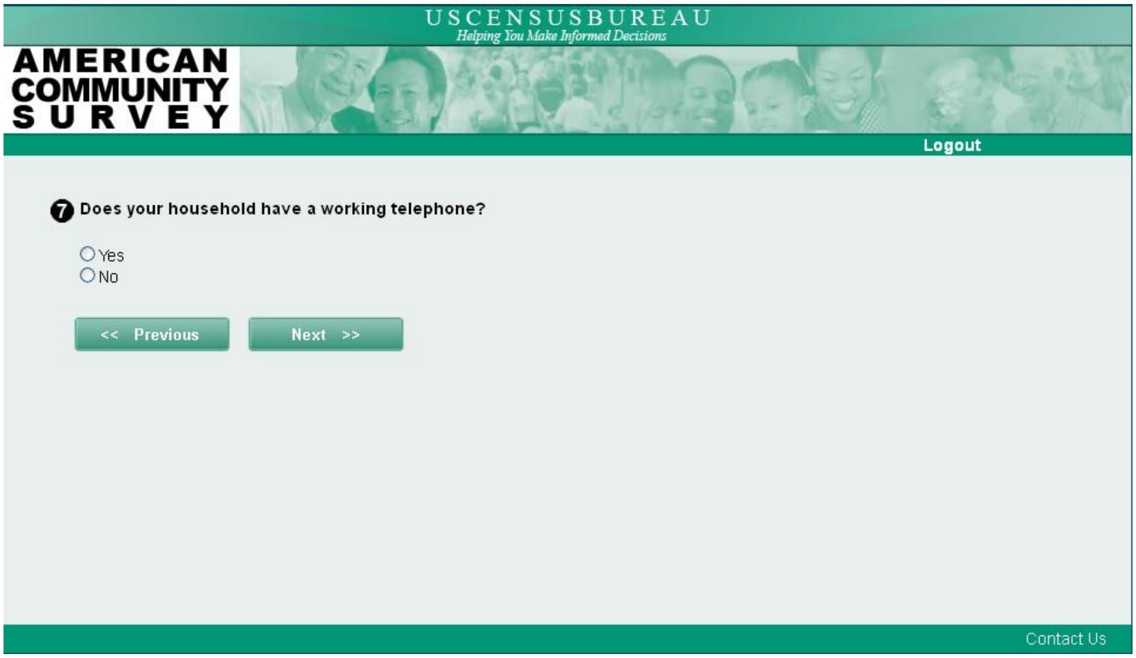
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

b. **How many of these rooms are bedrooms?** *Count as bedrooms those rooms you would list if this unit were for sale or rent.*

Number of bedrooms

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Telephone



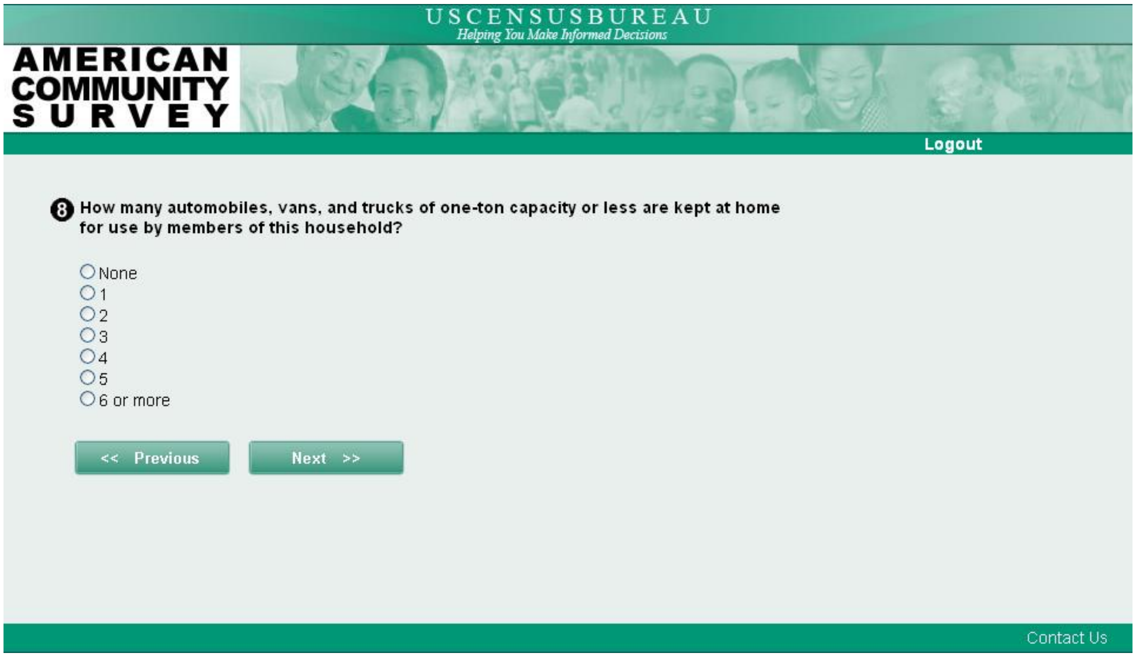
USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

❼ **Does your household have a working telephone?**

○ Yes
○ No

<< Previous
Next >>
Contact Us

Vehicles

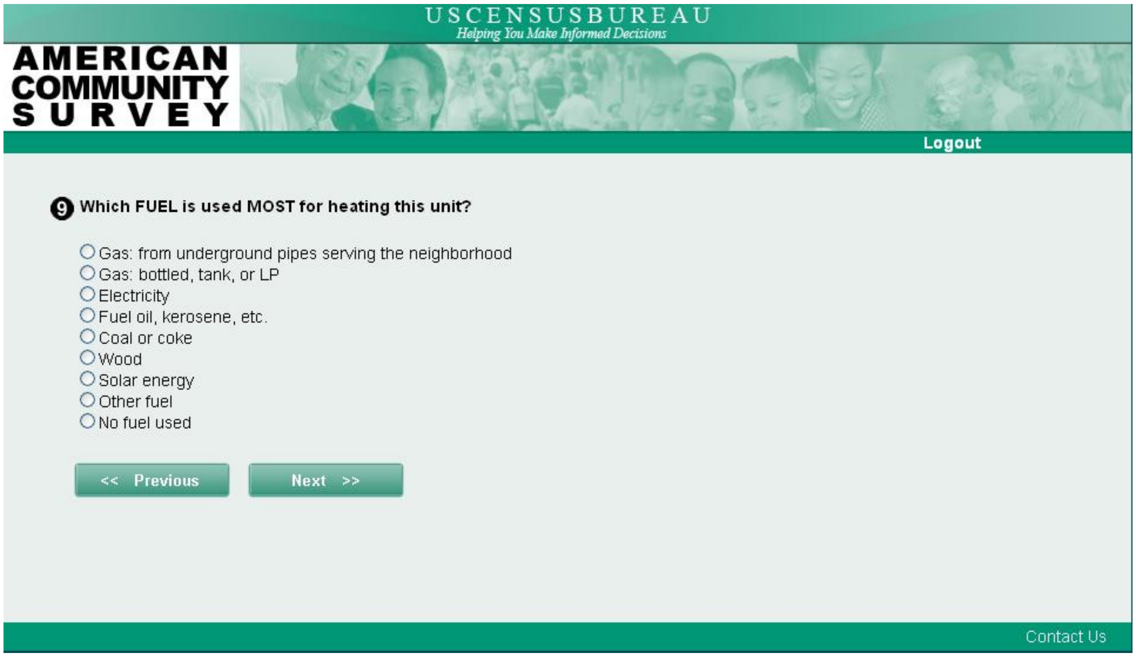

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

❽ How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of this household?

○ None
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

<< Previous
Next >>
Contact Us

Fuel

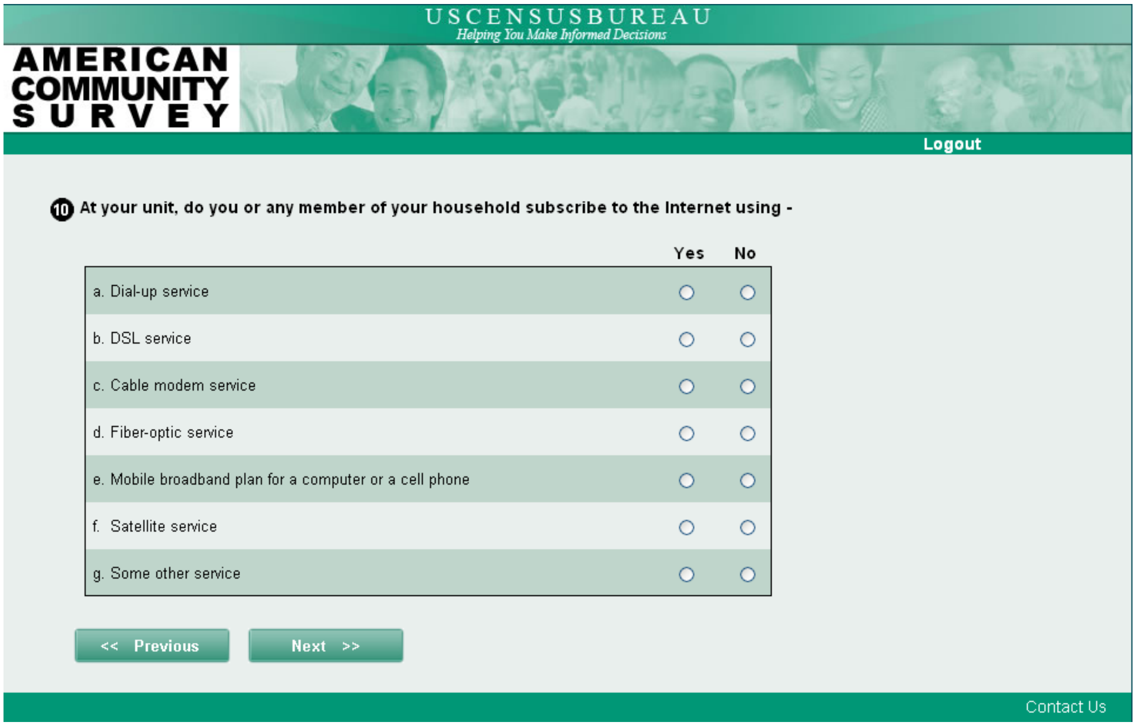

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

❾ Which FUEL is used MOST for heating this unit?

○ Gas: from underground pipes serving the neighborhood
○ Gas: bottled, tank, or LP
○ Electricity
○ Fuel oil, kerosene, etc.
○ Coal or coke
○ Wood
○ Solar energy
○ Other fuel
○ No fuel used

<< Previous
Next >>
Contact Us

Internet

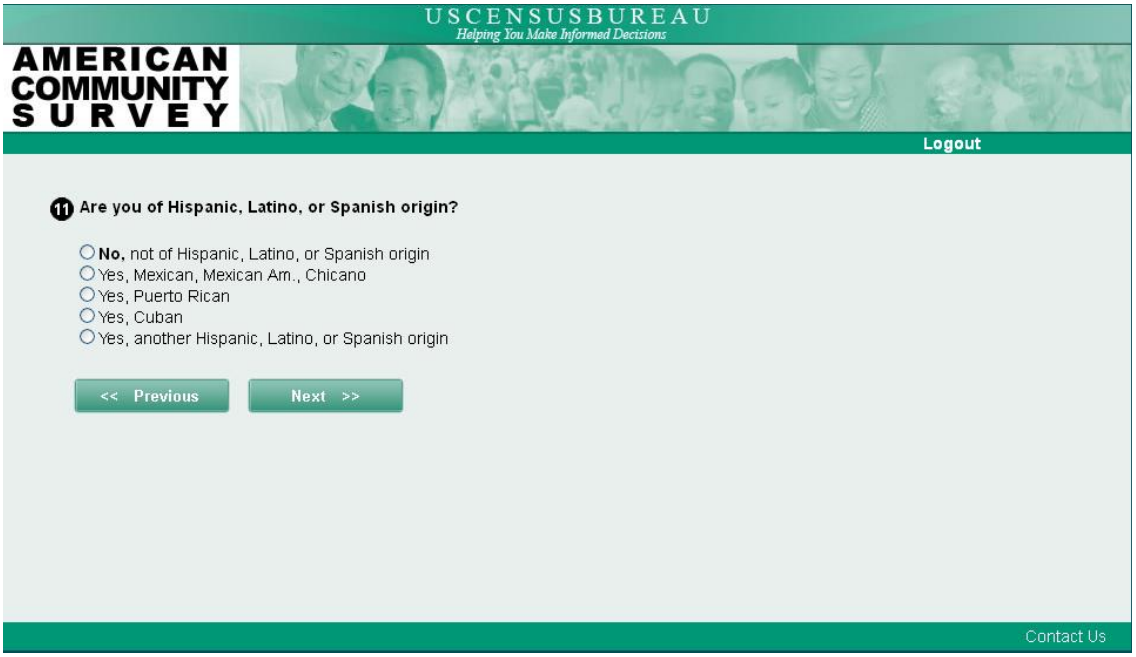

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

10 At your unit, do you or any member of your household subscribe to the Internet using -

	Yes	No
a. Dial-up service	○	○
b. DSL service	○	○
c. Cable modem service	○	○
d. Fiber-optic service	○	○
e. Mobile broadband plan for a computer or a cell phone	○	○
f. Satellite service	○	○
g. Some other service	○	○

<< Previous
Next >>
Contact Us

Hispanic

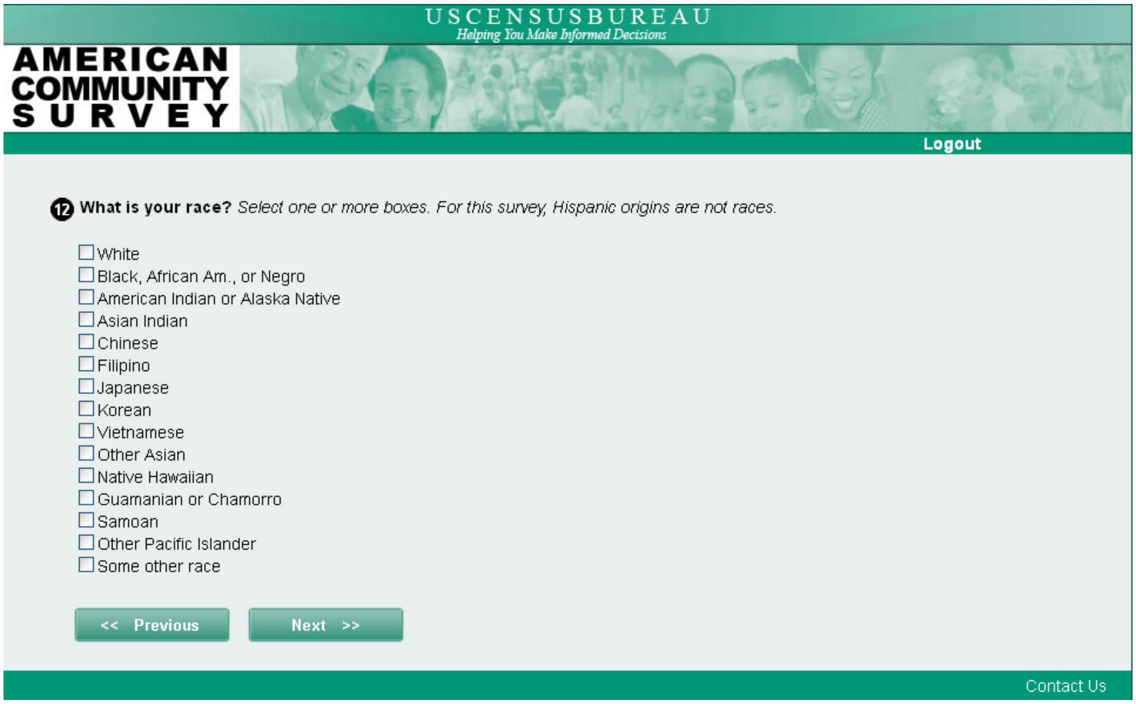

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

11 Are you of Hispanic, Latino, or Spanish origin?

○ **No,** not of Hispanic, Latino, or Spanish origin
○ Yes, Mexican, Mexican Am., Chicano
○ Yes, Puerto Rican
○ Yes, Cuban
○ Yes, another Hispanic, Latino, or Spanish origin

<< Previous
Next >>
Contact Us

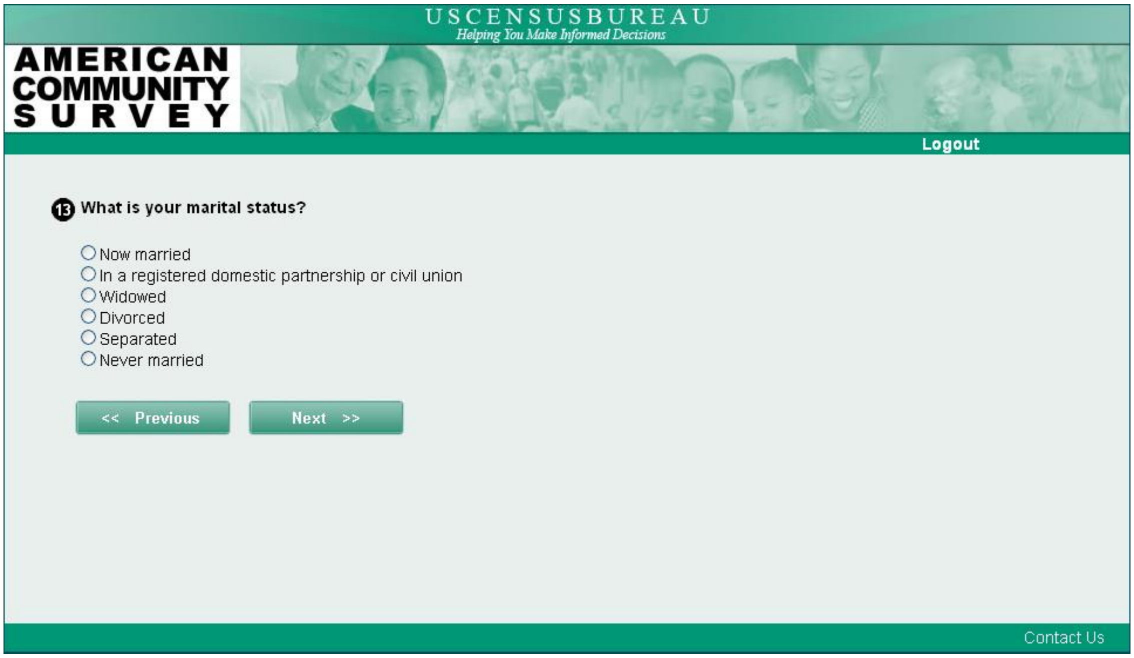
Race


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

12 **What is your race?** *Select one or more boxes. For this survey, Hispanic origins are not races.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native
□ Asian Indian
□ Chinese
□ Filipino
□ Japanese
□ Korean
□ Vietnamese
□ Other Asian
□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander
□ Some other race

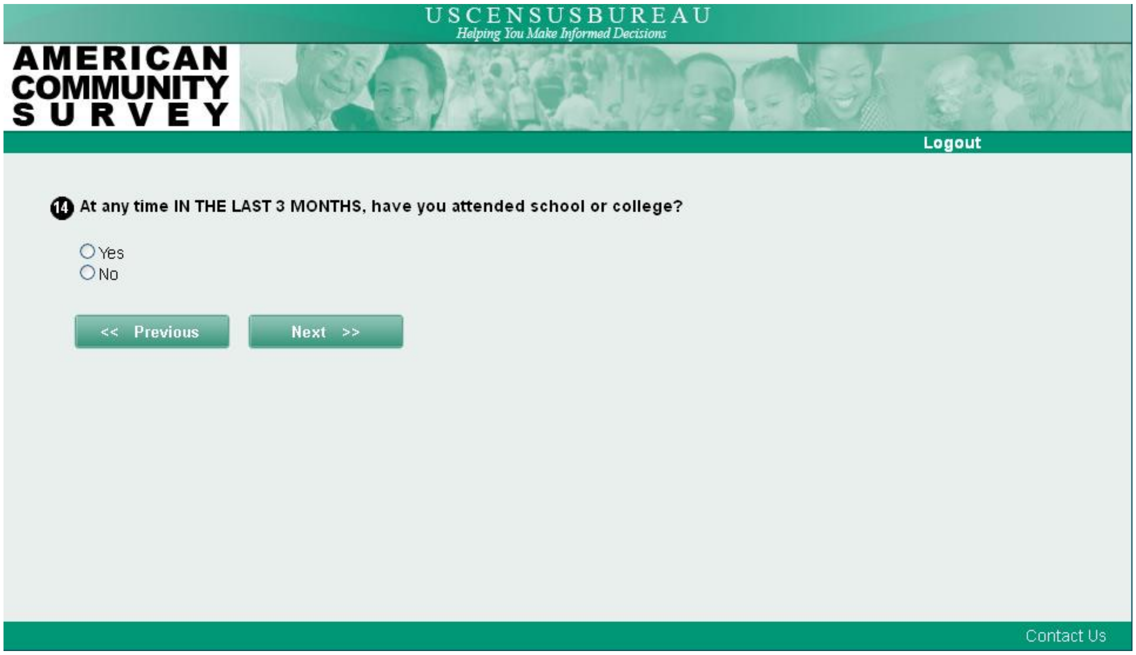
Marital status


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

13 **What is your marital status?**

○ Now married
○ In a registered domestic partnership or civil union
○ Widowed
○ Divorced
○ Separated
○ Never married

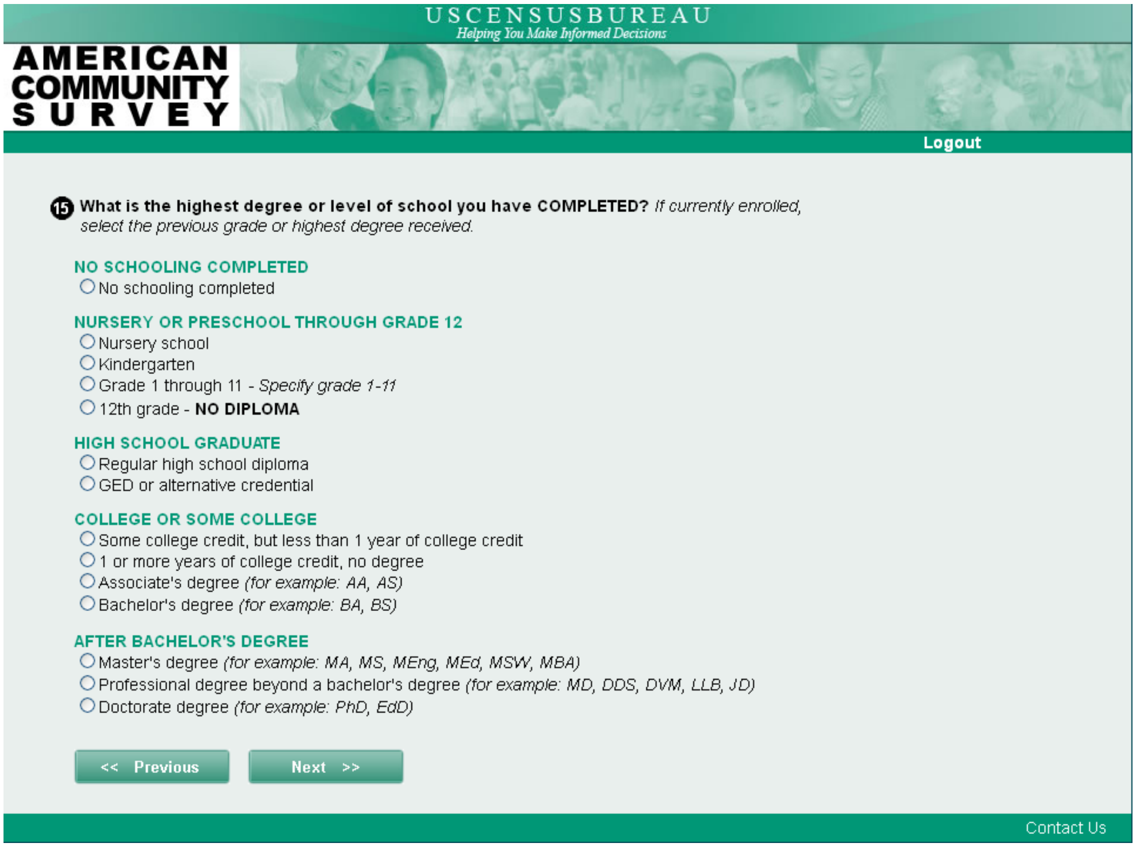
Attend school


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

14 **At any time IN THE LAST 3 MONTHS, have you attended school or college?**

○ Yes
○ No

Educational attainment


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

15 **What is the highest degree or level of school you have COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.*

NO SCHOOLING COMPLETED
○ No schooling completed

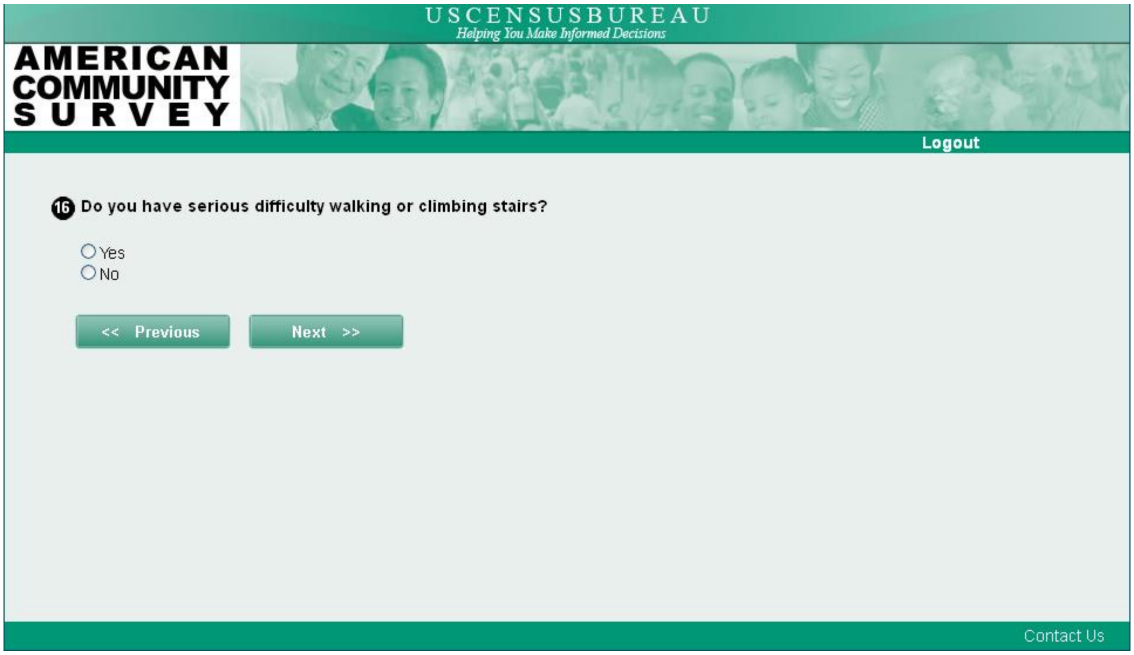
NURSERY OR PRESCHOOL THROUGH GRADE 12
○ Nursery school
○ Kindergarten
○ Grade 1 through 11 - *Specify grade 1-11*
○ 12th grade - **NO DIPLOMA**

HIGH SCHOOL GRADUATE
○ Regular high school diploma
○ GED or alternative credential

COLLEGE OR SOME COLLEGE
○ Some college credit, but less than 1 year of college credit
○ 1 or more years of college credit, no degree
○ Associate's degree *(for example: AA, AS)*
○ Bachelor's degree *(for example: BA, BS)*

AFTER BACHELOR'S DEGREE
○ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
○ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
○ Doctorate degree *(for example: PhD, EdD)*

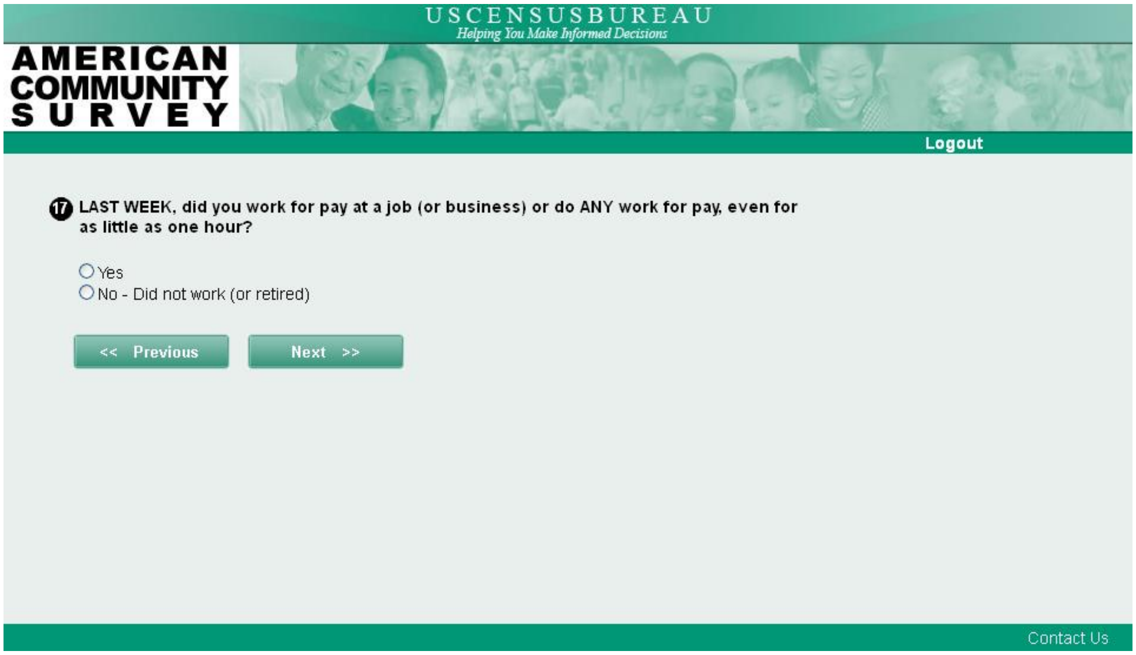
Difficulty walking


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

16 Do you have serious difficulty walking or climbing stairs?

○ Yes
○ No

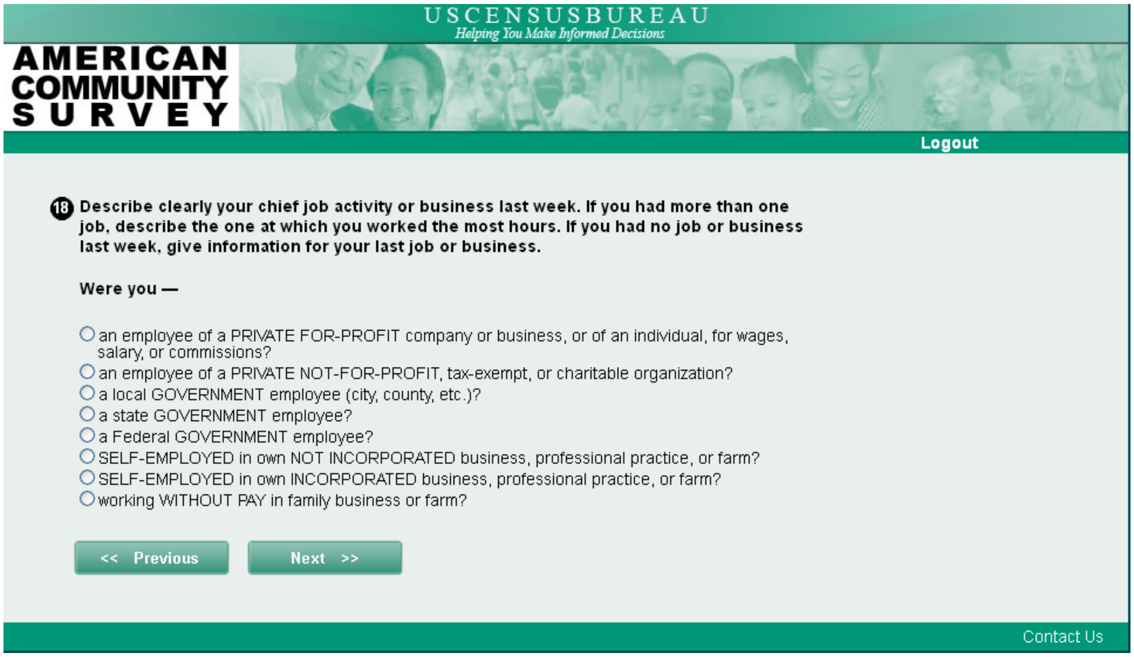
Work last week


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

17 LAST WEEK, did you work for pay at a job (or business) or do ANY work for pay, even for as little as one hour?

○ Yes
○ No - Did not work (or retired)

Employee type

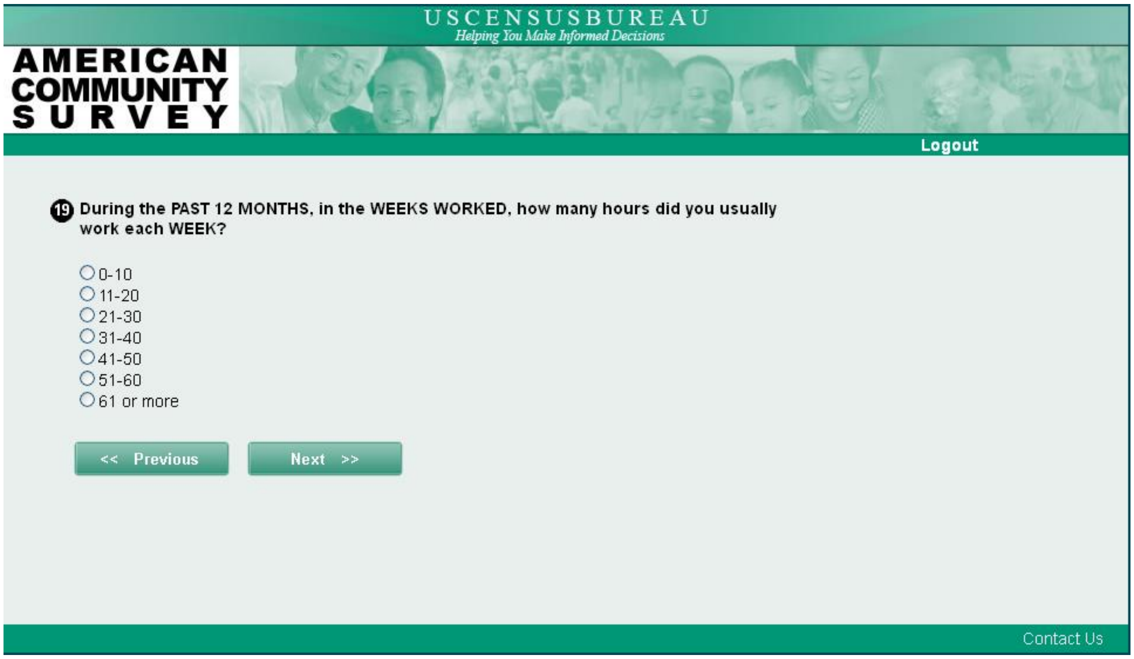

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

18 Describe clearly your chief job activity or business last week. If you had more than one job, describe the one at which you worked the most hours. If you had no job or business last week, give information for your last job or business.

Were you —

○ an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
○ an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
○ a local GOVERNMENT employee (city, county, etc.)?
○ a state GOVERNMENT employee?
○ a Federal GOVERNMENT employee?
○ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
○ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
○ working WITHOUT PAY in family business or farm?

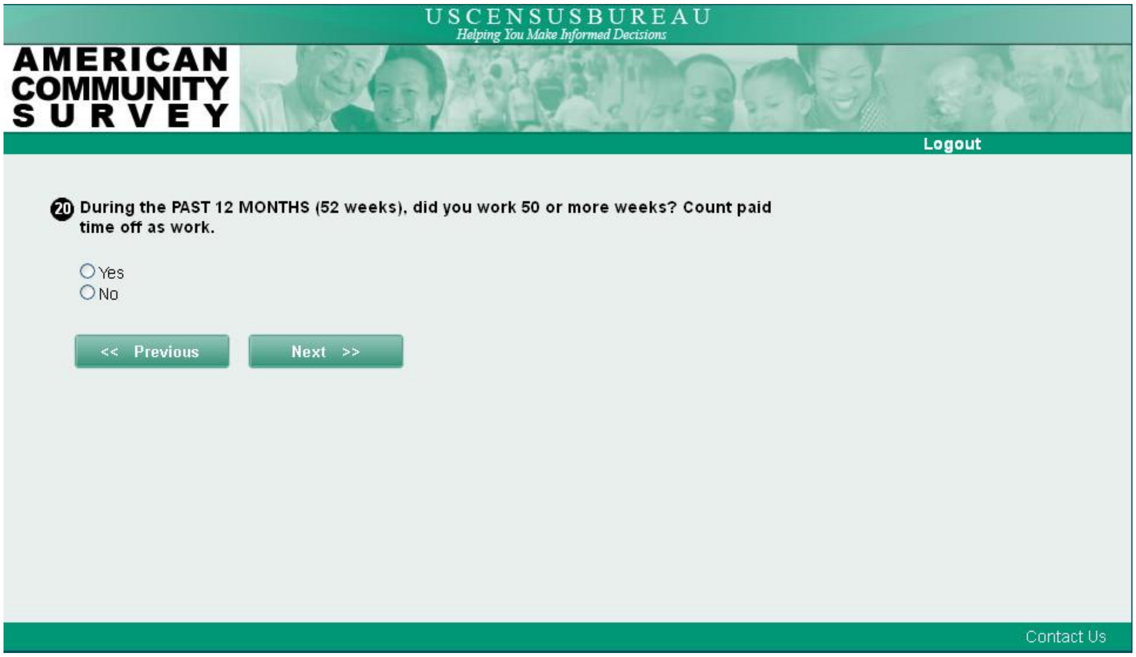
Hours worked


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
<< Previous
Next >>
Contact Us

19 During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did you usually work each WEEK?

○ 0-10
○ 11-20
○ 21-30
○ 31-40
○ 41-50
○ 51-60
○ 61 or more

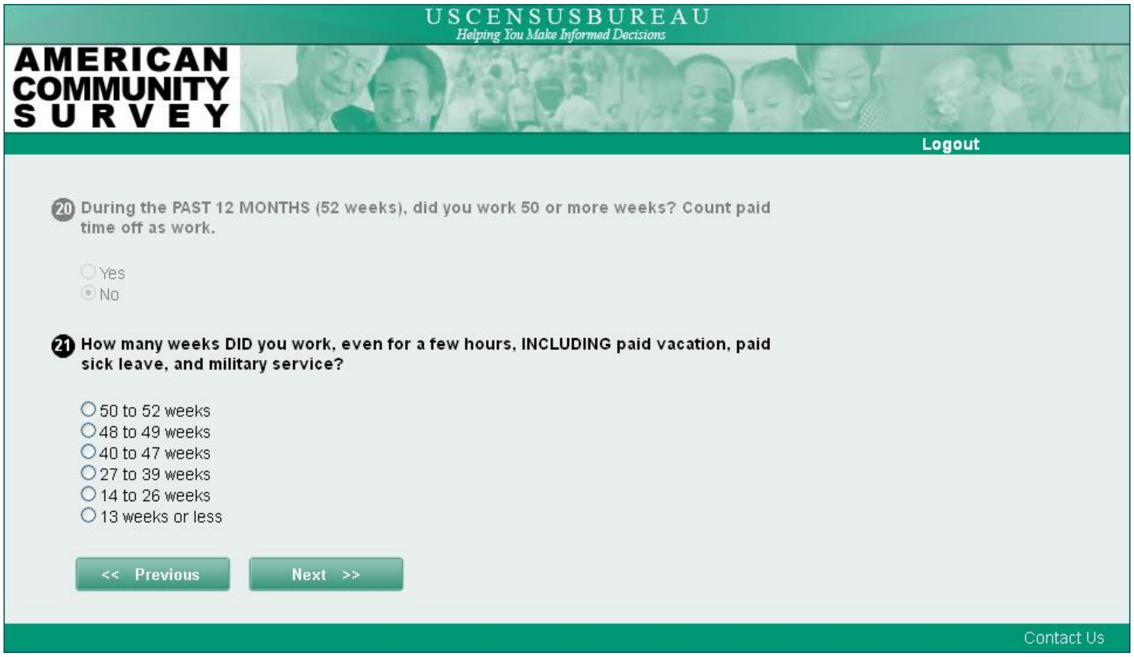
Weeks worked


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

20 During the PAST 12 MONTHS (52 weeks), did you work 50 or more weeks? Count paid time off as work.

○ Yes
○ No

<< Previous
Next >>
Contact Us


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

20 During the PAST 12 MONTHS (52 weeks), did you work 50 or more weeks? Count paid time off as work.

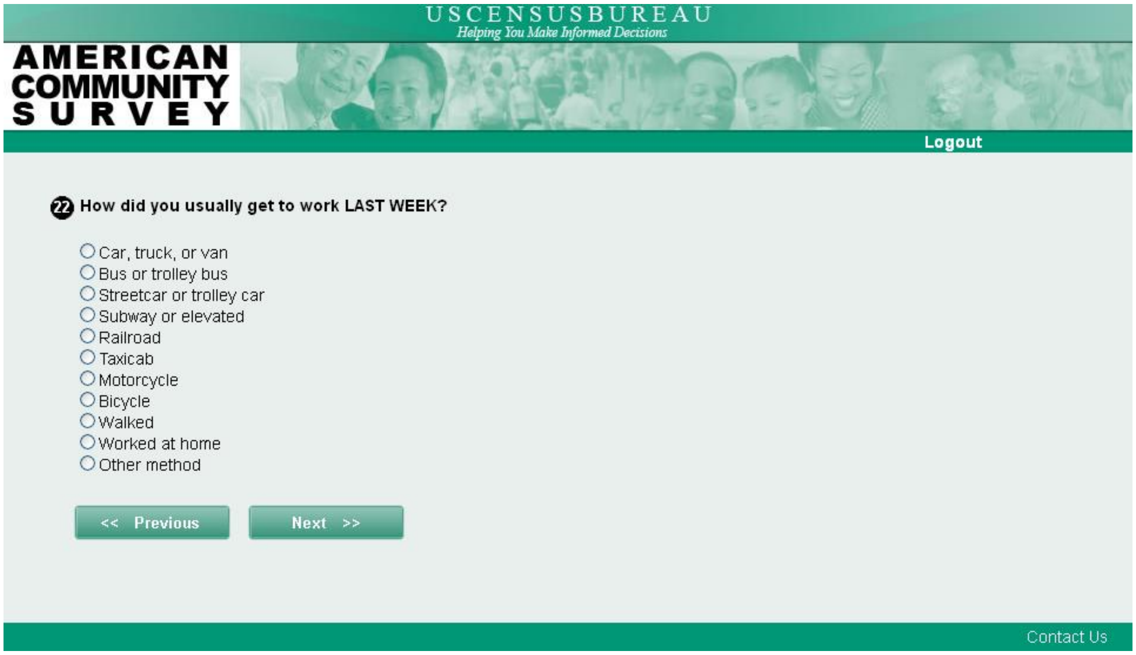
○ Yes
● No

21 How many weeks DID you work, even for a few hours, INCLUDING paid vacation, paid sick leave, and military service?

○ 50 to 52 weeks
○ 48 to 49 weeks
○ 40 to 47 weeks
○ 27 to 39 weeks
○ 14 to 26 weeks
○ 13 weeks or less

<< Previous
Next >>
Contact Us

Transport to work



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

22 How did you usually get to work LAST WEEK?

○ Car, truck, or van
○ Bus or trolley bus
○ Streetcar or trolley car
○ Subway or elevated
○ Railroad
○ Taxicab
○ Motorcycle
○ Bicycle
○ Walked
○ Worked at home
○ Other method

<< Previous
Next >>
Contact Us

Appendix D. Protocol for Study 1

Date______________________; Participant #________; Experimenter:_______________

General Introduction: Measuring Question Difficulty on the American Community Survey Internet Instrument

Thank you for your time today. My name is Rachel Horwitz and I am a student here in the Joint Program of Survey Methodology and am also employed at the US Census Bureau in the American Community Survey Data Collection area. I will be working with you today. If you have a cell phone, please turn it off or put it in vibrate. In order to help us improve our surveys, we turn to people like you to find out if our questions make sense and are fairly easy to understand and answer. We have found that the best way to do that is to actually conduct the survey with people and see how it works for them. So you will be helping us test a questionnaire from one of our surveys. I did not create the survey, so please share both your positive and negative reactions to it. The entire session should last 30-45 minutes. Your comments and feedback will be given to the developers of the survey and may be used to improve it.

First, I would like to ask you to read and sign this consent form. It explains the purpose of today's session and informs you of your rights as a participant. It also tells you that we would like to record the session, with your permission. Only those of us connected with the project will review the recording and any other data collected during the session, and it will be used solely for research purposes. We may also use clips from the recording to illustrate key points about the survey to the Web design team.

Hand the participant the consent form; give time to read and sign; sign own name and date if you have not already done so.

Start the tape.
While you are completing the survey, we will record the movements of your eyes with our eye-tracking monitor to get a record of where you are looking on the screen and we will record your mouse movements to see how you are interacting with the survey.
Now I am going to calibrate your eyes for the eye-tracking.

Do Calibration

A random five to ten respondents will take the survey using a think aloud procedure:

I would like you to tell me your impressions and thoughts about the questions as you read and answer them. I would like you to \think aloud" and talk to me about your impressions. If you expect to see some piece of information, tell me about that expectation as well. Finally, during the session, I may remind you to talk to me if you get quiet, not to interrupt your thought process simply to remind you to talk to me. Please focus on verbalizing what you are thinking as you read.

Before we get started, let's practice thinking aloud, since it's not something that you would normally do while reading letters. Can you tell me how many windows are in your house or apartment? [PROBE as appropriate to the participant's responses to this question.]

After think aloud practice is complete:

Now that we have your eyes calibrated, we are ready to begin. Please respond to the survey questions online. Please answer the questions as they apply to you in your real life.

Do you have any questions?

Start the eye-tracking software: Tobii Studio.

Leave room. Once in control room do a sound check and Start the eye-tracking software: Tobii Studio.

The mouse tracing software will start when Studio opens Internet Explorer.

Do not offer any assistance to respondents or answer questions other than "Please use your best judgment" or "how would you answer if you were at home?"

Overall Probe: Make a note if a person left a page with a blank answer, asked a question to the researcher, or displayed signs of confusion (hovers, regressions, using the mouse as a marker).

What was your overall impression of the survey?

Were there any questions you found to be difficult or challenging to answer?

> If yes, show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.

Were there any responses you were unsure of?

> If yes show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.

If respondents reported being confused by a question or displayed signs of confusion, ask them

> What was it about this question that you had trouble with (understanding the question, understanding the response options, applying their situation to one of the response options)?
>
> How did you come up with your answer?

For any questions the respondent displayed signs of confusion, ask if they had any trouble answering the question and what type of trouble they had. If they were not sure which answer category to select, ask them to describe their situation.

If they mentioned that someone in their household uses more than one mode of transportation to get to work (such as bus and subway) and they chose one, ask why they chose that one.

For attended school in the past 3 months

What does attended school mean to you?

For work questions

Do you currently have more than one job?

If yes, did you answer about one of those jobs or all of them?

For Relationship Question

Why did you answer the relationship question the way you did?
Were there any other choices that you considered?
Do you think your answer adequately describes your situation? Why or why not?

For Marital Status Question

Why did you answer the marital status question the way you did?

If R reports now married:

- What was the date of your marriage?
- In what city and state did the marriage take place?
- Was this a commitment ceremony or a legal marriage ceremony?

If R reports registered domestic partner or civil union:

- What was the date of your domestic partnership/civil union?
- In what city and state did the registration take place?

Do you think your answer adequately describes your situation? Why or why not?

What does “registered domestic partnership or civil union” mean to you in Q. 23?

If necessary, ask respondents in opposite-sex relationships if they know anyone who might select they are in a domestic partnership or civil union and have them explain what type of living arrangement/relationship that was.

Thank you again for your participation today. It is greatly appreciated.

Appendix E. Consent Form, Demographic, and Computer Experience Questions (all studies)

Usability Study of the American Community Survey

Each year, the Census Bureau and the Joint Program in Survey Methodology at the University of Maryland separately conduct many different usability evaluations. For example, they routinely test the wording, layout and behavior of products, such as Web sites, online surveys, and letters sent through the mail in order to obtain the best information possible from respondents.

You have volunteered to take part in a study to identify confusing questions in an Internet version of the American Community Survey (ACS). In order to have a complete record of your comments, your usability session will be videotaped. We plan to use the tapes to improve the design of the product. Staff directly involved in the usable design research project will have access to the tapes. We also plan to perform an eye-tracking analysis of your session. Your participation is voluntary and your answers will remain strictly confidential. You may skip any questions you do not wish to answer.

This usability study is being conducted under the authority of Title 13 USC. The OMB control number for this study is 0607-0725. This valid approval number legally certifies this information collection.
I have volunteered to participate in this usability study sponsored by the Census Bureau and the University of Maryland, and I give permission for my tapes to be used for the purposes stated above.

Participant's Name: ___________________________________
Participant's Signature:_____________________________ Date: ___________________
Researcher's Name: _______________________________ Date: ___________________
Researcher's Name: _______________________________ Date: ___________________

Investigator Contact Information
Rachel Horwitz
301-314-9916
rhorwitz@survey.umd.edu
IRB Office Contact Information
1204 Marie Mount
College Park, MD 20742
301-504-4212

Questionnaire on Computer-and-Internet Experience and Demographics

YOUR ANSWERS ARE CONFIDENTIAL

Demographics

1. What is your age? ______________

2. Are you male or female? _____________

3. Are you of Hispanic, Latino, or Spanish origin?
   □ Yes
   □ No

4. What is your race? Choose **one or more** races.
   □ White
   □ Black or African American
   □ American Indian or Alaska Native
   □ Asian
   □ Native Hawaiian or Other Pacific Islander

5. What is your level of education?
   __grade school
   __some high school
   __high school degree
   __some college
   __2-year college degree
   __4-year college degree
   __some postgraduate study (e.g., M.A., M.B.A., J.D., Ph.D., M.D., programs)
   __postgraudate degree (e.g., M.A., M.B.A., J.D., Ph.D., M.D.)

Computer and Internet Experience

1. During the last month, about how many hours did you use the Internet **during a typical week? __________**

2. How much experience have you had with computers **to use the Internet**? Please check one option.

   □ A great Deal
   □ A lot
   □ A moderate amount
   □ A little
   □ None

3. How much experience have you had with computers **to do things other than use the Internet**? Please check one option.

   □ A great Deal
   □ A lot
   □ A moderate amount
   □ A little
   □ None

4. During the last year, how many times did you complete a survey on the Internet? ______

5. During the last month, where were you when you used the Internet? Please select all that apply.

   □ Home
   □ Work
   □ School
   □ Library
   □ Another place (please specify: ______________)

6. Do you have a computer at your home?
   □ Yes
   □ No → (Skip to Question 7)
   6a. What type of Internet connection do you use at home? Please check all that apply.
      □ Cable Service
      □ Dial-up Service
      □ DSL Service
      □ Satellite Dish Service
      □ Fiber-Optic Service
      □ Mobile Broadband Plan for a Computer or Cell Phone
   6b. Browsers are software on a computer used to surf the Internet. Last month, which Internet browse did you typically use at home? Please check all that apply.
      □ Firefox
      □ Google Chrome
      □ Internet Explorer
      □ Safari
      □ Other: ___________
   6c. During the last month, what was the operating system on the computer you typically used at home?
      □ Linux
      □ MAC OS (version: ____________________)
      □ Windows (version: ____________________)
      □ Other (specify: ________________)
7. Name at least two things you use the computer to do.
   a. _______________
   b. _______________

8. Name at least two things you use the Internet to do.
   a. _______________
   b. _______________

9. How difficult is it for you to learn to use Web sites that you have not visited before?

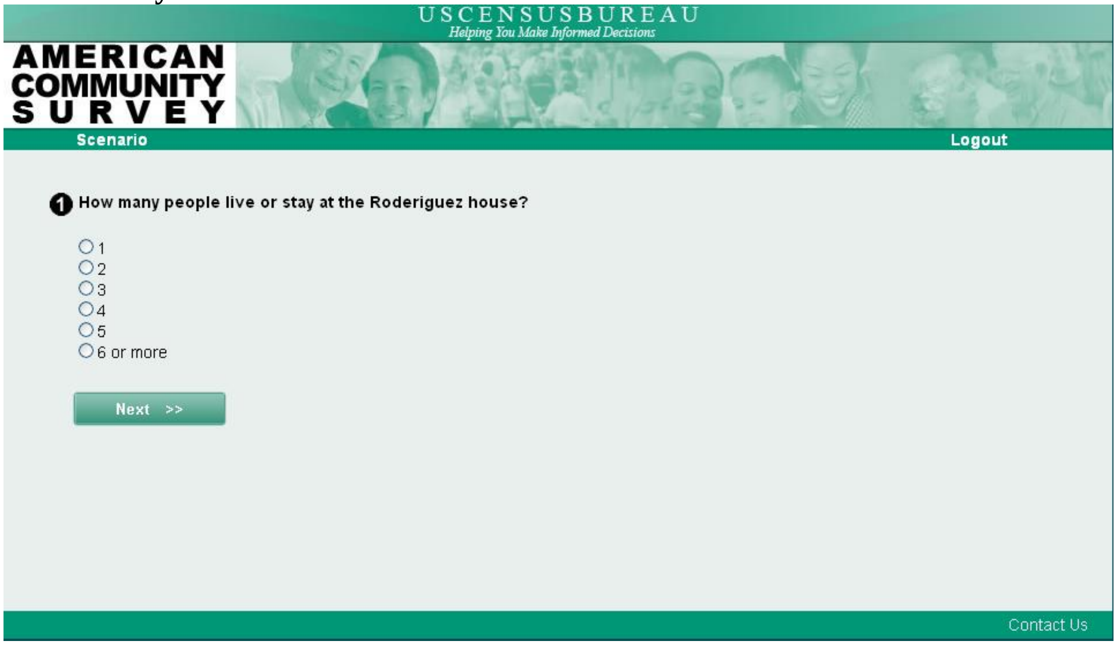
   □ Extremely difficult
   □ Very difficult
   □ Moderately difficult
   □ Not difficult at all

10. How difficult is it for you to use the Internet?

   □ Extremely difficult
   □ Very difficult
   □ Moderately difficult
   □ Not difficult at all

Appendix F. ACS Questions and Scenarios for Study 2

Live or Stay


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

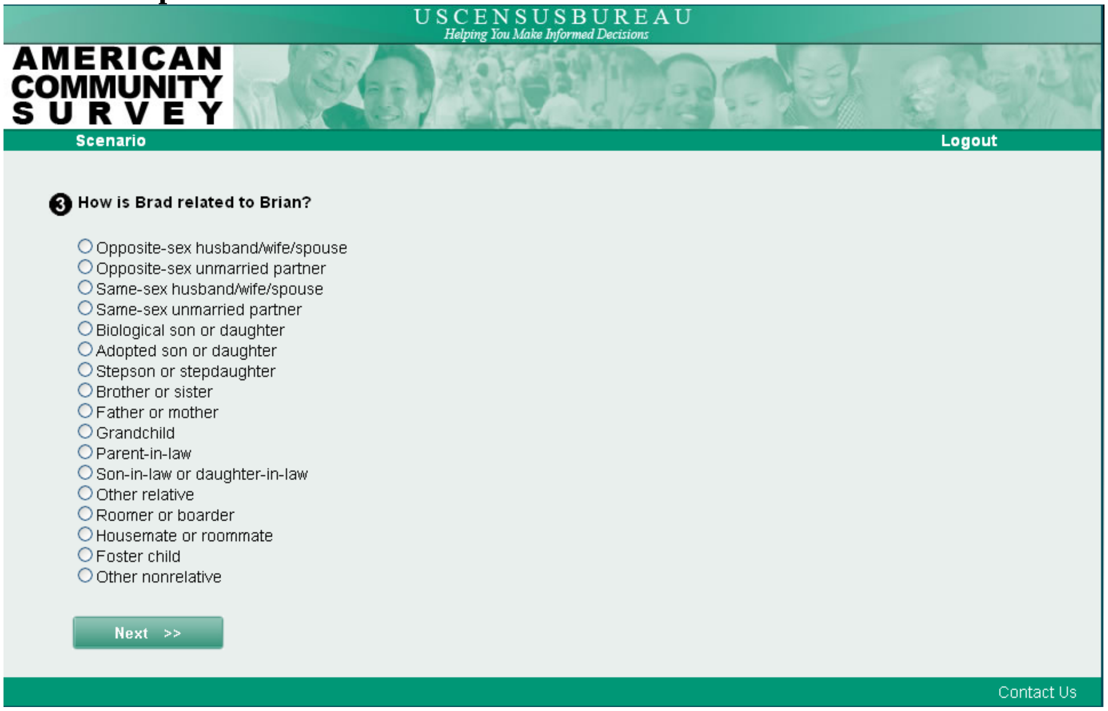
1 How many people live or stay at the Roderiguez house?

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

***Straightforward Scenario* –** Cassandra and Miguel Roderiguez just moved in to the 3-bedroom house at 4694 Main St. They previously lived in a larger house with their two children Inez and Jose. But since Inez and Jose moved out and started families of their own, Cassandra and Miguel wanted a smaller house because no one else lives or stays in the house with them.

***Complex Scenario* –** The Roderiguez family lives in the 3-bedroom house at 4694 Main St. The family has four members: Cassandra and Miguel Roderiguez, and their two children Inez and Jose. Cassandra and Miguel stay in the master bedroom, and Inez and Jose each have a room of their own. Inez always stays in her room, but Jose is a college student. Although his legal address is 4694 Main St., he lives at the college dorms all year, including summers, and only comes home for holidays and vacations.

Relationship


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

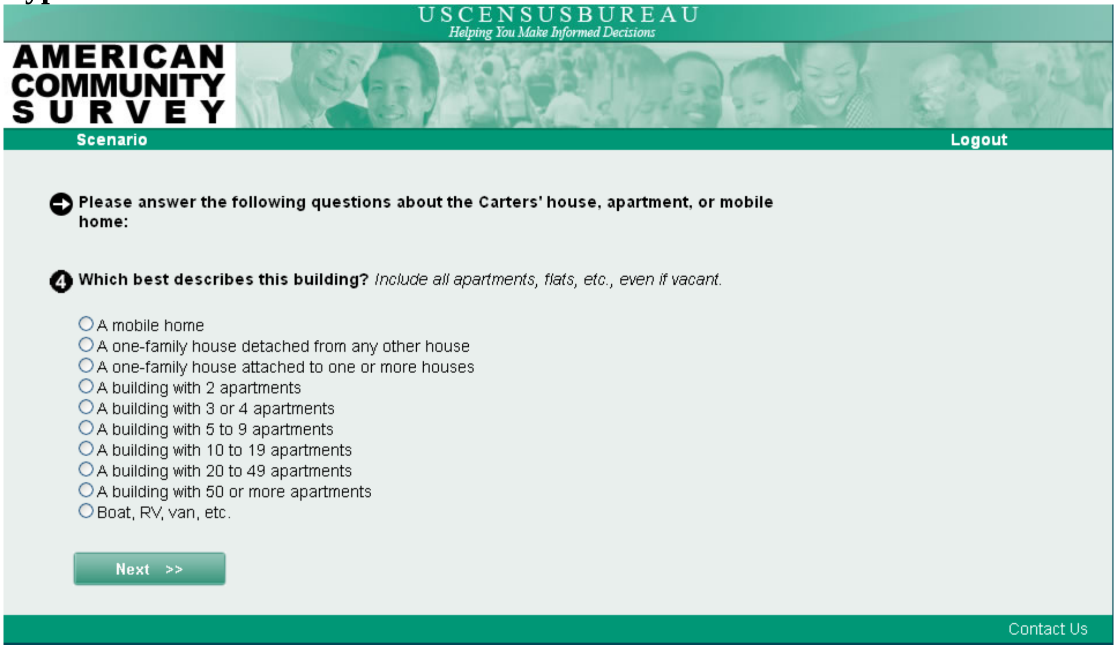
3 How is Brad related to Brian?

○ Opposite-sex husband/wife/spouse
○ Opposite-sex unmarried partner
○ Same-sex husband/wife/spouse
○ Same-sex unmarried partner
○ Biological son or daughter
○ Adopted son or daughter
○ Stepson or stepdaughter
○ Brother or sister
○ Father or mother
○ Grandchild
○ Parent-in-law
○ Son-in-law or daughter-in-law
○ Other relative
○ Roomer or boarder
○ Housemate or roommate
○ Foster child
○ Other nonrelative

Straightforward Scenario – Brian and Brad went to college together and after graduation they decided to live together in a rented 2-bedroom apartment. They share rent, are not related, and they are not in a relationship

Complex Scenario – Brian and Brad live together in a 2 bedroom house. Brian owns the house and Brad pays rent to Brian. Brad has his own room and access to certain parts of the house. They are not related.

Type of Unit



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

Please answer the following questions about the Carters' house, apartment, or mobile home:

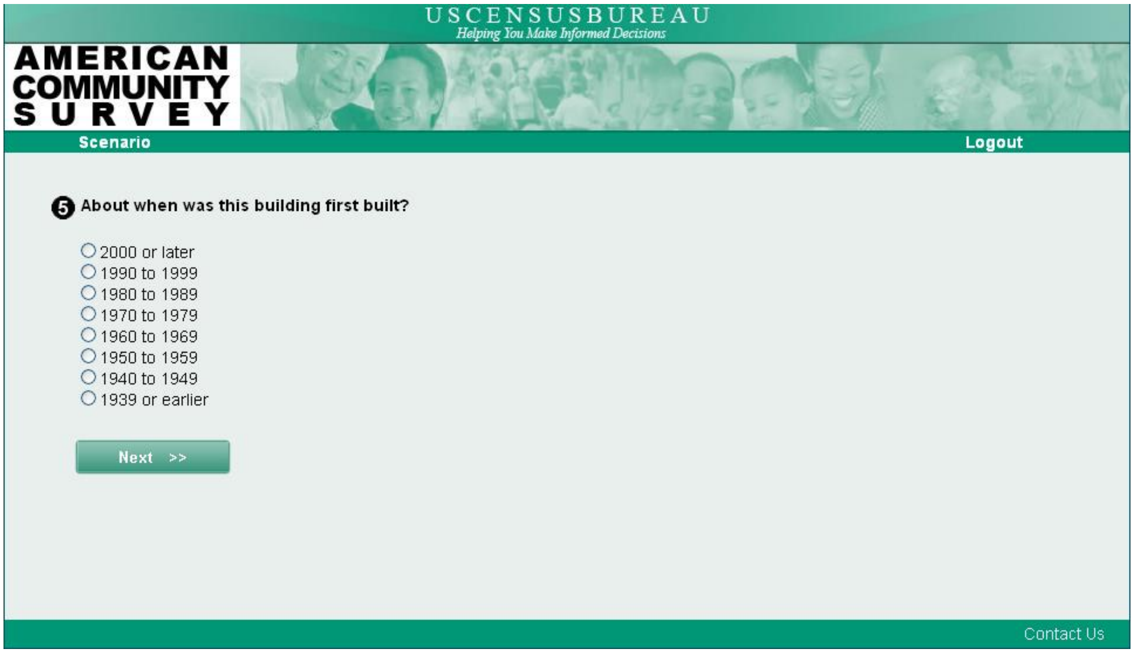
4 **Which best describes this building?** *Include all apartments, flats, etc., even if vacant.*

○ A mobile home
○ A one-family house detached from any other house
○ A one-family house attached to one or more houses
○ A building with 2 apartments
○ A building with 3 or 4 apartments
○ A building with 5 to 9 apartments
○ A building with 10 to 19 apartments
○ A building with 20 to 49 apartments
○ A building with 50 or more apartments
○ Boat, RV, van, etc.

***Straightforward Scenario* –** The Carter family lives in a stand-alone 4 bedroom house on 3 acres of land.

***Complex Scenario* –** The Carter family lives in a 3 story townhouse, attached to the houses on either side. The townhouse is broken into two condominiums, one on each floor. There is also a basement condominium that is not currently inhabited, so the Carters and the other families use it for storage.

Year Built


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

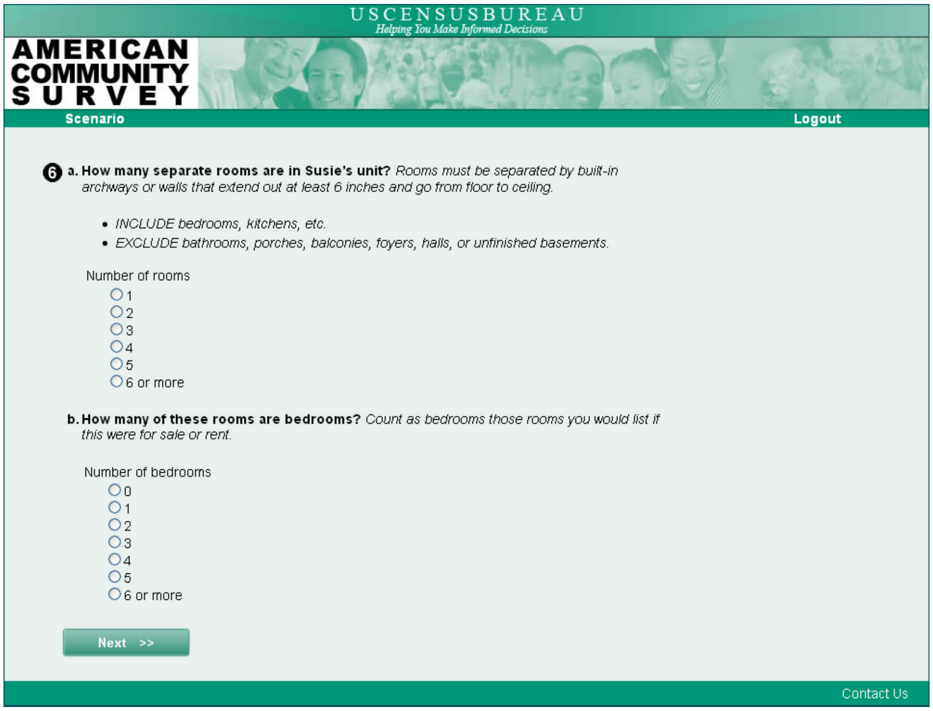
5 About when was this building first built?

○ 2000 or later
○ 1990 to 1999
○ 1980 to 1989
○ 1970 to 1979
○ 1960 to 1969
○ 1950 to 1959
○ 1940 to 1949
○ 1939 or earlier

***Straightforward Scenario* –** The Carters designed and began building their own home after they were married in 1973. It was finished in 1974.

***Complex Scenario* –** The Carters designed and began building their own home after they were married in 1998. It was finished in 2000.

Rooms


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

6 a. **How many separate rooms are in Susie's unit?** *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms

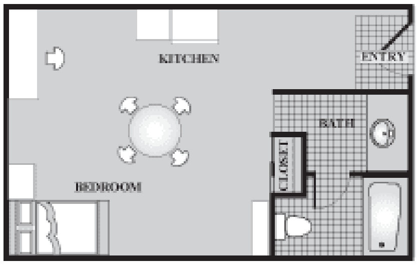
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

b. **How many of these rooms are bedrooms?** *Count as bedrooms those rooms you would list if this were for sale or rent.*

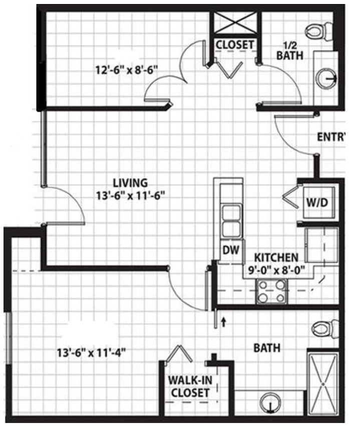
Number of bedrooms

○ 0
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Straightforward Scenario –

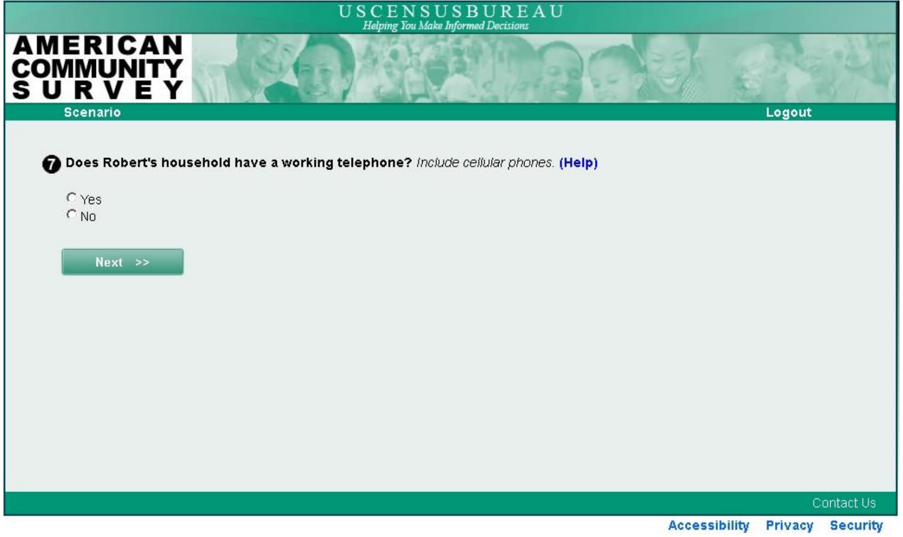


Complex Scenario –



Telephone

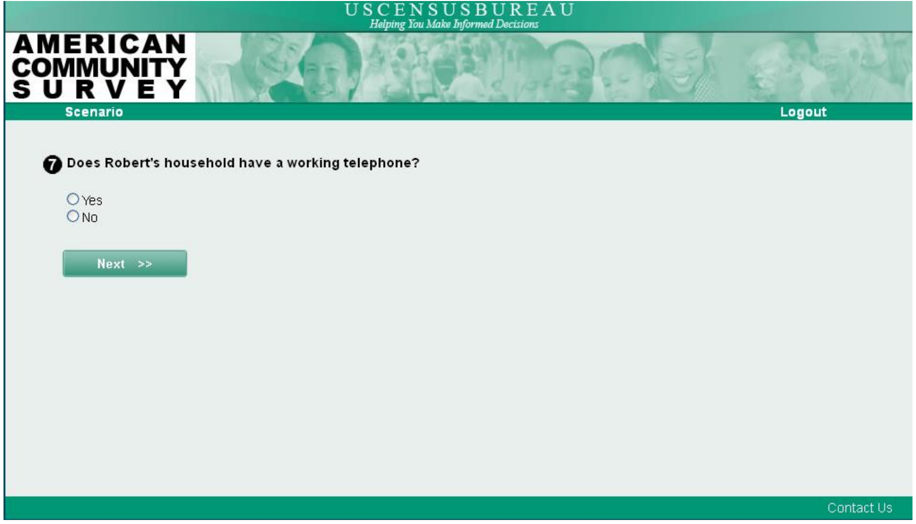
Straightforward – Provides instruction to include cellular phones



❼ **Does Robert's household have a working telephone?** *Include cellular phones.* **(Help)**

○ Yes
○ No

Complex – Instruction not included

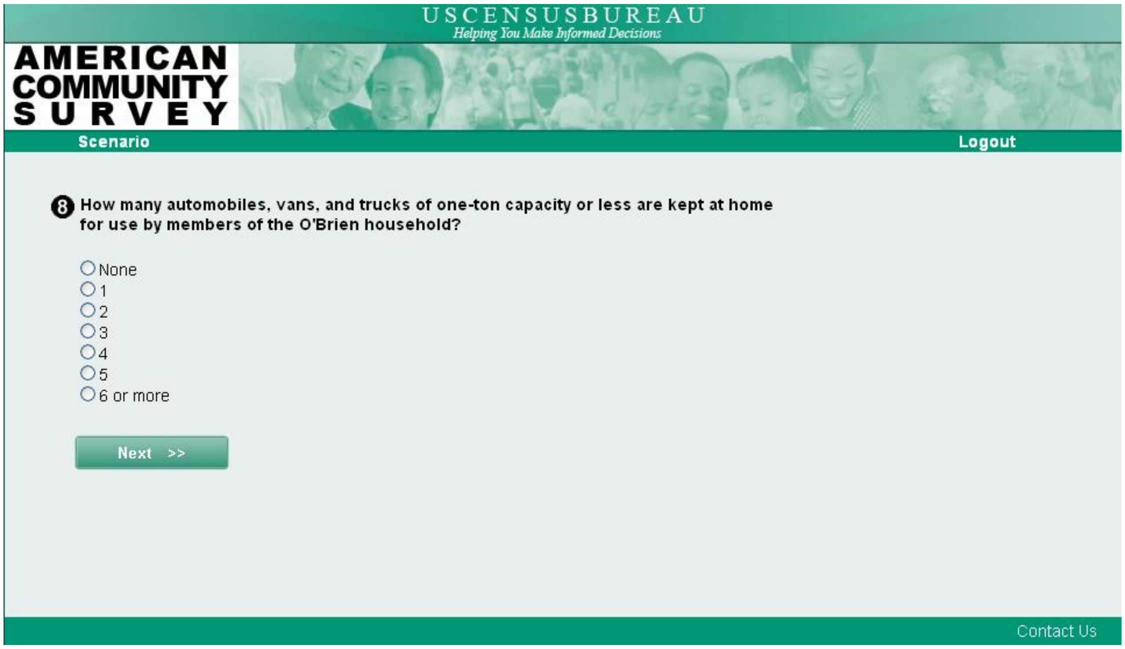

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

❼ **Does Robert's household have a working telephone?**

○ Yes
○ No

Scenario – Robert lives alone in Chicago, Il. He carries a cellular phone with him but does not have a landline at his apartment.

Vehicles



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

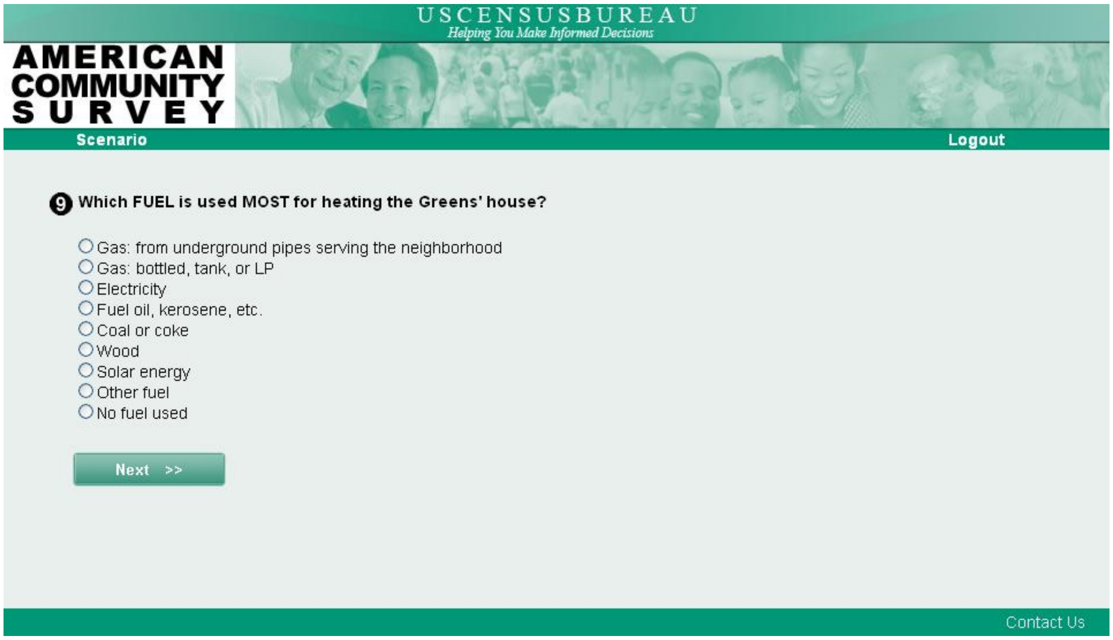
8 How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of the O'Brien household?

○ None
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Straightforward Scenario – The O'Brien family has 2 cars, which they use daily to drive to work. They keep the cars parked in their garage when they are not in use.

Complex Scenario – The O'Brien family has an RV that they use for vacations. They keep it parked in their driveway when it is not in use. They have no other vehicles.

Fuel


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

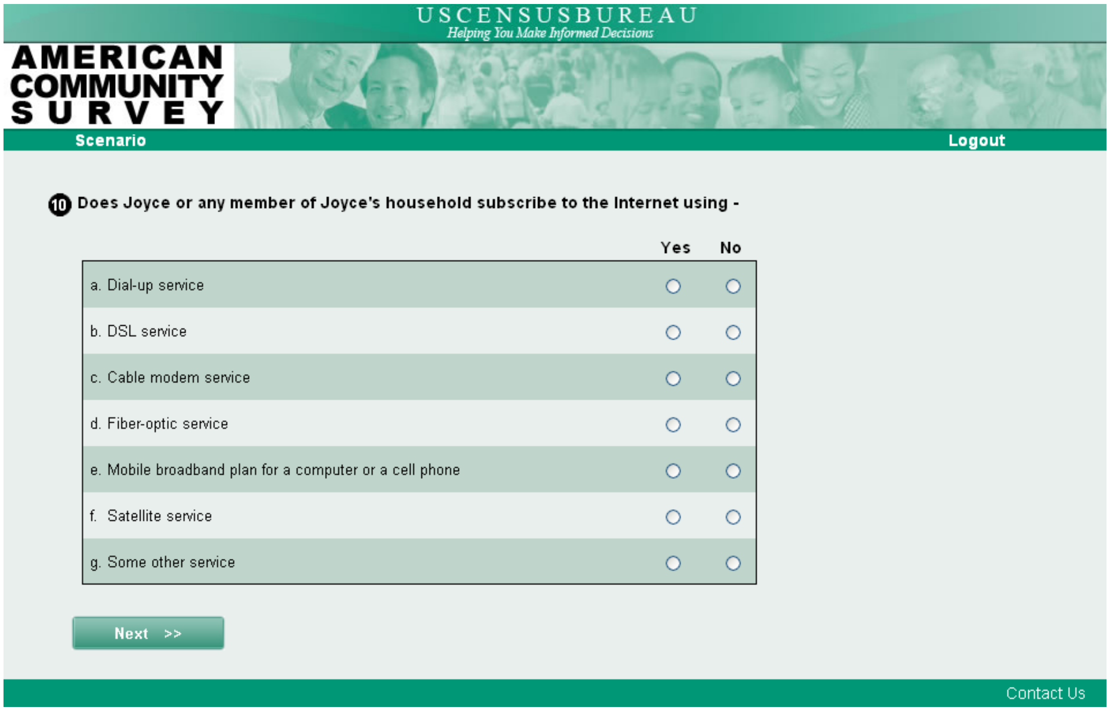
9 Which FUEL is used MOST for heating the Greens' house?

○ Gas: from underground pipes serving the neighborhood
○ Gas: bottled, tank, or LP
○ Electricity
○ Fuel oil, kerosene, etc.
○ Coal or coke
○ Wood
○ Solar energy
○ Other fuel
○ No fuel used

***Straightforward Scenario* –** The Greens live in Phoenix, Arizona. Since Phoenix is generally warm year round, central heating is unnecessary. However, when they need to heat their home they use an electric radiator.

***Complex Scenario* –** The Greens use a geothermal heat pump to heat their home. Geothermal heat pumps force the transfer of heat from the earth instead of the air because the earth temperature is more constant than the air temperature

Internet


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

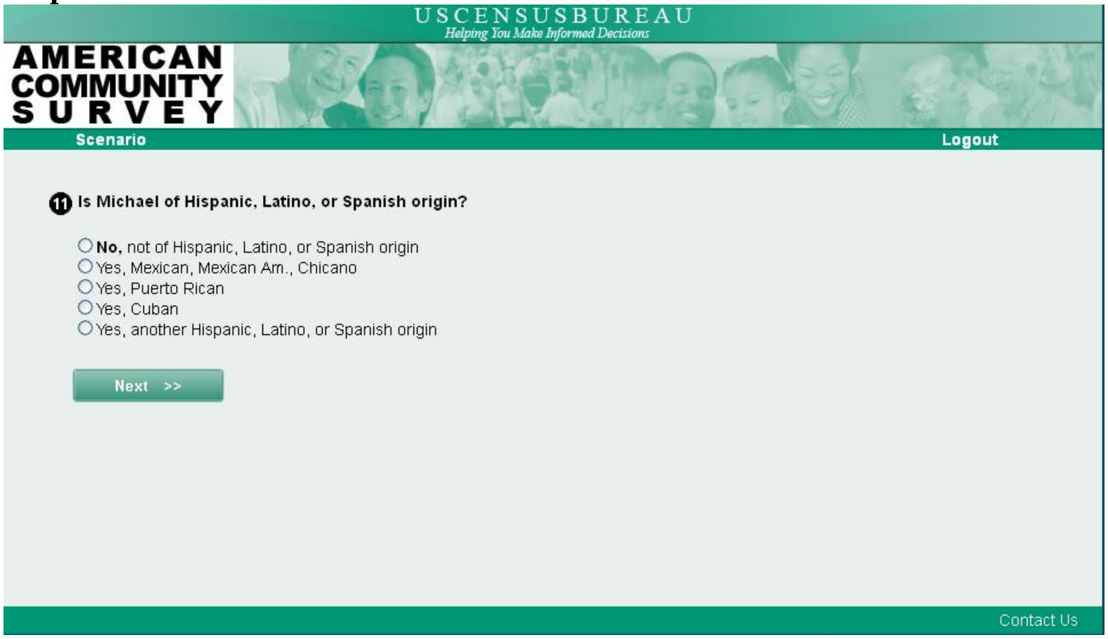
10 Does Joyce or any member of Joyce's household subscribe to the Internet using -

	Yes	No
a. Dial-up service	○	○
b. DSL service	○	○
c. Cable modem service	○	○
d. Fiber-optic service	○	○
e. Mobile broadband plan for a computer or a cell phone	○	○
f. Satellite service	○	○
g. Some other service	○	○

Straightforward Scenario – Joyce has a cable modem and pays her Internet bill to Comcast

Complex Scenario – Joyce has a package deal through Verizon for phone, high speed Internet, and cable television

Hispanic


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

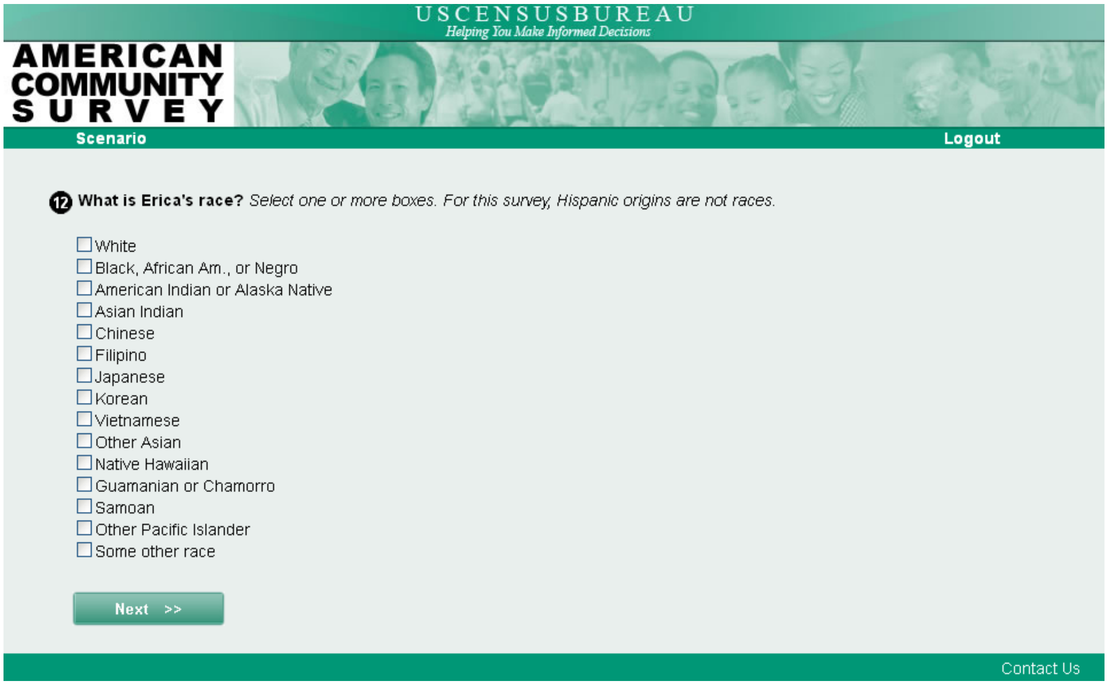
11 Is Michael of Hispanic, Latino, or Spanish origin?

○ **No,** not of Hispanic, Latino, or Spanish origin
○ Yes, Mexican, Mexican Am., Chicano
○ Yes, Puerto Rican
○ Yes, Cuban
○ Yes, another Hispanic, Latino, or Spanish origin

Straightforward Scenario – Both of Michael's biological parents were born and raised in Mexico and consider themselves to be Mexican. Michael was raised by his biological parents and also identifies as Mexican.

Complex Scenario – Michael's grandmother on his father's side was of Spanish descent. His mother's family is from Germany. Michael does not identify with his father's side of the family.

Race


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

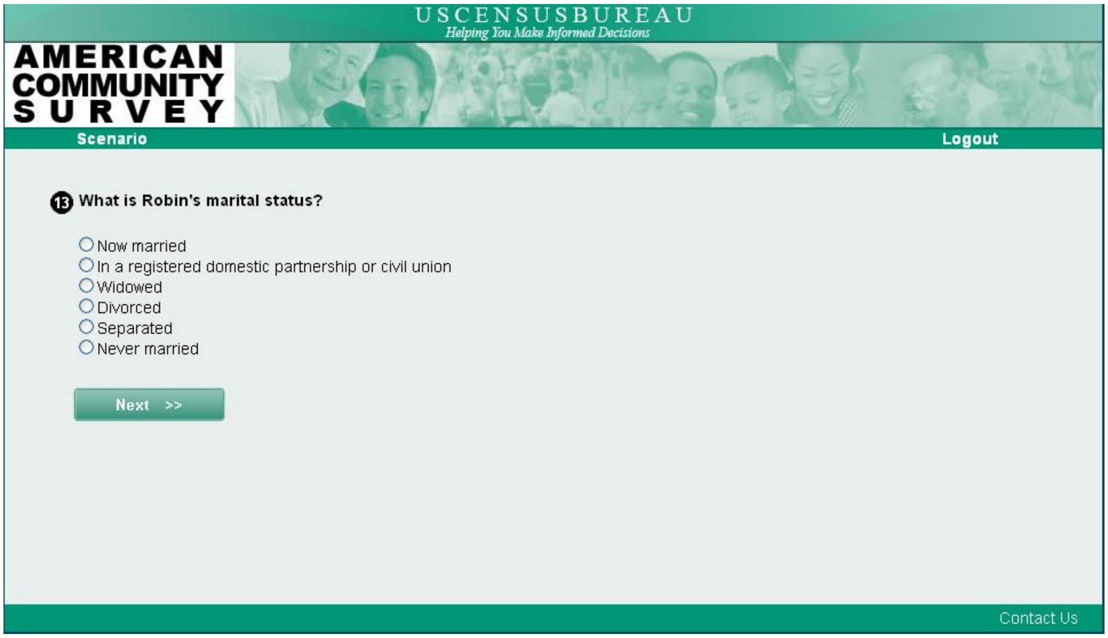
12 **What is Erica's race?** *Select one or more boxes. For this survey, Hispanic origins are not races.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native
□ Asian Indian
□ Chinese
□ Filipino
□ Japanese
□ Korean
□ Vietnamese
□ Other Asian
□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander
□ Some other race

Straightforward Scenario – Erica's biological mother is Caucasian, as is her biological father. She was raised by her biological parents. She also identifies as White.

Complex Scenario – Erica's biological mother was African American and her biological father was Caucasian. She was adopted by an African American family and identifies as African American.

Marital Status


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

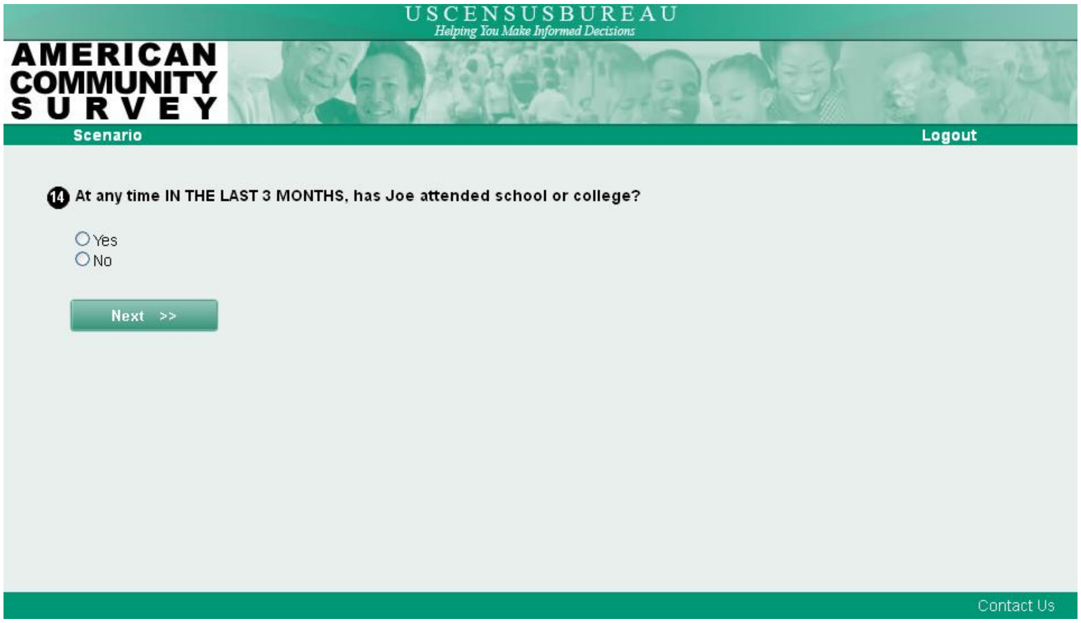
13 What is Robin's marital status?

○ Now married
○ In a registered domestic partnership or civil union
○ Widowed
○ Divorced
○ Separated
○ Never married

Straightforward Scenario – Robin was married for 10 years, but her husband died in a car accident two years ago. She has not remarried nor had any prior marriages.

Complex Scenario – Robin and her partner Jen have lived together for 10 years and had a wedding ceremony in Illinois last year. Illinois does not recognize same sex marriages. Neither Robin nor her partner has been previously married.

Attend School


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

14 At any time IN THE LAST 3 MONTHS, has Joe attended school or college?

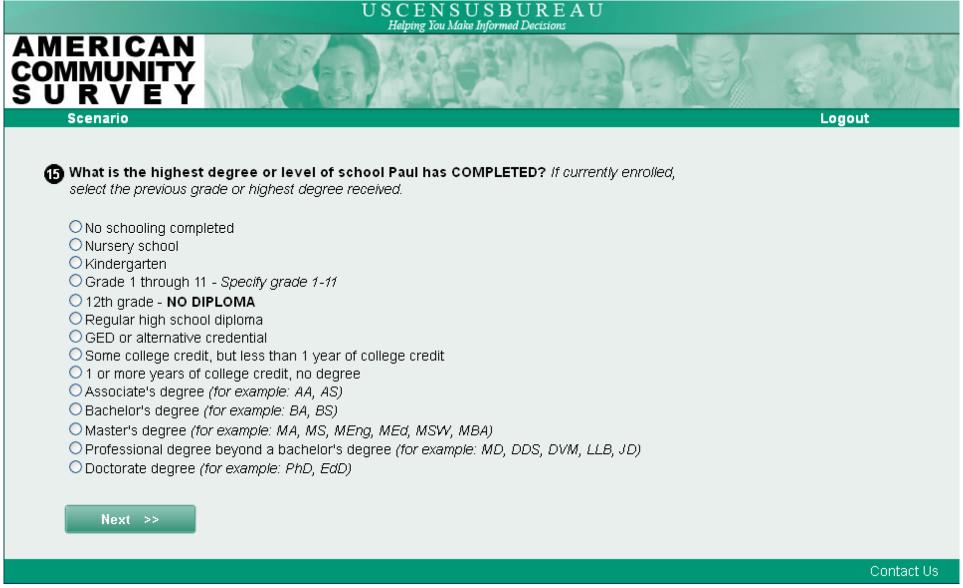
○ Yes
○ No

***Straightforward Scenario* –** Joe Smith is currently a junior at George Washington University, majoring in Business Administration. He is registered for five classes in the current semester and regularly attends all of them. Classes began in September and it is currently October.

***Complex Scenario* –** Joe Smith has a Master's degree in electrical engineering which he received in 2004. Last month, at the request of his employer, he attended a 2-day seminar in communication cabling which earns him continuing education credits.

Educational Attainment

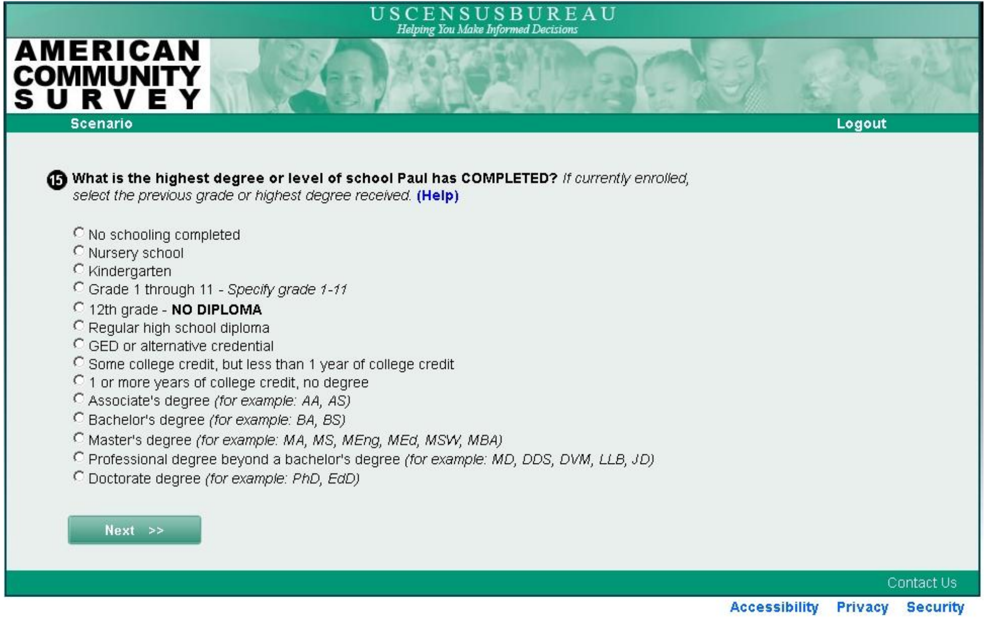
Straightforward – No labels


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

⓯ **What is the highest degree or level of school Paul has COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.*

○ No schooling completed
○ Nursery school
○ Kindergarten
○ Grade 1 through 11 - *Specify grade 1-11*
○ 12th grade - **NO DIPLOMA**
○ Regular high school diploma
○ GED or alternative credential
○ Some college credit, but less than 1 year of college credit
○ 1 or more years of college credit, no degree
○ Associate's degree *(for example: AA, AS)*
○ Bachelor's degree *(for example: BA, BS)*
○ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
○ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
○ Doctorate degree *(for example: PhD, EdD)*

Complex – Labels

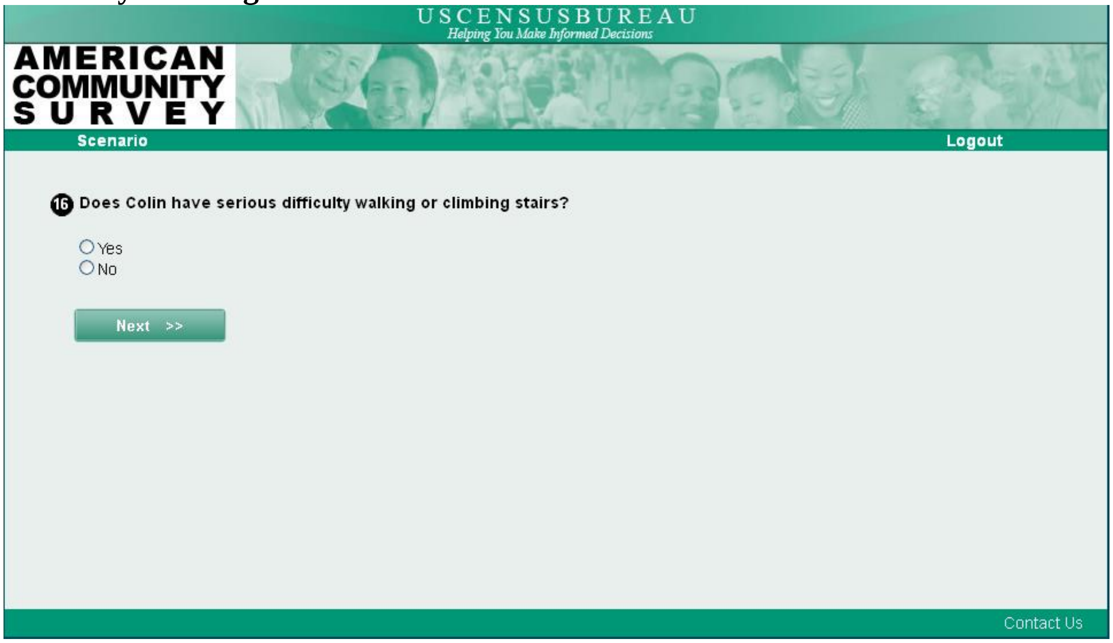

U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us
Accessibility
Privacy
Security

⓯ **What is the highest degree or level of school Paul has COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.* **(Help)**

○ No schooling completed
○ Nursery school
○ Kindergarten
○ Grade 1 through 11 - *Specify grade 1-11*
○ 12th grade - **NO DIPLOMA**
○ Regular high school diploma
○ GED or alternative credential
○ Some college credit, but less than 1 year of college credit
○ 1 or more years of college credit, no degree
○ Associate's degree *(for example: AA, AS)*
○ Bachelor's degree *(for example: BA, BS)*
○ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
○ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
○ Doctorate degree *(for example: PhD, EdD)*

Scenario **–** Dylan graduated from a 4-year university last year. He will be starting a Masters program next year.

Difficulty Walking


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

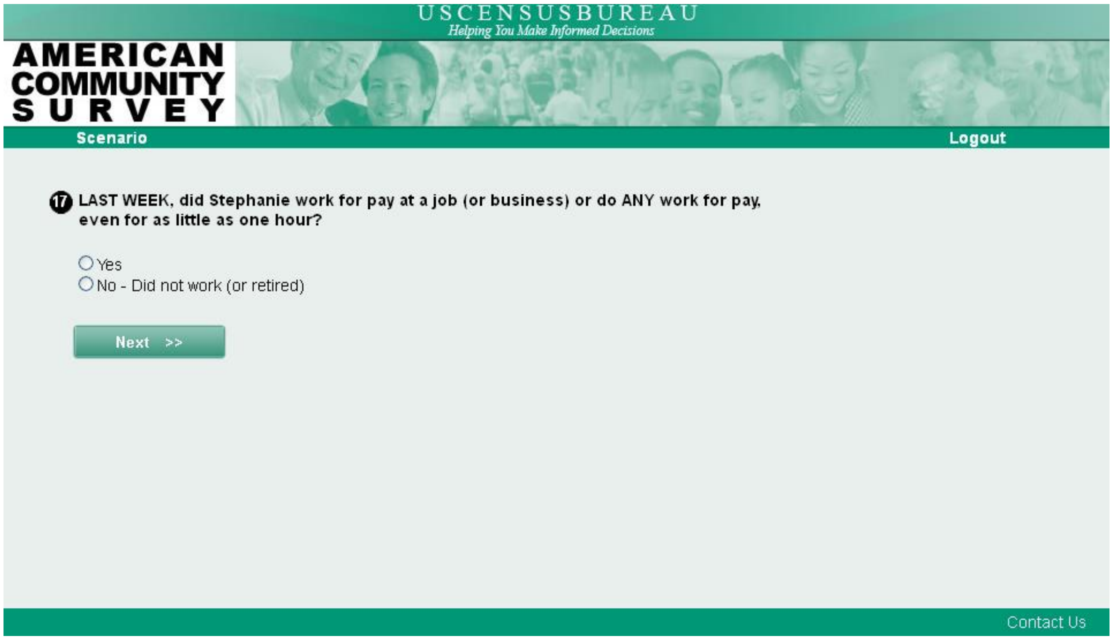
16 Does Colin have serious difficulty walking or climbing stairs?

○ Yes
○ No

Straightforward Scenario – Colin is a healthy, active young adult. For exercise, he walks 2 miles every day and also jogs up and down the bleachers at the local high school.

Complex Scenario – Colin sprained his ankle during a soccer game. He will be on crutches for the next week.

Work Last Week


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

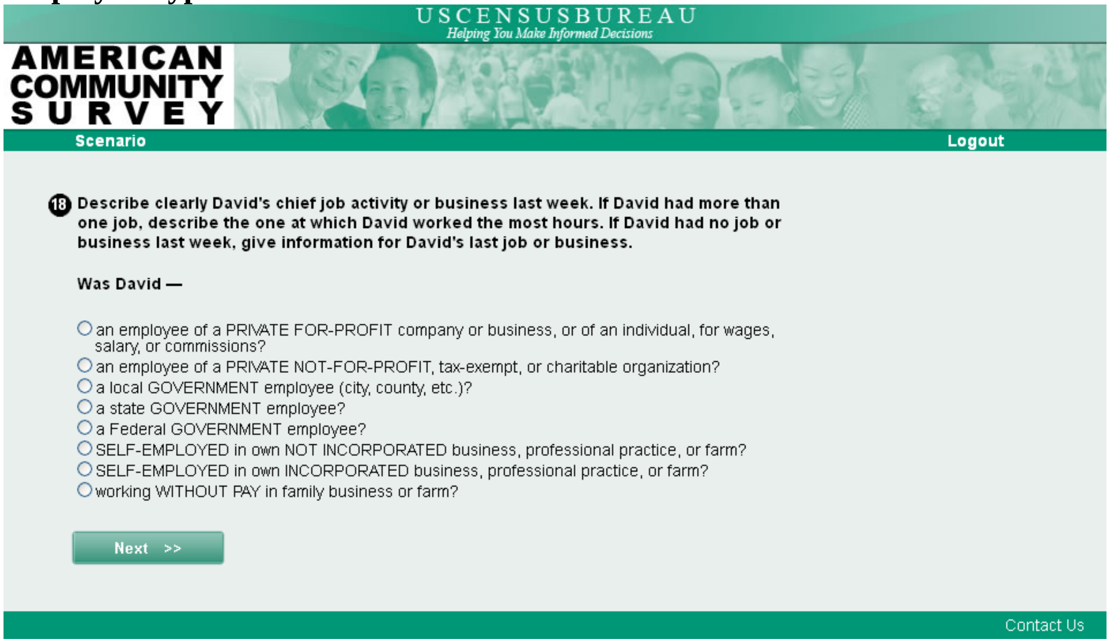
17 LAST WEEK, did Stephanie work for pay at a job (or business) or do ANY work for pay, even for as little as one hour?

○ Yes
○ No - Did not work (or retired)

Straightforward Scenario – Stephanie works as a high school teacher and is paid on an annual 52-week contract. The school year runs from September through June and consists of two 18-week semesters. It is currently October.

Complex Scenario – Stephanie works as a high school teacher and is paid on an annual 52-week contract. The school year runs from September through June and consists of two 18-week semesters. It is currently August.

Employee Type


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

18 Describe clearly David's chief job activity or business last week. If David had more than one job, describe the one at which David worked the most hours. If David had no job or business last week, give information for David's last job or business.

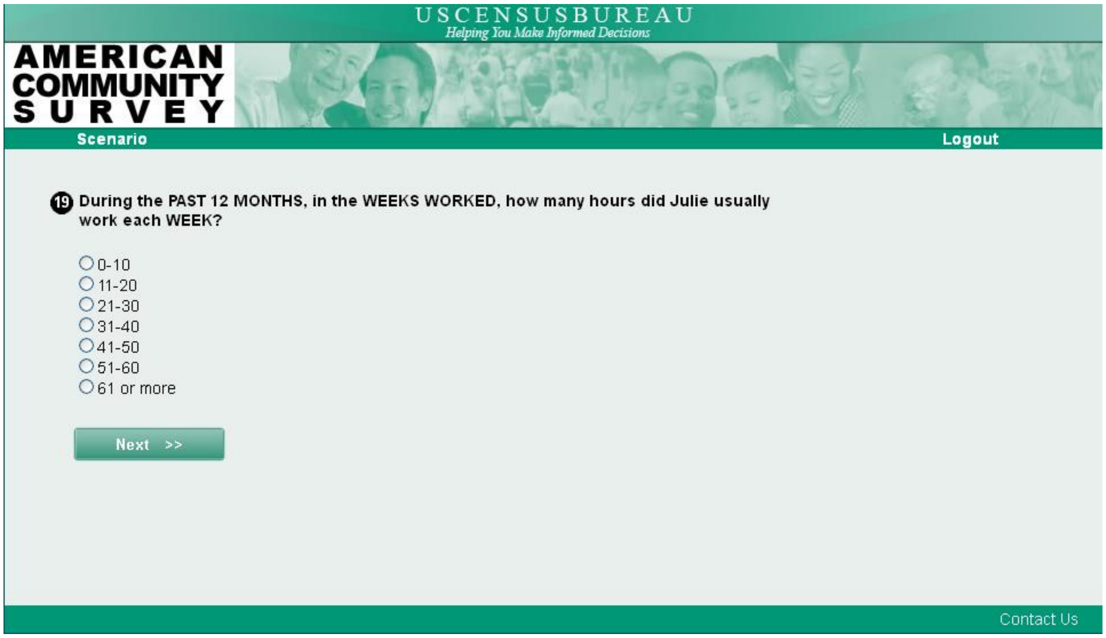
Was David —

○ an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
○ an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
○ a local GOVERNMENT employee (city, county, etc.)?
○ a state GOVERNMENT employee?
○ a Federal GOVERNMENT employee?
○ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
○ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
○ working WITHOUT PAY in family business or farm?

***Straightforward Scenario* –** David has been directly employed by the U.S. Department of Justice for the past 5 years in Washington, DC. David worked at the Department of Justice last week and has not worked at any other job.

***Complex Scenario* –** David worked as a government employee for the Department of Justice for 35 years before retiring in 2009. After he retired, he became an independent contractor and obtained a contract with his old agency, the Department of Justice. This is the only job he worked last week.

Hours Worked


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

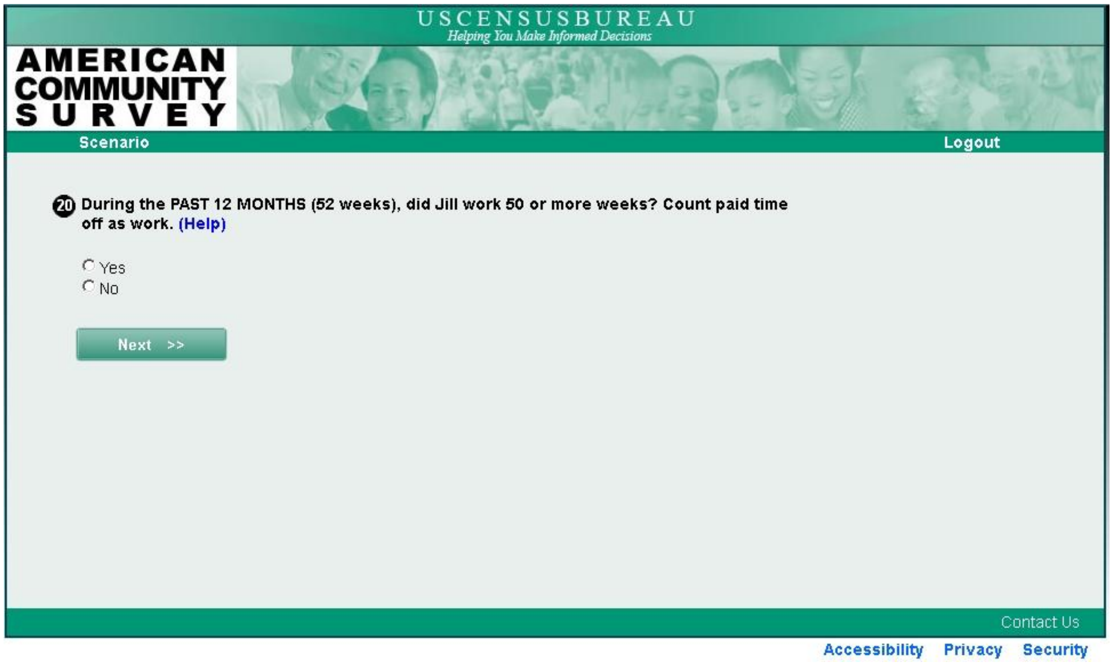
19 During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did Julie usually work each WEEK?

○ 0-10
○ 11-20
○ 21-30
○ 31-40
○ 41-50
○ 51-60
○ 61 or more

Straightforward Scenario – For the past year, Julie has worked a regular 9 to 5 job with the Federal government, which works out to 40 hours per week. She never works overtime, has not taken sick leave, and has not been on vacation this year. She does not have another job.

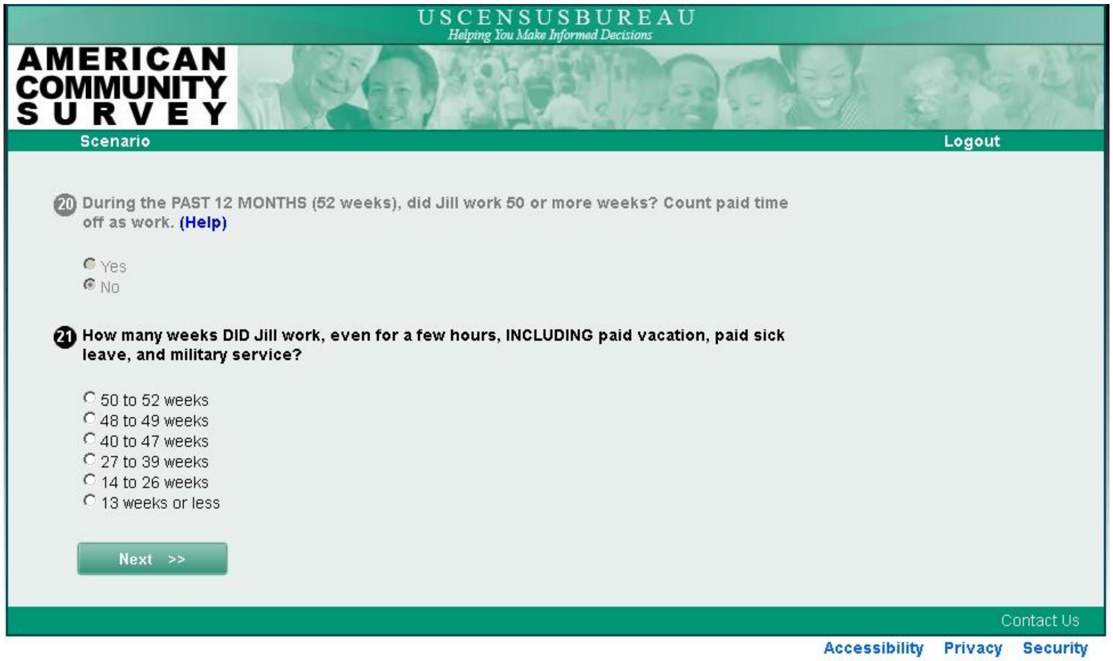
Complex Scenario – For the past year, Julie has spent 40 hours per week in the office at her government job. In the last six months Julie has been trying to start her own business in addition to her government job. So when she gets home, she spends 2 hours each weeknight working on starting her own business.

Weeks Worked


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

20 During the PAST 12 MONTHS (52 weeks), did Jill work 50 or more weeks? Count paid time off as work. (Help)

○ Yes
○ No

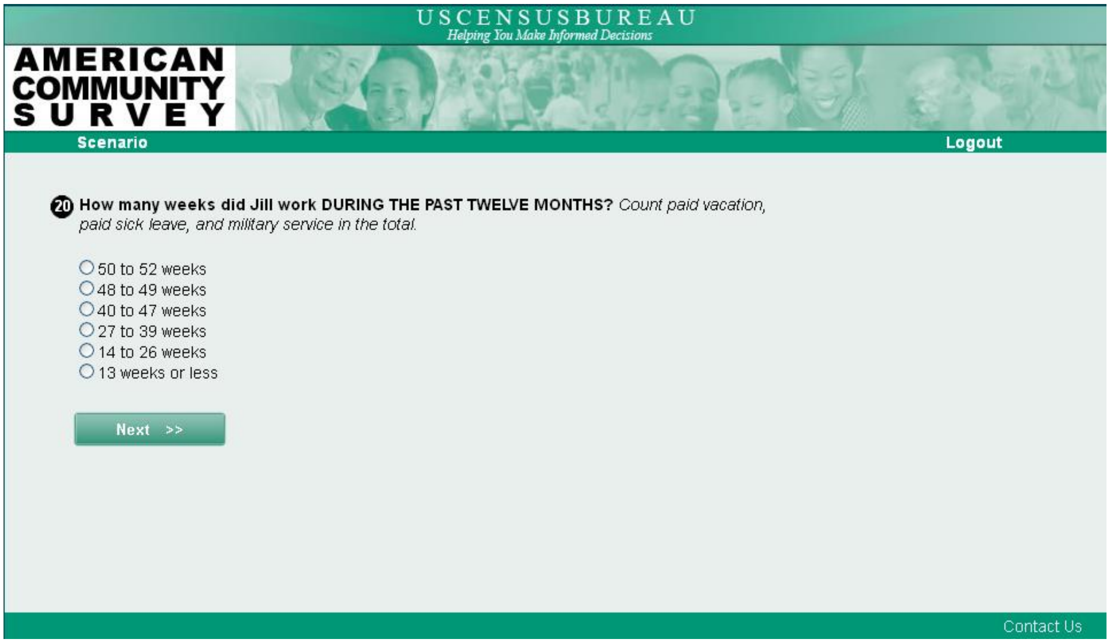

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

20 During the PAST 12 MONTHS (52 weeks), did Jill work 50 or more weeks? Count paid time off as work. (Help)

○ Yes
● No

21 How many weeks DID Jill work, even for a few hours, INCLUDING paid vacation, paid sick leave, and military service?

○ 50 to 52 weeks
○ 48 to 49 weeks
○ 40 to 47 weeks
○ 27 to 39 weeks
○ 14 to 26 weeks
○ 13 weeks or less

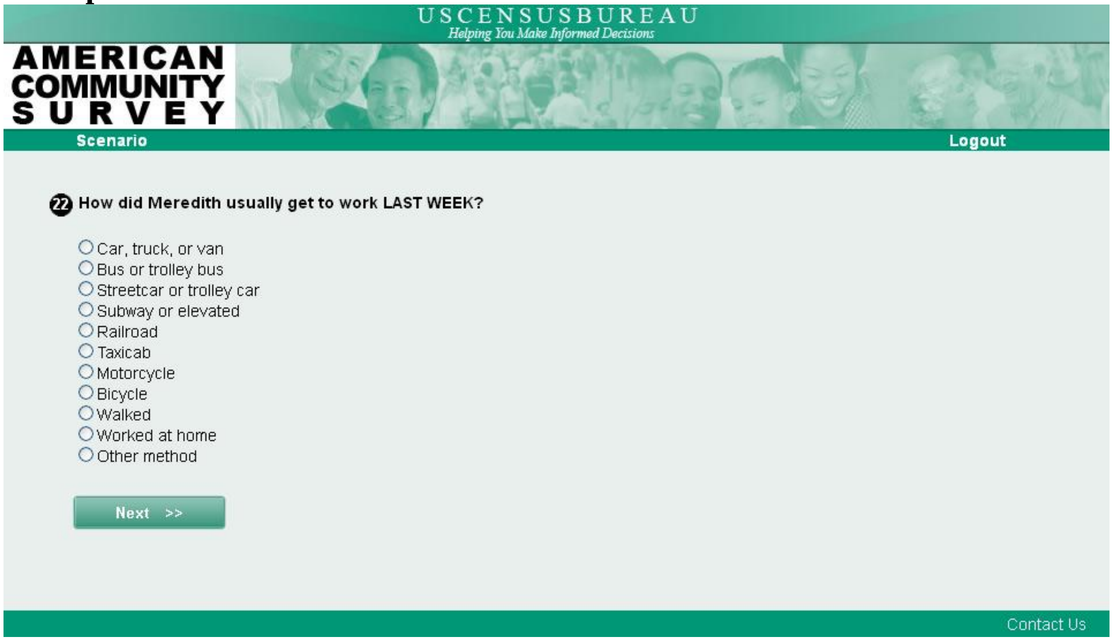

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

⓴ **How many weeks did Jill work DURING THE PAST TWELVE MONTHS?** *Count paid vacation, paid sick leave, and military service in the total.*

○ 50 to 52 weeks
○ 48 to 49 weeks
○ 40 to 47 weeks
○ 27 to 39 weeks
○ 14 to 26 weeks
○ 13 weeks or less

***Scenario* –** Jill worked most of the year as a librarian but was laid off for the last three weeks of the year and did not find other employment.

Transport to Work



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

22 How did Meredith usually get to work LAST WEEK?

○ Car, truck, or van
○ Bus or trolley bus
○ Streetcar or trolley car
○ Subway or elevated
○ Railroad
○ Taxicab
○ Motorcycle
○ Bicycle
○ Walked
○ Worked at home
○ Other method

***Straightforward Scenario* –** Meredith lives over an hour away from her company's office and there is no convenient public transportation near her home. Therefore, she has to drive to work every day.

***Complex Scenario* –** Meredith does sales for a pharmaceutical company. There is no corporate office nearby, so her office is at her home. She makes daily sales visits to doctors' offices around the metro area. She drives a car to these appointments.

Appendix G. Protocol for Study 2

Date______________________; Participant #________; Experimenter:_______________

General Introduction: Measuring Question Difficulty on the American Community Survey Internet Instrument

Thank you for your time today. My name is Rachel Horwitz and I am a student here in the Joint Program of Survey Methodology and am also employed at the US Census Bureau in the American Community Survey Data Collection area. I will be working with you today. If you have a cell phone, please turn it off or put it in vibrate. In order to help us improve our surveys, we turn to people like you to find out if our questions make sense and are fairly easy to understand and answer. We have found that the best way to do that is to actually conduct the survey with people and see how it works for them. So you will be helping us test a questionnaire from one of our surveys. Since it is difficult to recruit participants with a wide range of experience to come in to take these studies, we will be asking you to read a scenario for each question and to answer the question based on the information you have read, not your own personal experience. I did not create the survey, so please share both your positive and negative reactions to it. The entire session should last 30-45 minutes. Your comments and feedback will be given to the developers of the survey and may be used to improve it.

First, I would like to ask you to read and sign this consent form. It explains the purpose of today's session and informs you of your rights as a participant. It also tells you that we would like to record the session, with your permission. Only those of us connected with the project will review the recording and any other data collected during the session, and it will be used solely for research purposes. We may also use clips from the recording to illustrate key points about the survey to the Web design team.

Hand the participant the consent form; give time to read and sign; sign own name and date if you have not already done so.

Start the tape.
While you are completing the survey, we will record the movements of your eyes with our eye-tracking monitor to get a record of where you are looking on the screen and we will record your mouse movements to see how you are interacting with the survey.

Now I am going to calibrate your eyes for the eye-tracking.

Do Calibration
Now that we have your eyes calibrated, we are ready to begin. Please respond to the survey questions online. Please read each scenario carefully and then answer the questions as they apply to the scenario. To view the scenario, click the "Scenario" link at the top of the question page (*show screenshot of where the link is located*).

Do you have any questions?

Start the eye-tracking software: Tobii Studio. The survey opens in Internet Explorer when Tobii is started. Tell the respondent to answer as if they were home alone. Leave room.

Overall Probe: Make a note if a person left a page with a blank answer, asked a question to the researcher, or displayed signs of confusion (hovers, regressions, using the mouse as a marker).

What was your overall impression of the survey?

Were there any questions you found to be difficult or challenging to answer?

> If yes, show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.

Were there any responses you were unsure of?

> If yes, show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.

Were there any questions for which you debated between two or more response options?

> If yes, show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.
>
> How did you decide on an answer?

If respondents reported being confused by a question or displayed signs of confusion, ask them

> What was it about this question that you had trouble with (understanding the question, understanding the response options, applying their situation to one of the response options)?
>
> How did you come up with your answer?

For any questions the respondent displayed signs of confusion, ask if they had any trouble answering the question and what type of trouble they had. If they were not sure which answer category to select, ask them to describe their situation.

For any questions that seemed straightforward but the respondent displayed signs of confusion, ask what they were thinking about while answering the question

Retrospective Think-Aloud – pull up video of the respondent answering any questions on which they displayed signs of confusion.

I would like you to watch this recording of you answering a question and tell me what you were thinking about.

Thank you again for your participation today. It is greatly appreciated.

Appendix H. Pretest Results for Study 2

Prior to beginning data collection for the second study, we wanted to make sure the complex scenarios actually made the questions more difficult to answer than the straightforward scenarios. Therefore, we conducted three rounds of testing where we asked participants to rate the difficulty of each question/scenario combination on a five-point scale. We then compared the ratings provided for the straightforward scenario to the complex scenario for each question. Table H.1 shows this difference for each question, along with the associated p-value. After all the rounds of testing, there was sufficient differentiation in scenario versions for all questions other than 'Attend school.' However, we ran out of time to retest this question.

Table H.1. Pretest Results Comparing Scenario Versions

Question	Difference in Ratings Between Straightforward and Complex Scenarios	P-value
Live or Stay	1.36	0.0162
Relationship	1.49	0.0022
Type of Unit	1.51	0.0029
Year built	1.03	0.0468
Rooms	1.38	0.0131
Vehicles	1.85	0.0051
Fuel	2.41	0.0001
Internet	1.82	0.0067
Hispanic	1.58	0.0007
Race	1.24	0.0039
Marital status	1.55	0.0038
Attend school	0.67	0.1830
Difficulty walking	1.37	0.0116
Work last week	1.42	0.0115
Employee type	1.60	0.0005
Hours worked	1.75	0.0005
Weeks worked	1.52	0.0060
Transport to work	1.07	0.0134

Appendix I. Advertisements for Studies 2 and 3

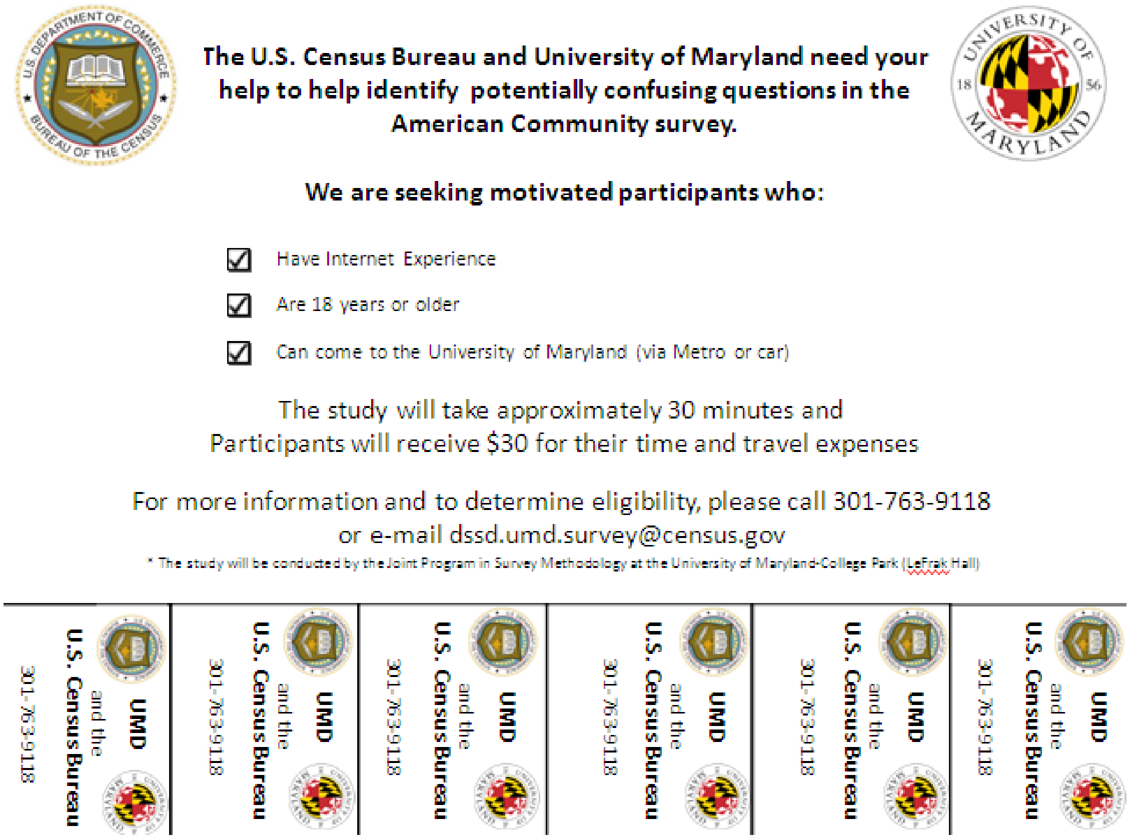
Craigslist Advertisement

The US Census Bureau and the Joint Program in Survey Methodology at the University of Maryland need your help in identifying potentially confusing questions on the American Community Survey. We are seeking participants who:

- Are 18 or older
- Have completed High School
- Have at least one year of Internet experience
- Can come to the University of Maryland (via metro or car)

The study involves answering 20 questions from the American Community Survey and rating their difficulty. It will take about 30 minutes to complete and eligible participants will be reimbursed $30 cash for time and travel. The study will be conducted at LeFrak Hall at the University of Maryland – College Park.
If interested, please respond to this posting, email rhorwitz@survey.umd.edu, or call 301-314-9916.

UMD Flyer Advertisement



The U.S. Census Bureau and University of Maryland need your help to help identify potentially confusing questions in the American Community survey.

We are seeking motivated participants who:

☑ Have Internet Experience

☑ Are 18 years or older

☑ Can come to the University of Maryland (via Metro or car)

The study will take approximately 30 minutes and
Participants will receive $30 for their time and travel expenses

For more information and to determine eligibility, please call 301-763-9118
or e-mail dssd.umd.survey@census.gov

* The study will be conducted by the Joint Program in Survey Methodology at the University of Maryland-College Park (LeFrak Hall)

UMD and the U.S. Census Bureau 301-763-9118

UMD and the U.S. Census Bureau 301-763-9118

UMD and the U.S. Census Bureau 301-763-9118

UMD and the U.S. Census Bureau 301-763-9118

UMD and the U.S. Census Bureau 301-763-9118

UMD and the U.S. Census Bureau 301-763-9118

Appendix J. Phone Screener for Studies 2 and 3

Phone Screener: Human Factors and Usability Research Group

Thanks for your interest in participating in a Research Study with the U.S. Census Bureau and the University of Maryland. Our study will take about between 45 minutes and one hour to complete and if you are eligible you will receive a $30 cash reimbursement for traveling expenses to the University of Maryland, which is located in College Park, MD. Prior to scheduling you to participate in this study, I will need to gather some information from you to determine whether you are eligible for this study. The information that you provide is confidential and you can decline to answer any questions if you do not want to provide the information at this time.

General

7. **Let's start with the spelling of your first and last name:**

   First name   Last name

8. **Are you male/female? (if necessary)**

   □ Male

   □ Female

9. **Have you participated in any research studies in the past year?**

   □ Yes

   How many? _____

   □ No

10. **Are you currently employed?**

    □ Yes

    □ No

Demographics

11. **What is your date of birth – just including the month and year?**

    MM   YYYY

12. **Choose one or more of the following races:**

    Note: If reluctant, inform them that this does not impact their ability to participate in our studies-we gather this information to ensure that we have a diverse group of participants for each of our studies

    □ White

    □ Black or African American

    □ Asian

    □ American Indian or Alaska Native

    □ Native Hawaiian or other Pacific Islander

13. **What is your highest level of education COMPLETED?**
□ Never completed any school
□ Completed ninth grade or below
□ Some high school
□ Completed high school, or received a GED
□ Vocational training beyond high school
□ Some college
□ Completed college
□ Any graduate or professional education

Computer/Internet Experience

14. **Do you have at least one year internet experience?**
□ Yes
□ No (*Inform them that most of our studies require at least one year internet experience and that we will keep them in mind for other studies that do not require internet experience.*) *SKIP to Contact Information*

15. **Name at least two things that you use the internet to do?**
b. ____________________
b.____________________

16. **Do you currently use the internet at least twice per week for searching for information, filling out forms, or shopping online?**
□ Yes
□ No

Contact Information

17. **Is there a daytime phone number where you can be reached? (Include preferred contact times if any)**
Home_______________
Mobile_______________
Work________________

18. **How did you hear about our study?**
□ Craigslist
□ Newspaper ad_____________
□ Flier posting____________
□ Referral
□ Other______________

19. **What city do you live in?**

Appendix K. Response Times

As past research has studied the relationship between response times and difficulty, we wanted to see if our hypotheses regarding mouse movements supported this relationship. In other words, are response times longer when more mouse movements are used and are they longer for the same movements that are associated with difficulty? Answering these questions will help determine whether mouse movements can provide more information than response times alone. At a high level, mean response times were significantly longer when there was at least one movement compared to no movements (22.82 seconds compared to 11.45 seconds). Table K.1 provides a comparison of the mean response time when at least one movement occurred on a question to when no movements occurred. We can see that movements resulted in longer response times across all movements. However, the differences are generally larger for movements we predicted may be related to difficulty, which could again suggest a relationship between difficulty and mouse movements.

Table K.1. Comparison of Response Times when no Movement Occurred to when at least one Movement Occurred.

Movement	**Resp time - no move (se)**	**Resp time – move (se)**	**Difference**
Response-to-Next	17.67 (.572)	29.97 (2.539)	12.30
Hover – Question	17.92 (.563)	34.54 (3.023)	16.62
Response-to-Response	17.65 (.582)	27.4 (1.683)	9.75
Marker	14.6 (.328)	27.06 (1.66)	12.46
Horizontal – Response	17.98 (.564)	32.68 (2.905)	14.70
Response-to-Question	17.98 (.559)	35.65 (4.753)	17.67
Hover – Next	17.38 (.623)	24.29 (.917)	6.91
Horizontal – Question	13.47 (1.197)	23.67 (1.566)	10.2
Vertical Reading	17.89 (.617)	21.44 (.866)	3.55

Appendix L. Sensitivity Analysis of Probing Questions

The difficulty analyses used in this dissertation defined a question as difficult if the probing question received a rating of ‘Very difficult,’ ‘Somewhat difficult,’ or ‘Neither easy nor difficult.’ The middle category, ‘Neither easy nor difficult’ was included because there was not much differentiation in participants’ responses to the probing questions and also because some participants indicated in the debriefings that they had trouble answering questions that they did not necessarily rate as difficult.

To check how susceptible our findings are to counting the middle category as difficult, we reran the analysis only including reports of ‘Somewhat difficult’ and ‘Very difficult.’ Table L.1 provides the results of this analysis for the final model. The results are similar to those found in the original analysis. All of the odds ratios change somewhat, but the only large change is to Hover – Question. Even where this difference exists, the results from this analysis yield the same conclusions as our original analysis.

Figure L.1 Odds Ratios from Final Regression Model using Recoded Measure of Difficulty

Fit Statistics		
Participants	100	
Observations	2037	
-2 Res Log Pseudo-Likelihood	10386.55	
Generalized Chi-Square	1635.73	
Generalized Chi-Square/DF	0.81	
Movement	**Odds Ratio Estimate**	**Confidence Interval**
Marker - Multiple vs Zero	5.93	3.76-9.35
Marker - One vs Zero	3.17	2.35-4.28
Response-to-Response	2.05	1.30-3.22
Response-to-Next	6.95	4.23-11.41
Response-to-Space	1.59	1.30-3.22
Hover – Question	4.80	2.26-10.20
Horizontal Reading – Question	1.75	1.12-2.72

Appendix M. ANOVA Test Results of Frequency of Mouse Movements across Demographic Characteristics

To determine whether any demographic groups were more likely to engage in each movement, we ran one-way ANOVAs to test for relationships between each movement and gender (Male, Female), age (18-25, 26-35, 36-45, 46-55, 55+), education (high school, some college Associates, Bachelors, some graduate study, Post graduate degree), and race (Asian, Black, White, other). We first determined whether there were any differences in how frequently participants in each demographic subgroup engaged in each movement (Table L.1).

Table M.1 ANOVA Test Results of Frequency of Mouse Movements across Demographic Characteristics

	Gender	Age	Education	Race
Movement	**(DF =1, 2049))**	**(DF = 4, 2046)**	**(DF = 5, 2045)**	**(DF = 3,2047)**
Horizontal reading - Response	F=4.68, p=.0306*	F=1.96, p=.0982^{+}	F=1.15, p=.3313	F=1.24, p=.2922
Horizontal reading - Question	F=20.23, p<.0001*	F=16.33, p<.0001*	F=6.07, p<.0001*	F=2.45, p=.0615^{+}
Hover – Next	F=15.53, p<.0001*	F=3.30, p=.0105*	F=4.43, p=.0005*	F=.18, p=.9132
Hover – Question	F=1.49, p=.2229	F=2.06, p=.0841^{+}	F=2.28, p=.0445*	F=1.0, p=.3921
Marker	F=.14, p=.7037	F=1.69, p=.1503	F=.85, p=.5168	F=1.57, p=.1947
Vertical reading	F=.01, p=.9029	F=4.16, p=.0020*	F=1.46, p=.2000	F=.39, p=.7587
Response-to-Response	F=.86, p=.3535	F=3.84, p=.0041*	F=1.73, p=.1236	F=.43, p=.7311
Response-to-Next	F=.07, p=.7979	F=4.62, p=.0010*	F=1.60, p=.1579	F=.25, p=.8641
Response-to-Space	F=.81, p=.3675	F=2.35, p=.0525^{+}	F=2.08, p=.0654^{+}	F=.19, p=.9055
Response-to-Question	F=.19, p=.6628	F=2.29, p=.0577^{+}	F=1.03, p=.4004	F=.62, p=.6012

*Significant at the α=0.05 level

$^{+}$Significant at the α=0.10 level

From Table M.1, we identified the questions and demographic characteristics where there were some differences between the subgroups. For each of these movement/characteristic combinations, we used a Newman Keul test to determine whether which specific subgroups moved the mouse statistically more than other groups. Unfortunately, for most characteristics we were unable to uncover a meaningful pattern.

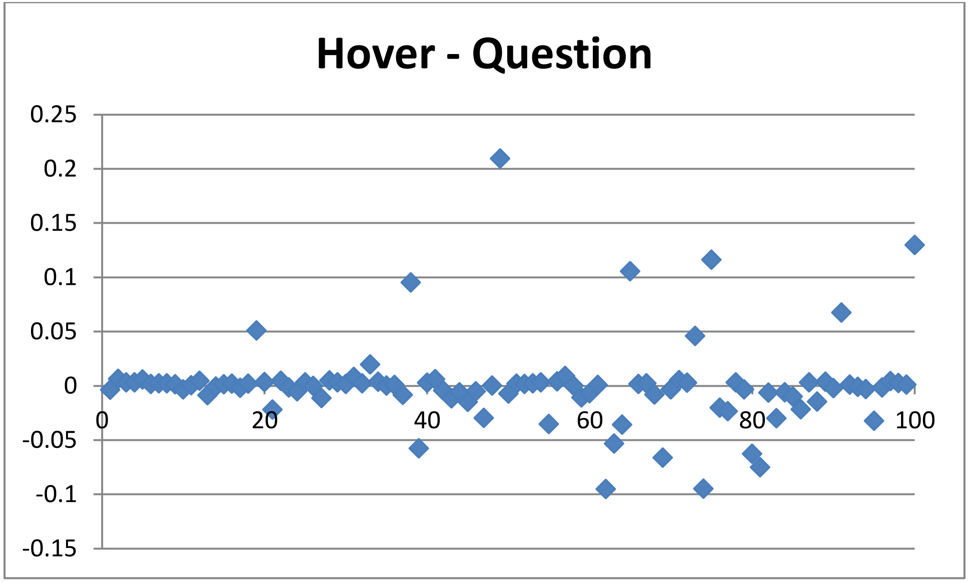
For the questions where there were differences, we found that males moved the mouse significantly more than females, but there is no obvious reason why this would be the case. Similarly, we found no support that participants with more education moved the mouse differently than participants with less education. For example, participants that had a post graduate degree engaged in Horizontal reading – Question more frequently than participants who had completed some college, but participants who completed high school also moved the mouse more frequently than participants who had completed some college. Therefore, it does not appear that we can make generalizations relating movements to different levels of education.

For age, we found that participants between the age of 26 and 35 moved the mouse more frequently than many of the other age groups, across all movements where there were differences. Additionally, for three of the five movements, the 18-25 age-group also moved the mouse more frequently than 46 to 55 year olds. Although the younger age groups did not move the mouse significantly more often compared to all of the older age groups for each of the movements, they lower age groups were almost always nominally higher, providing some support that younger respondents engage in these movements more often than older respondents.

Appendix N. Identifying Influential Participants – Differences between Final Parameter Estimate and Estimates after Removing each Participant for Study 2

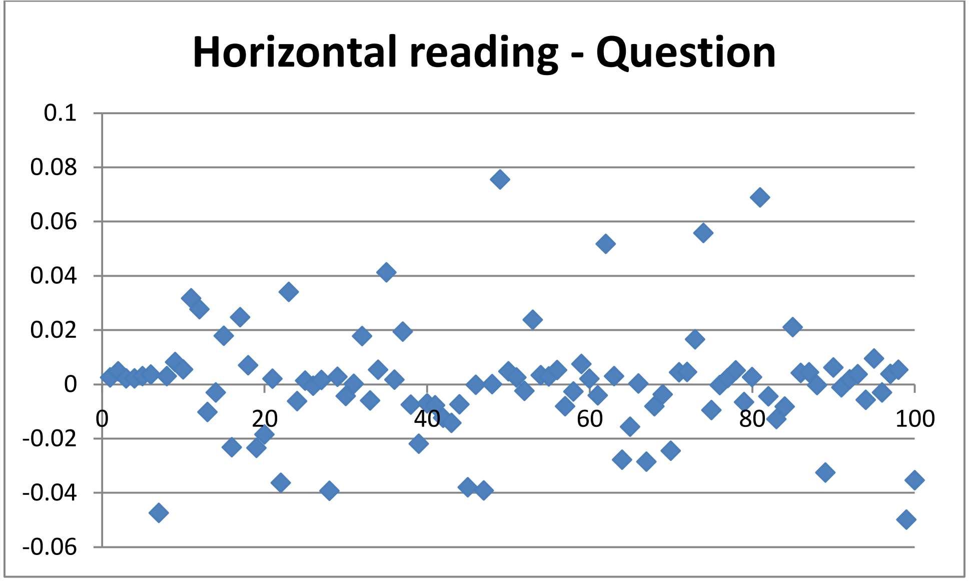
There was some concern in the second study that if a handful of participants engaged in a movement far more often than other participants, they may drive the results of the analysis and we may find significant findings where none truly exits. Therefore, we reran the final model after systematically removing each participant one at a time. The figures in this appendix provide a plot of these differences across the 100 participants for each movement (horizontal axis). The vertical axis provides the difference in parameter estimate after removing each participant. Therefore, a value of zero means there the parameter estimate did not change after removing a participant from the analysis. We did not find any outliers[31] that suggest a single participant was driving the results.

Hover – Question

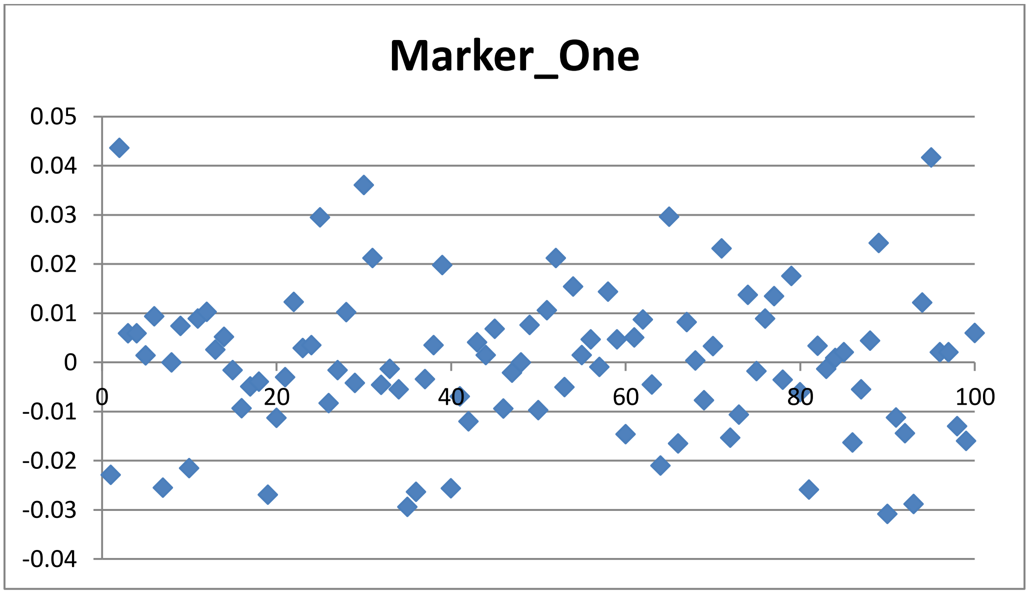


[31] For Hover – Question there is one difference after removing the participant at the 0.2 mark that could be considered an outlier. However, removing this participant did not have an effect on the data, so they were kept in the analysis.

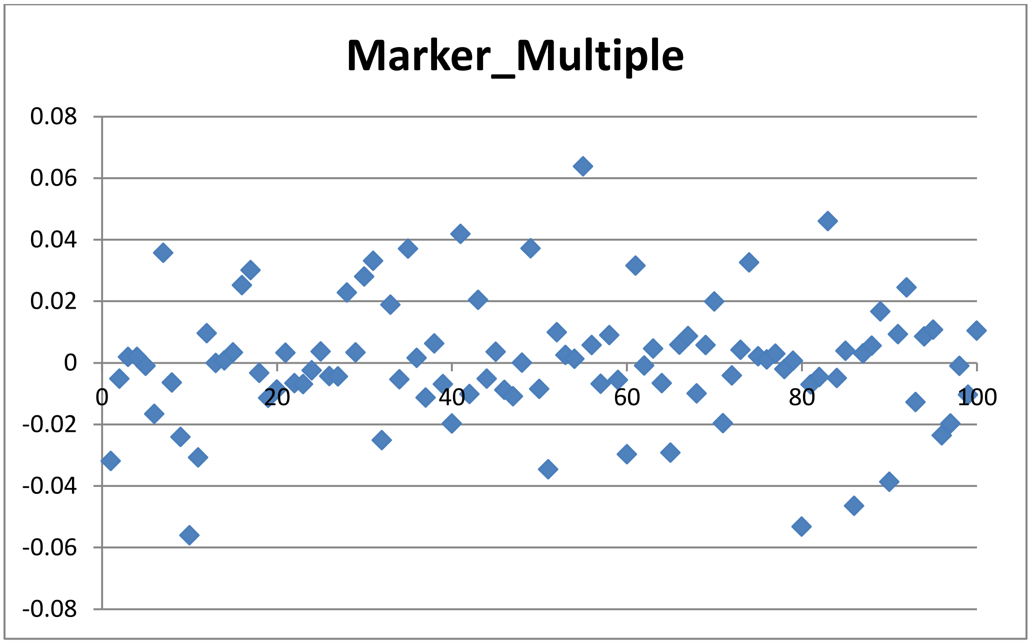
Horizontal reading – Question



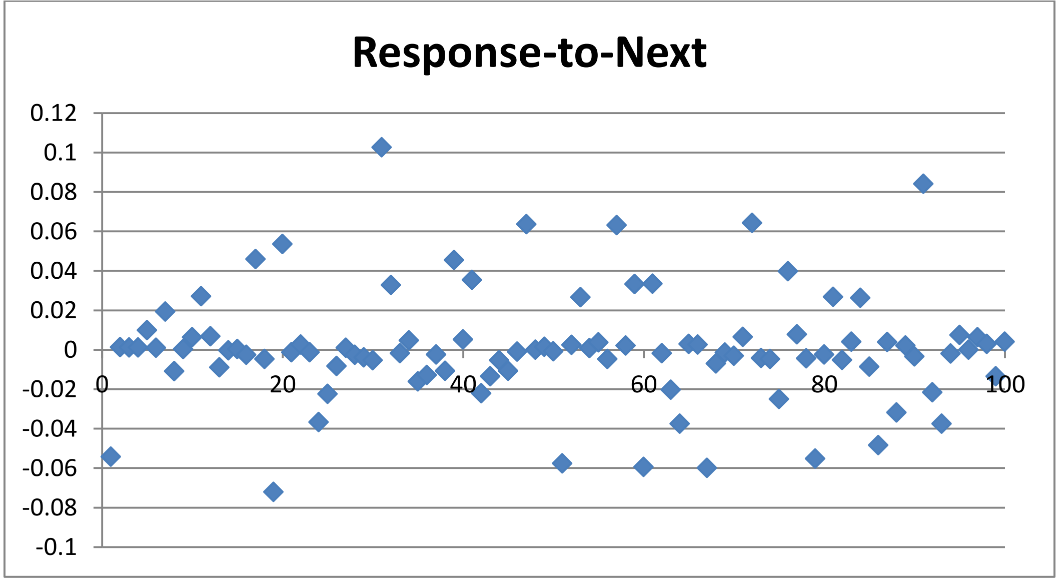
Marker – One occurrence



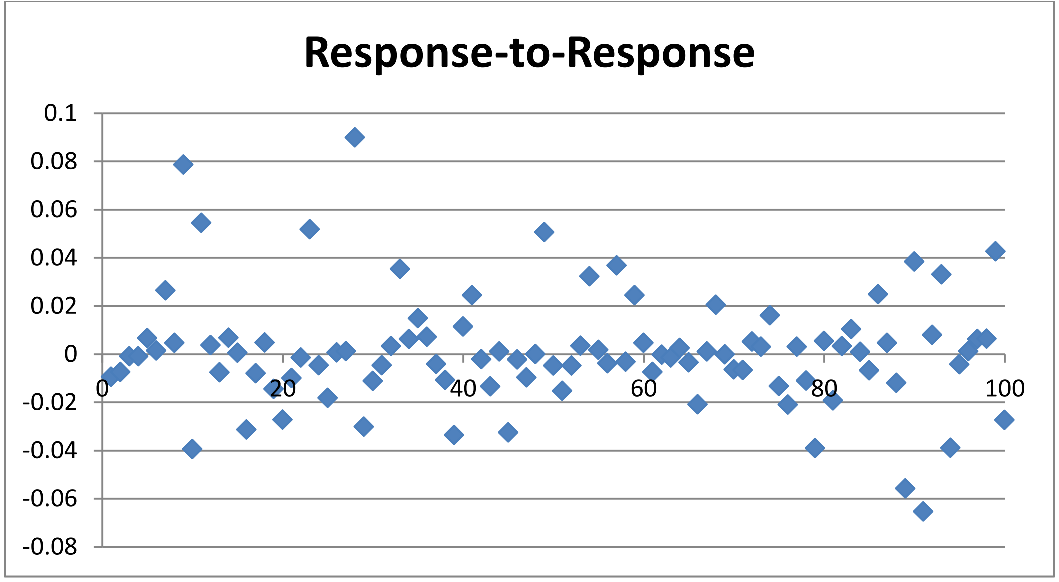
Marker – More than one occurrence



Response-to-Next

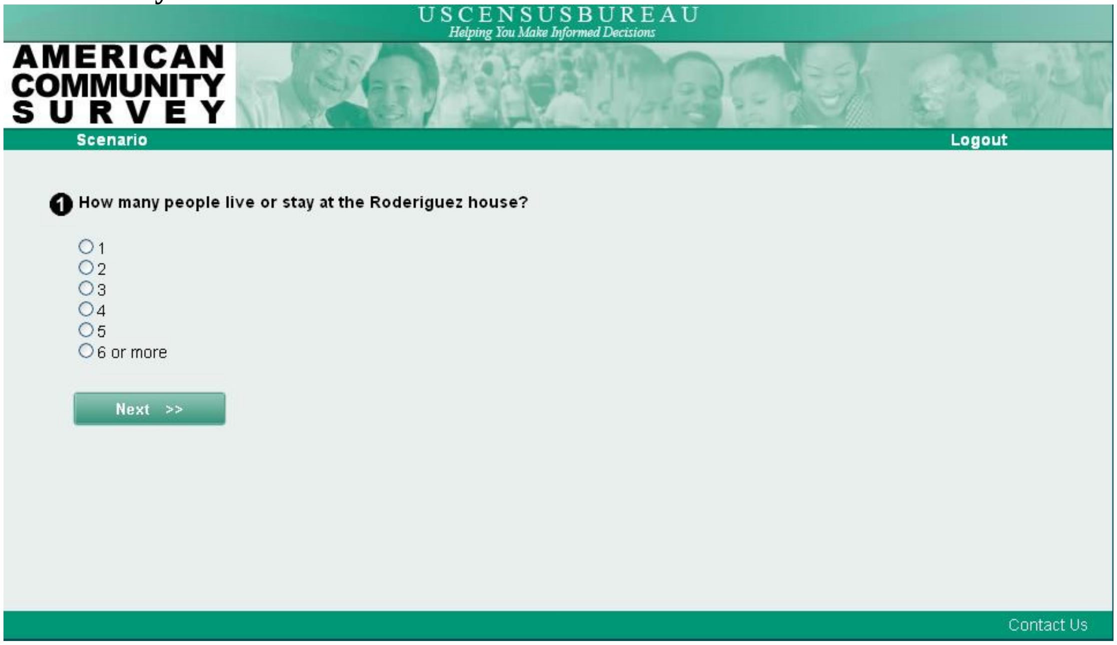


Response-to-Response



Appendix O. ACS Questions, Scenarios, and Help Content for Study 3

Live or Stay


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

1 How many people live or stay at the Roderiguez house?

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Straightforward Scenario – Cassandra and Miguel Roderiguez just moved in to the 3-bedroom house at 4694 Main St. They previously lived in a larger house with their two children Inez and Jose. But since Inez and Jose moved out and started families of their own, Cassandra and Miguel wanted a smaller house because no one else lives or stays in the house with them.

Complex Scenario – The Roderiguez family lives in the 3-bedroom house at 4694 Main St. The family has four members: Cassandra and Miguel Roderiguez, and their two children Inez and Jose. Cassandra and Miguel stay in the master bedroom, and Inez and Jose each have a room of their own. Inez always stays in her room, but Jose is a college student. Although his legal address is 4694 Main St., he lives at the college dorms all year, including summers, and only comes home for holidays and vacations.

Help Information

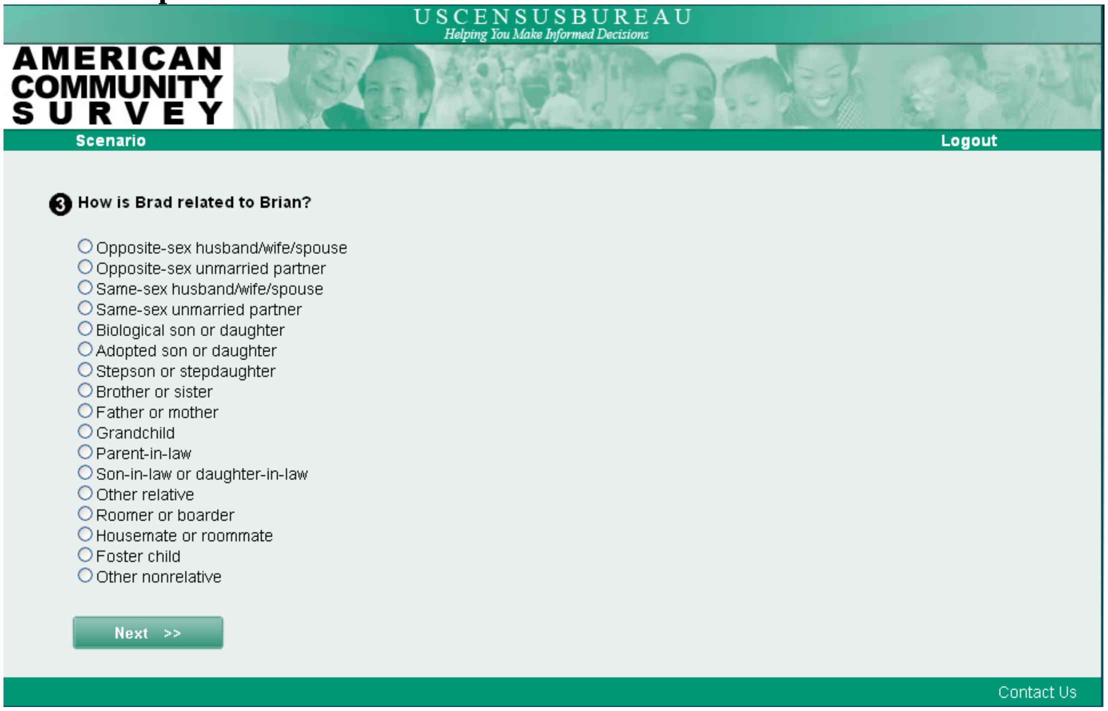
Include in count anyone who:

- Is currently living or staying there for more than 2 months
- Intends to be there for more than 2 months, even if they have been there less time than that as of today
- Is away now but is not planning to be away for more than 2 months
- Is a child living at boarding schools or at summer camp

Do NOT include anyone who:

- Is currently a college student living away at school for more than two months
- Is armed forces personnel who are living away for more than two months
- Has another place to live or stay

Relationship


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

❸ How is Brad related to Brian?

○ Opposite-sex husband/wife/spouse
○ Opposite-sex unmarried partner
○ Same-sex husband/wife/spouse
○ Same-sex unmarried partner
○ Biological son or daughter
○ Adopted son or daughter
○ Stepson or stepdaughter
○ Brother or sister
○ Father or mother
○ Grandchild
○ Parent-in-law
○ Son-in-law or daughter-in-law
○ Other relative
○ Roomer or boarder
○ Housemate or roommate
○ Foster child
○ Other nonrelative

***Straightforward Scenario* –** Brian and Brad went to college together and after graduation they decided to live together in a rented 2-bedroom apartment. They share rent, are not related, and they are not in a relationship

***Complex Scenario* –** Brian and Brad live together in a 2 bedroom house. Brian owns the house and Brad pays rent to Brian. Brad has his own room and access to certain parts of the house. They are not related.

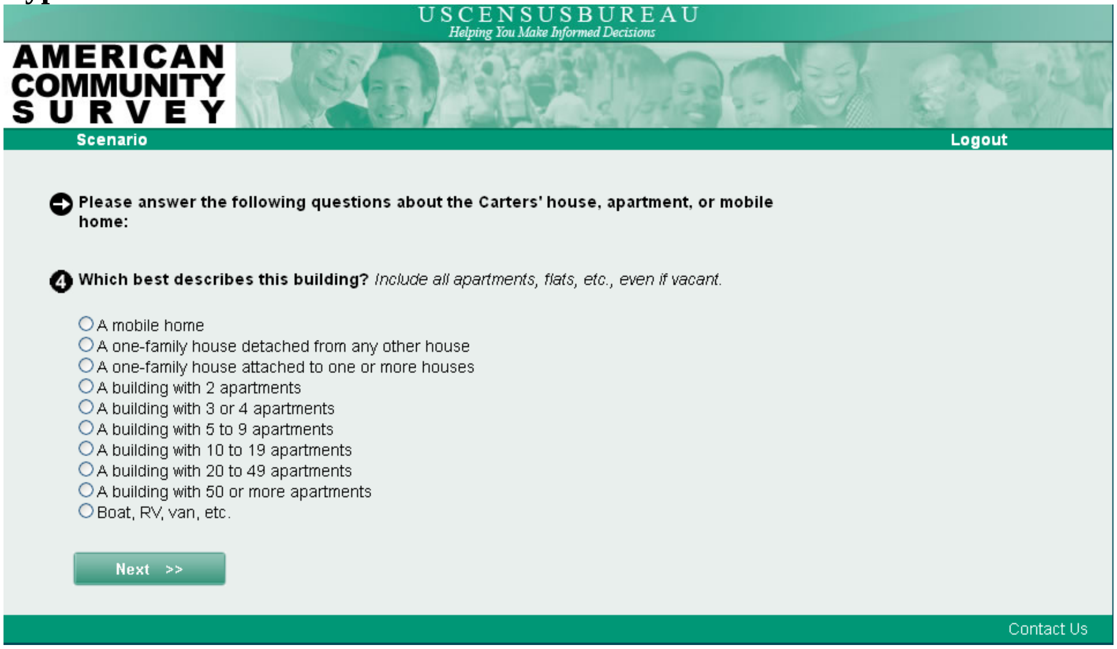
Help Information

Roomer or boarder – Occupies the same residence as the owner and pays rent directly to the owner

Housemate or roommate – Shares living quarters primarily to share expenses, including rent, BUT does not share a close personal relationship

Other nonrelative – Not related AND not one of the options listed

Type of Unit


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

➔ Please answer the following questions about the Carters' house, apartment, or mobile home:

❹ **Which best describes this building?** *Include all apartments, flats, etc., even if vacant.*

○ A mobile home
○ A one-family house detached from any other house
○ A one-family house attached to one or more houses
○ A building with 2 apartments
○ A building with 3 or 4 apartments
○ A building with 5 to 9 apartments
○ A building with 10 to 19 apartments
○ A building with 20 to 49 apartments
○ A building with 50 or more apartments
○ Boat, RV, van, etc.

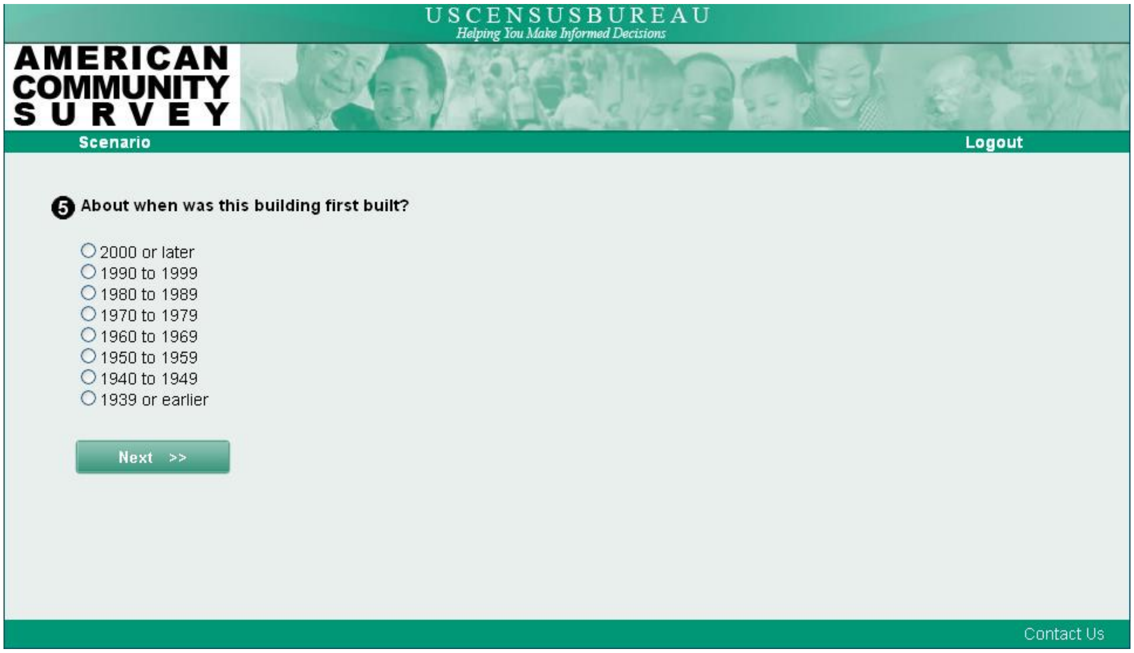
***Straightforward Scenario (UPDATED)* –** The Carter family lives in a suburban area in the midwest. They own a 3-bedroom house, which has a small yard that surrounds the house.

***Complex Scenario* –** The Carter family lives in a 3 story townhouse, attached to the houses on either side. The townhouse is broken into two condominiums, one on each floor. There is also a basement condominium that is not currently inhabited, so the Carters and the other families use it for storage.

Help Information

- If house has open space on all sides, then select "A one-family house detached from any other house"
- Count both vacant and occupied units
- Condominiums are equivalent to apartments
- If a townhouse has been converted into a condominium or apartment, count as an apartment

Year Built


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

5 About when was this building first built?

○ 2000 or later
○ 1990 to 1999
○ 1980 to 1989
○ 1970 to 1979
○ 1960 to 1969
○ 1950 to 1959
○ 1940 to 1949
○ 1939 or earlier

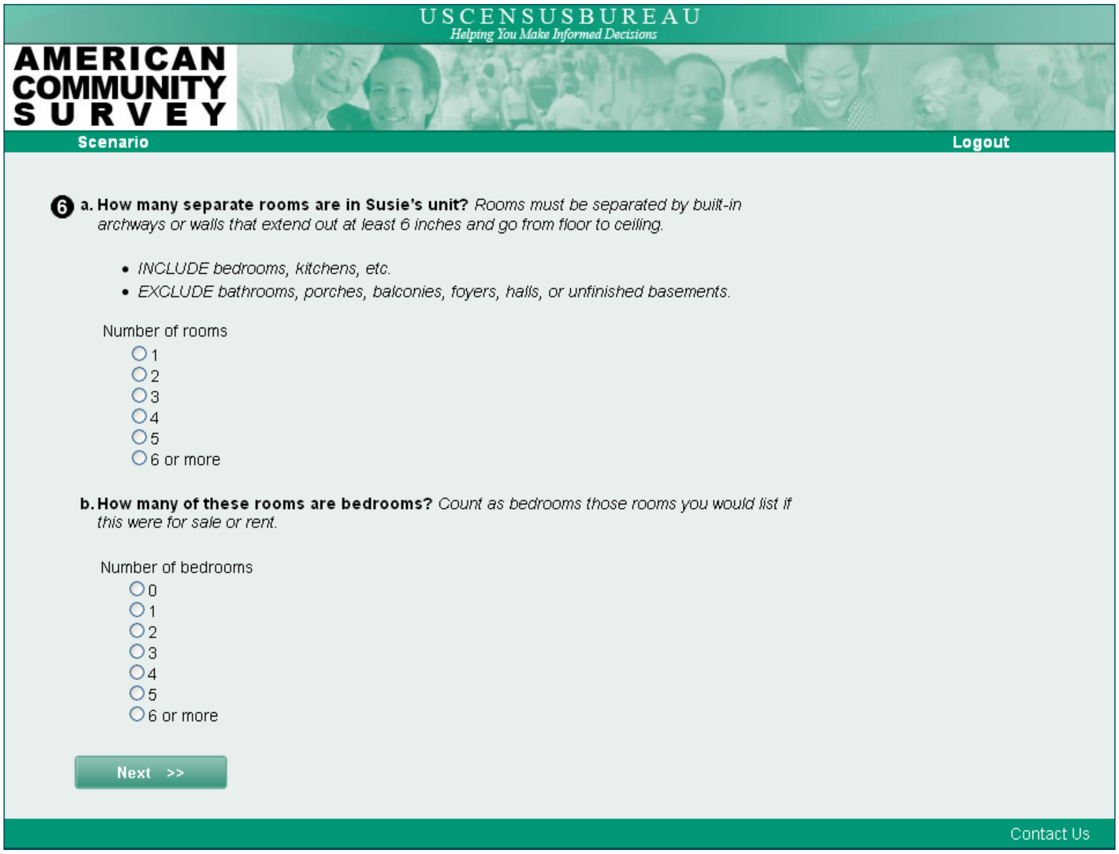
***Straightforward Scenario* –** The Carters designed and began building their own home after they were married in 1973. It was finished in 1974.

***Complex Scenario* –** The Carters designed and began building their own home after they were married in 1998. It was finished in 2000.

Help Information

Select the range in which the original construction was *finished*.

Rooms


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

❻ a. **How many separate rooms are in Susie's unit?** *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

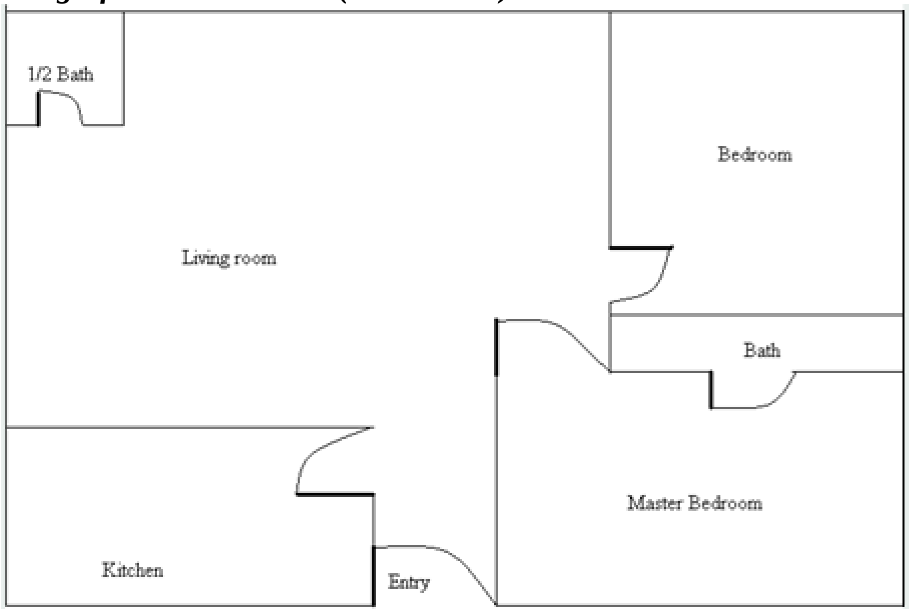
- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

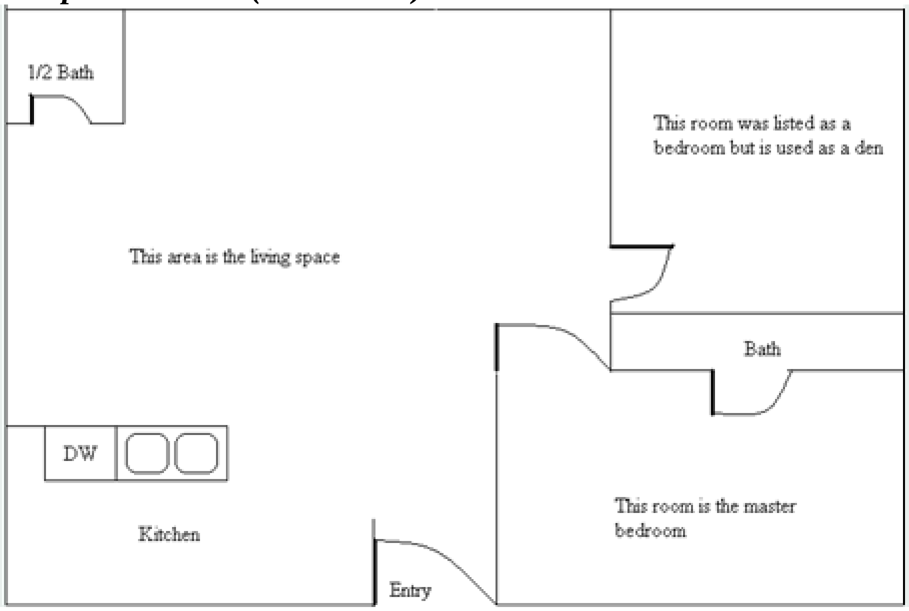
b. **How many of these rooms are bedrooms?** *Count as bedrooms those rooms you would list if this were for sale or rent.*

Number of bedrooms
○ 0
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Straightforward Scenario (UPDATED) –



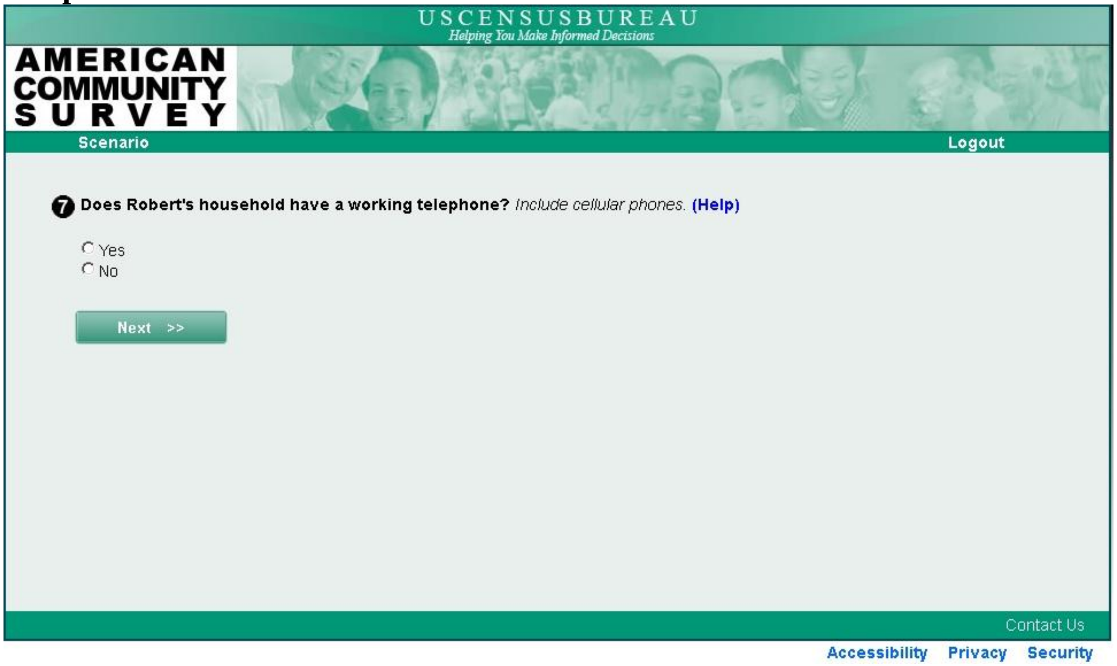
Complex Scenario (UPDATED) –


1/2 Bath
This room was listed as a bedroom but is used as a den
This area is the living space
Bath
DW
This room is the master bedroom
Kitchen
Entry

Help Information

- **Include** only whole rooms used for living purposes, such as living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms, family rooms, enclosed porches suitable for year-round use, etc. **DO NOT** count bathrooms
- Partially divided rooms, such as a dinette next to a kitchen or living room is a separate room ONLY if there is a built in partition or wall from floor to ceiling
- Count bedrooms as if this unit was for sale, not based on their current use

Telephone



7 **Does Robert's household have a working telephone?** *Include cellular phones.* **(Help)**

○ Yes
○ No

Straightforward Scenario (UPDATED) – Robert lives alone in Chicago, Il. He subscribes to Sprint’s cell phone service and carries a working cellular phone with him that can make and receive calls.

Complex Scenario (UPDATED) – Robert lives with is wife and two teenage children in Chicago, Il. Robert carries a working cellular phone but no one else in the family does and the house does not have a landline. He subscribes to Sprint’s phone service so his phone can make and receive calls.

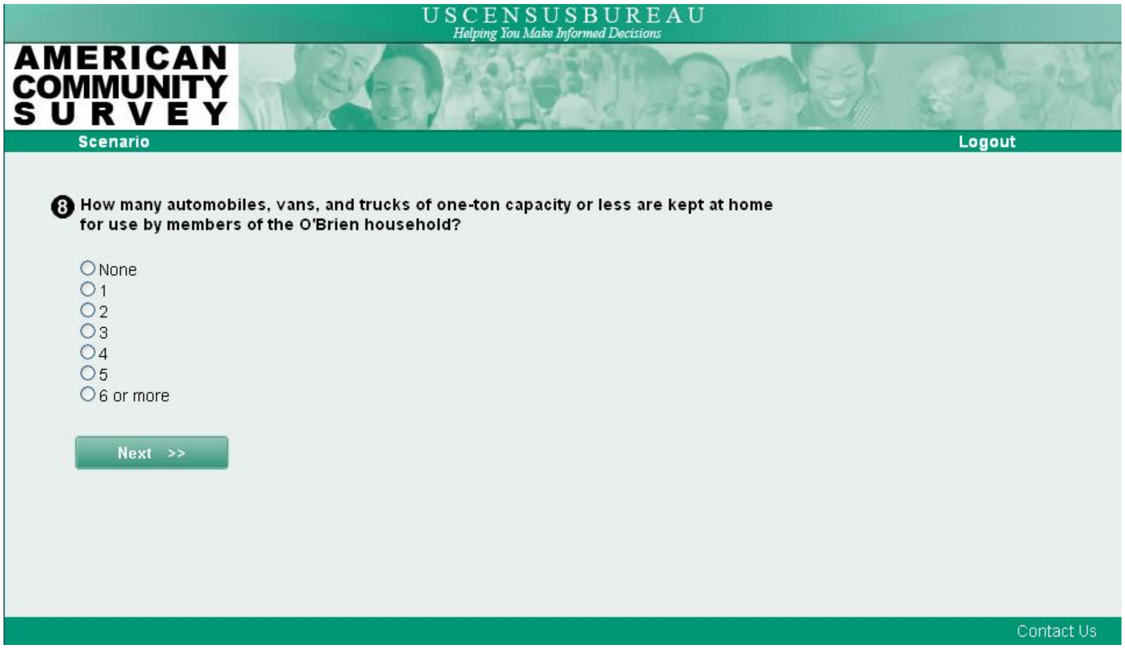
Help Information

Select ‘Yes’ under the following conditions:

- There is a telephone in working order AND someone receives service
- Someone has a cell phone that can both make AND receive calls

Select ‘No’ if the telephone service is disconnected for nonpayment or other reasons

Vehicles


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

8 How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of the O'Brien household?

○ None
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6 or more

Straightforward Scenario – The O'Brien family has 2 cars, which they use daily to drive to work. They keep the cars parked in their garage when they are not in use.

Complex Scenario – The O'Brien family has a recreational vehicle (RV) that they use for vacations. They keep it parked in their driveway when it is not in use. They have no other vehicles.

Help Information

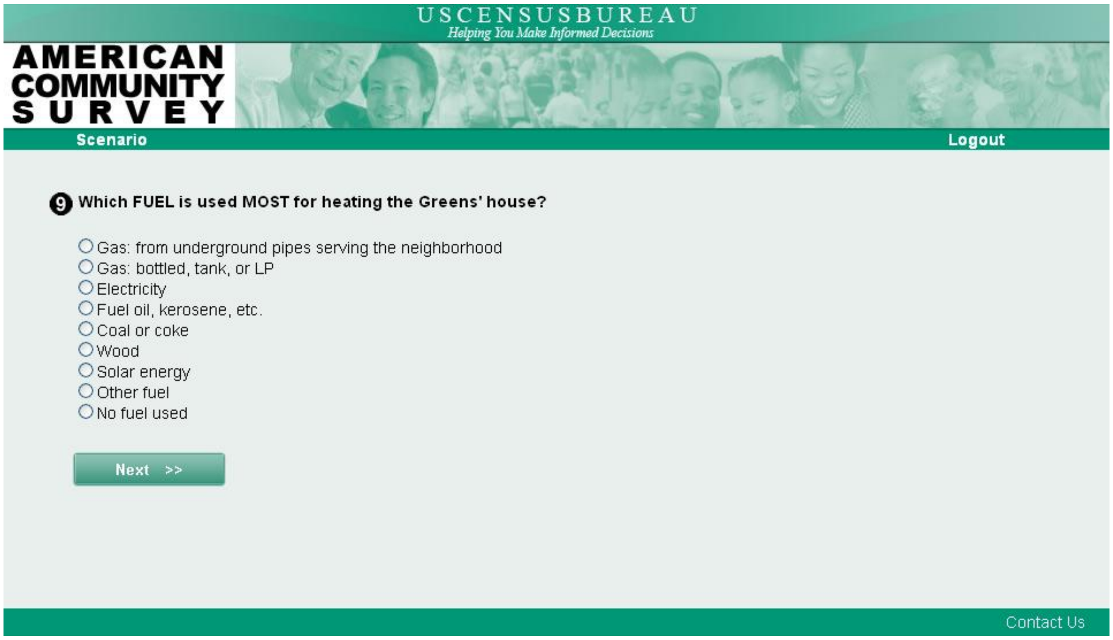
Include in count:

- Cars, vans or SUVs if they are
  - Regularly kept at home AND
  - Used by household for nonbusiness purposes

Do NOT include in count:

- Recreational Vehicles
- Motorcycles

Fuel


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

9 Which FUEL is used MOST for heating the Greens' house?

○ Gas: from underground pipes serving the neighborhood
○ Gas: bottled, tank, or LP
○ Electricity
○ Fuel oil, kerosene, etc.
○ Coal or coke
○ Wood
○ Solar energy
○ Other fuel
○ No fuel used

Straightforward Scenario (UPDATED) – The Smiths live in Phoenix, Arizona. Since Phoenix is generally warm year round, they do not heat their home using any type of fuel. Rather, they just use extra blankets if the temperature drops.

Complex Scenario (UPDATED) – The Smiths use a geothermal heat pump to heat their home. They use geothermal energy because they are environmentally conscious and the carbon monoxide emissions are far less than are emitted by fuel plants.

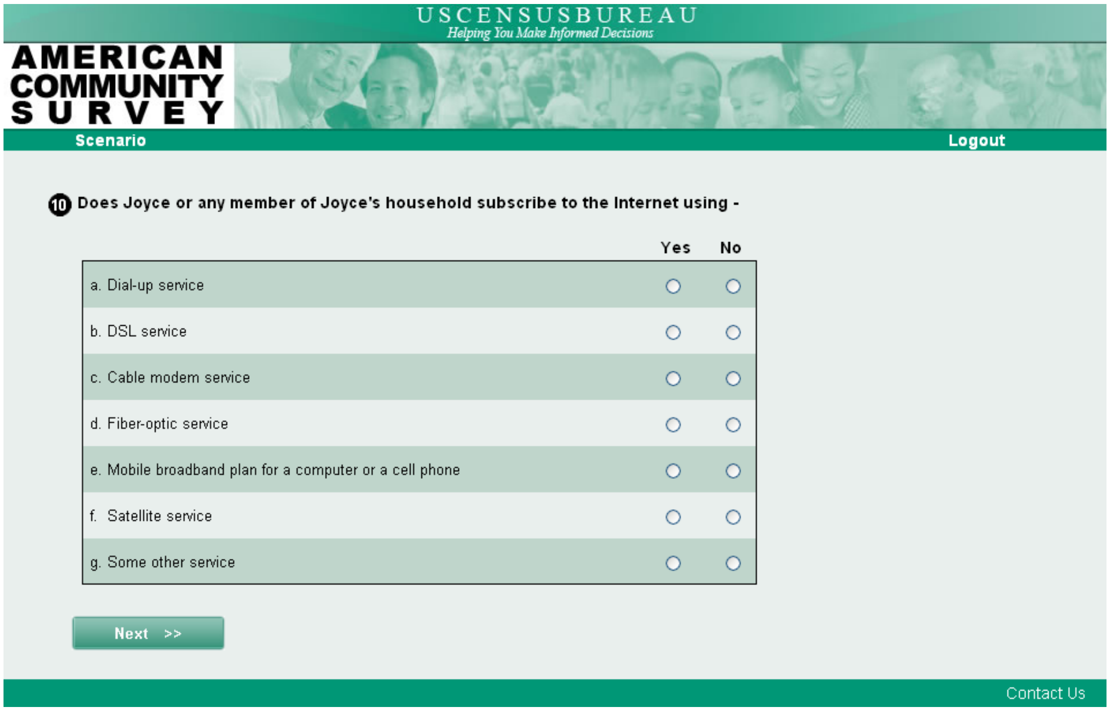
Help Information

Fuel oil or kerosene – Include any liquid petroleum product that is burned in a furnace or boiler for the generation of heat

Solar energy – Include any system that collects, stores, and distributes energy directly from the sun
Other fuel – Include fuel not listed separately, such as:

- Purchased steam
- Fuel briquettes
- Geothermal heat pumps
- Waste materials

Internet


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

10 Does Joyce or any member of Joyce's household subscribe to the Internet using -

	Yes	No
a. Dial-up service	○	○
b. DSL service	○	○
c. Cable modem service	○	○
d. Fiber-optic service	○	○
e. Mobile broadband plan for a computer or a cell phone	○	○
f. Satellite service	○	○
g. Some other service	○	○

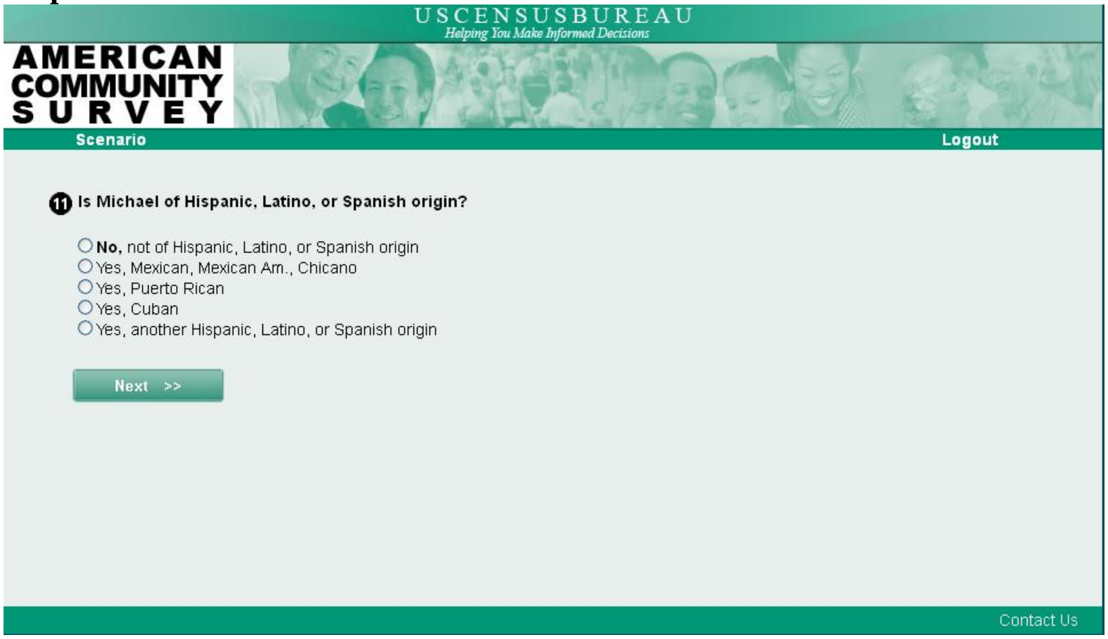
Straightforward Scenario (UPDATED) – Joyce has an Internet subscription through Comcast for a traditional cable modem. She pays a monthly bill to Comcast and she does not subscribe to any other Internet services.

Complex Scenario (UPDATED) – Joyce lives in an apartment building. Her neighbor, Jared, shared his password with her so she can access his fiber optic Internet connection. She has been using Jared's internet connection for 2 years and she pays him $20 a month for the access.

Help Information

Select 'Yes' if the user has directly entered into an agreement with a company to receive Internet service in exchange for payment.
Select 'No' if someone from another household entered into an agreement with a company to receive Internet service, even if multiple households use the service.

Hispanic


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

11 Is Michael of Hispanic, Latino, or Spanish origin?

○ **No,** not of Hispanic, Latino, or Spanish origin
○ Yes, Mexican, Mexican Am., Chicano
○ Yes, Puerto Rican
○ Yes, Cuban
○ Yes, another Hispanic, Latino, or Spanish origin

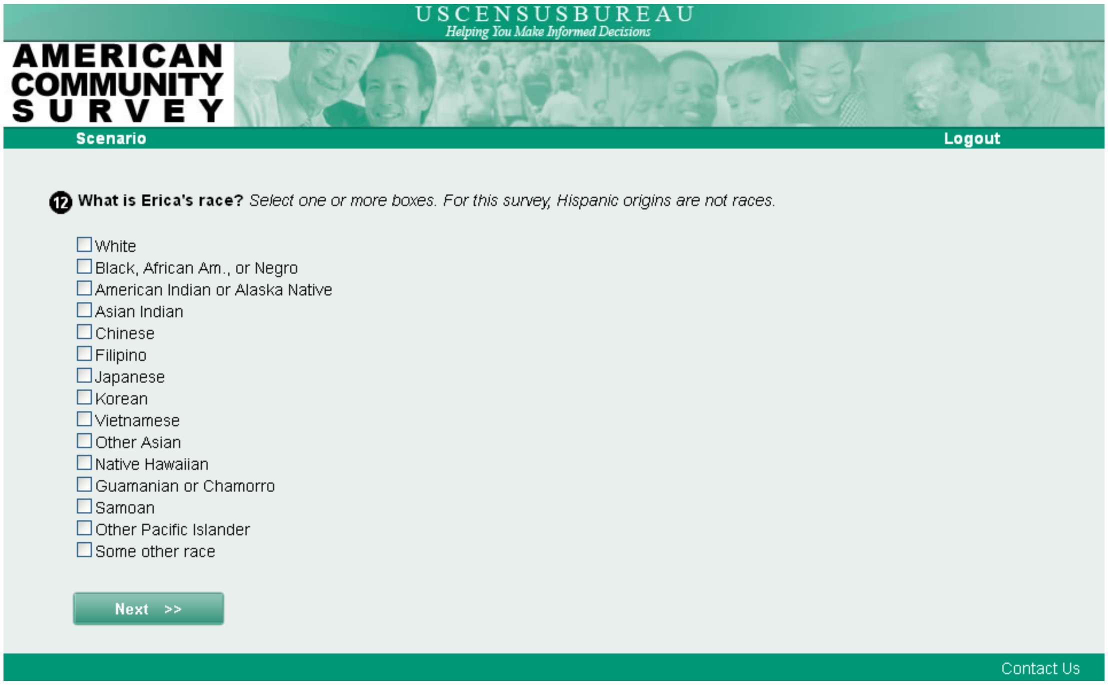
Straightforward Scenario (UPDATED) – Both of Michael's biological parents were born and raised in Russia. Michael was also born in Russia and lived there until moving to the United States two years ago. He also identifies as Russian.

Complex Scenario – Michael's grandmother on his father's side was of Spanish descent. His mother's family is from Germany. Michael does not identify with his father's side of the family.

Help Information

The concept of Hispanic, Latino, or Spanish origin as used by the Census Bureau reflects self-identification. It does not indicate any clear-cut scientific definition that is biological or genetic in reference.

Race


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

12 **What is Erica's race?** *Select one or more boxes. For this survey, Hispanic origins are not races.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native
□ Asian Indian
□ Chinese
□ Filipino
□ Japanese
□ Korean
□ Vietnamese
□ Other Asian
□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander
□ Some other race

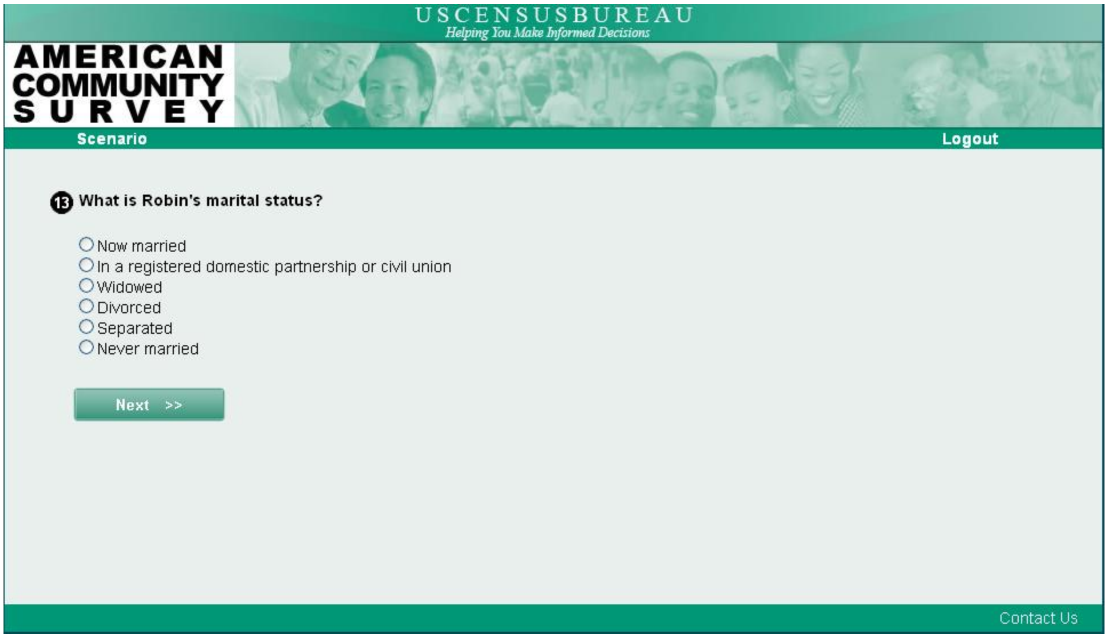
***Straightforward Scenario* –** Erica's biological mother is Caucasian, as is her biological father. She was raised by her biological parents. She also identifies as White.

***Complex Scenario* –** Erica's biological mother was African American and her biological father was Caucasian. She was adopted by an African American family and identifies as African American.

Help Information

The concept of race, as used by the Census Bureau, reflects self-identification by individuals according to the race or races with which they identify.

Marital Status


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

13 What is Robin's marital status?

○ Now married
○ In a registered domestic partnership or civil union
○ Widowed
○ Divorced
○ Separated
○ Never married

Straightforward Scenario – Robin was married for 10 years, but her husband died in a car accident two years ago. She has not remarried nor had any prior marriages.

Complex Scenario (UPDATED) – Robin and her partner Jen have lived together for 10 years. Last year they had a wedding ceremony in Illinois last year at their local church. Illinois does not recognize same sex marriages. Neither Robin nor her partner has been previously married, nor have the ever registered for a domestic partnership or any kind of civil union.

Help Information

Now married – currently *legally* married, regardless of whether or not their spouses live with them

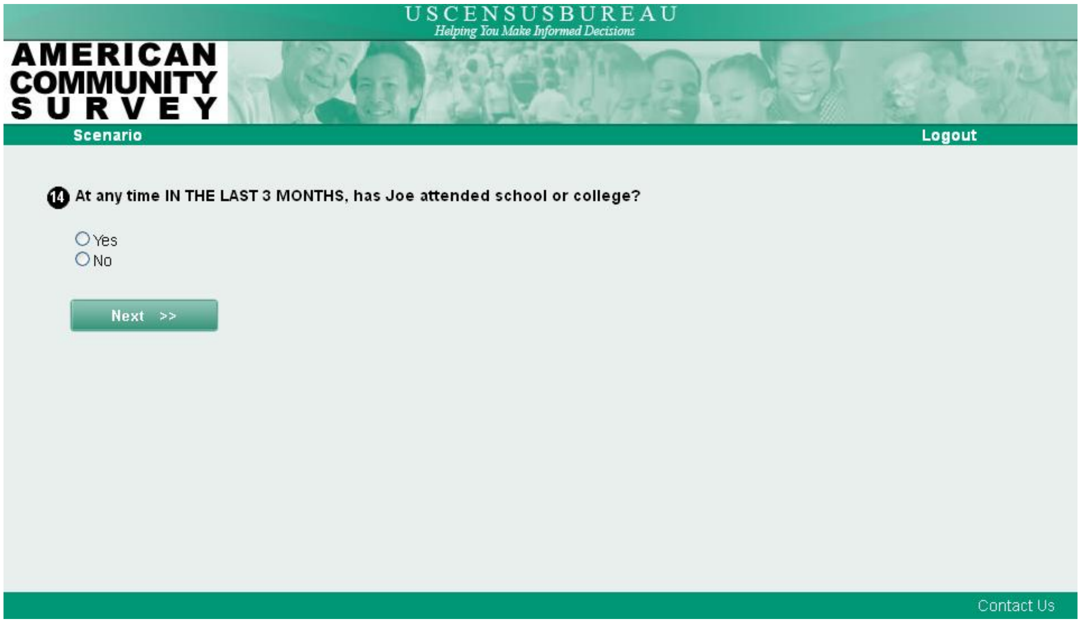
In a registered domestic partnership or civil union – currently *legally* registered by the state as a domestic partnership or civil union

Widowed – currently widowed

Divorced – currently divorced

Never married – never been married or whose only marriage has been legally annulled

Attend School


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

14 At any time IN THE LAST 3 MONTHS, has Joe attended school or college?

○ Yes
○ No

***Straightforward Scenario* –** Joe Smith is currently a junior at George Washington University, majoring in Business Administration. He is registered for five classes in the current semester and regularly attends all of them. Classes began in September and it is currently October.

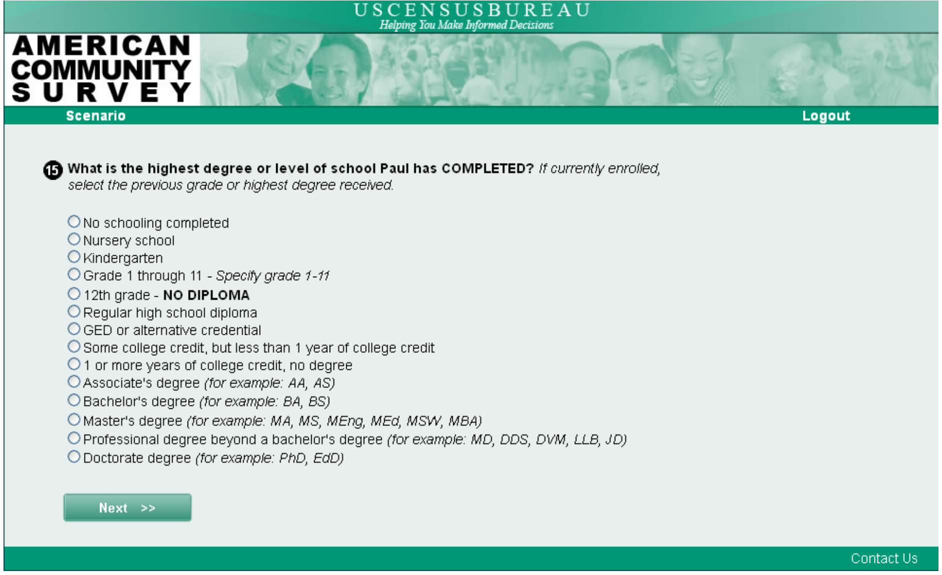
***Complex Scenario* –** Joe Smith has a Master's degree in electrical engineering which he received in 2004. Last month, at the request of his employer, he attended a 2-day seminar in communication cabling which earns him continuing education credits.

Help Information

Answer 'Yes' if the school leads to an elementary school certificate, high school diploma, or college, university, or professional school degree.
Answer 'No' for vocational or technical schools, adult education classes, and on-the-job-training.

Educational Attainment

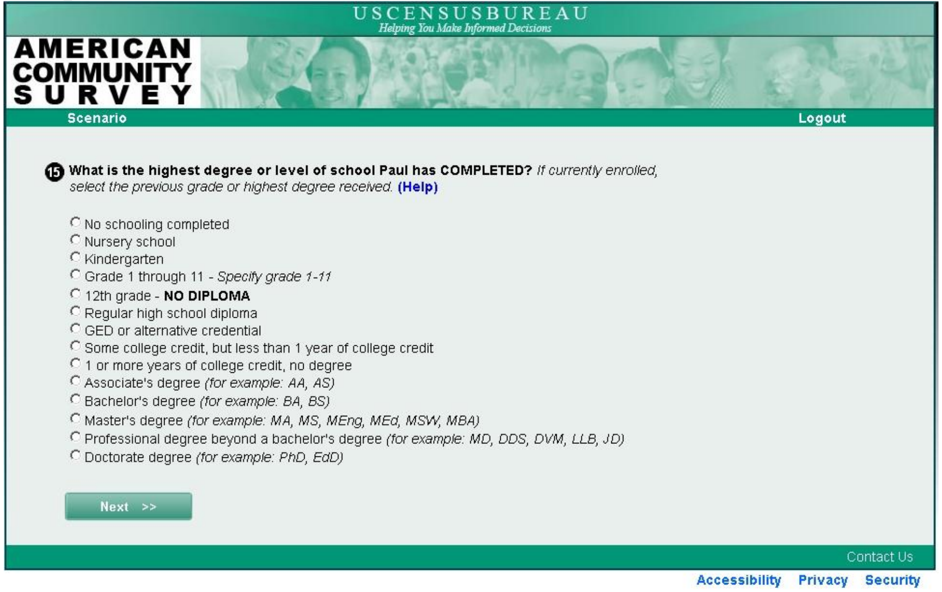
Straightforward – No labels


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

15 **What is the highest degree or level of school Paul has COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.*

○ No schooling completed
○ Nursery school
○ Kindergarten
○ Grade 1 through 11 - *Specify grade 1-11*
○ 12th grade - **NO DIPLOMA**
○ Regular high school diploma
○ GED or alternative credential
○ Some college credit, but less than 1 year of college credit
○ 1 or more years of college credit, no degree
○ Associate's degree *(for example: AA, AS)*
○ Bachelor's degree *(for example: BA, BS)*
○ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
○ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
○ Doctorate degree *(for example: PhD, EdD)*

Complex – Labels



15 **What is the highest degree or level of school Paul has COMPLETED?** *If currently enrolled, select the previous grade or highest degree received.* **(Help)**

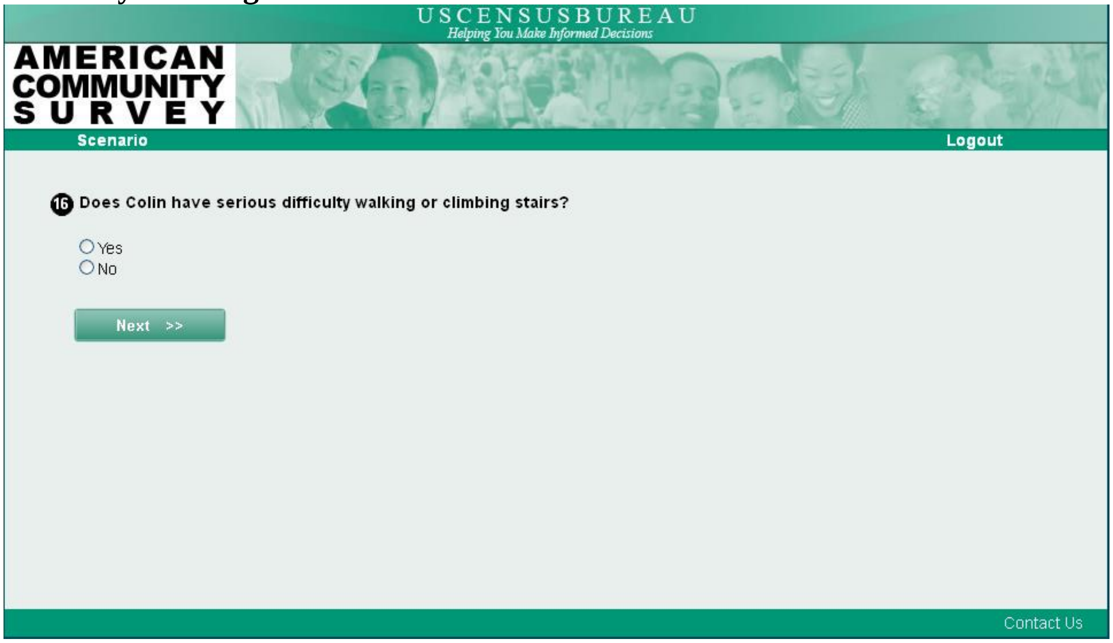
○ No schooling completed
○ Nursery school
○ Kindergarten
○ Grade 1 through 11 - *Specify grade 1-11*
○ 12th grade - **NO DIPLOMA**
○ Regular high school diploma
○ GED or alternative credential
○ Some college credit, but less than 1 year of college credit
○ 1 or more years of college credit, no degree
○ Associate's degree *(for example: AA, AS)*
○ Bachelor's degree *(for example: BA, BS)*
○ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
○ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
○ Doctorate degree *(for example: PhD, EdD)*

Scenario – Dylan graduated from a 4-year university last year. He will be starting a Masters program next year.

Help Information

- Only select one option
- If a person is enrolled in a grade or program but has not yet started, select the grade or highest level previously COMPLETED.

Difficulty Walking


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

16 Does Colin have serious difficulty walking or climbing stairs?

○ Yes
○ No

***Straightforward Scenario* –** Colin is a healthy, active young adult. For exercise, he walks 2 miles every day and also jogs up and down the bleachers at the local high school.

***Complex Scenario* –** Colin sprained his ankle during a soccer game. He will be on crutches for the next week.

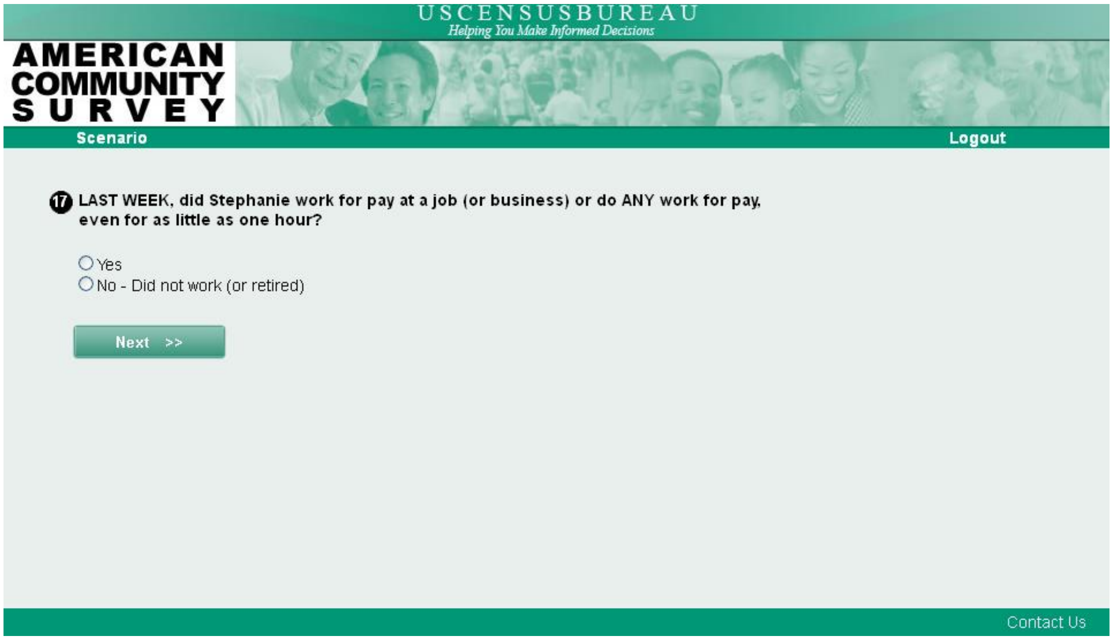
Help Information

Answer ‘Yes’ if it is sometimes or always very difficult or impossible for the person to:

- Walk three city blocks
- Climb a flight of stairs

Otherwise, answer ‘No’

Work Last Week


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

17 LAST WEEK, did Stephanie work for pay at a job (or business) or do ANY work for pay, even for as little as one hour?

○ Yes
○ No - Did not work (or retired)

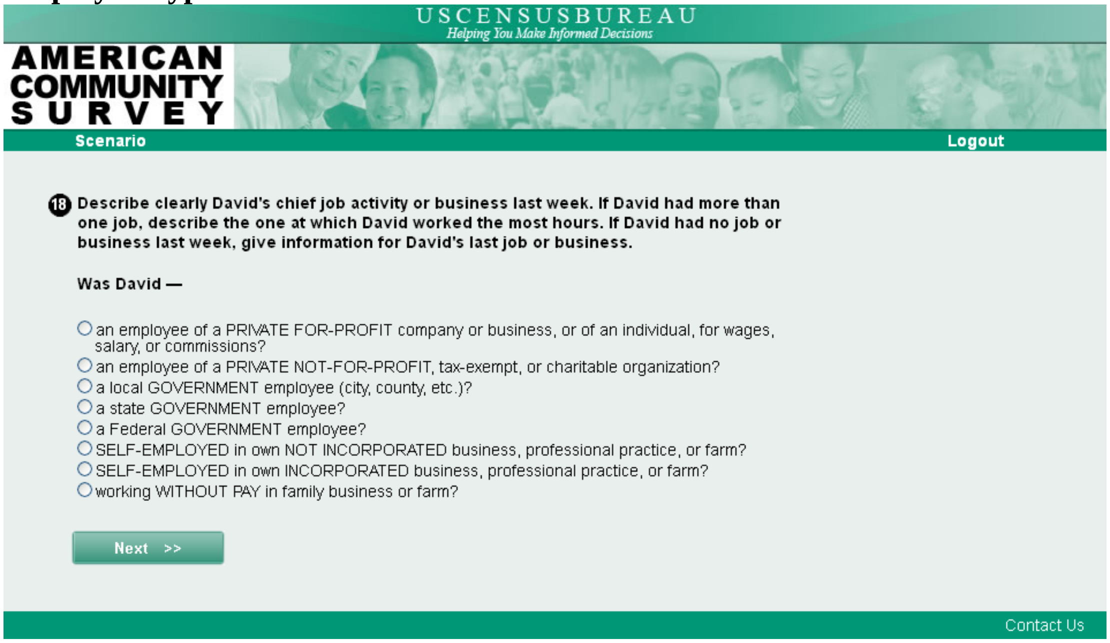
Straightforward Scenario – Stephanie works as a high school teacher and is paid on an annual 52-week contract. The school year runs from September through June and consists of two 18-week semesters. It is currently October.

Complex Scenario – Stephanie works as a high school teacher and is paid on an annual 52-week contract. The school year runs from September through June and consists of two 18-week semesters. It is currently August.

Help Information

Include any week the person spent on PAID vacation, PAID sick leave, or military service. Do not include weeks in which the person was on unpaid vacation or unpaid leave for the complete week.

Employee Type


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

18 Describe clearly David's chief job activity or business last week. If David had more than one job, describe the one at which David worked the most hours. If David had no job or business last week, give information for David's last job or business.

Was David —

○ an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
○ an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
○ a local GOVERNMENT employee (city, county, etc.)?
○ a state GOVERNMENT employee?
○ a Federal GOVERNMENT employee?
○ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
○ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
○ working WITHOUT PAY in family business or farm?

Straightforward Scenario – David has been directly employed by the U.S. Department of Justice for the past 5 years in Washington, DC. David worked at the Department of Justice last week and has not worked at any other job.

Complex Scenario (UPDATED) – David worked as a government employee for the Department of Justice for 35 years before retiring in 2009. After he retired, he became an independent contractor and obtained a contract with his old agency, the Department of Justice. This is the only job he worked last week.

Help Information

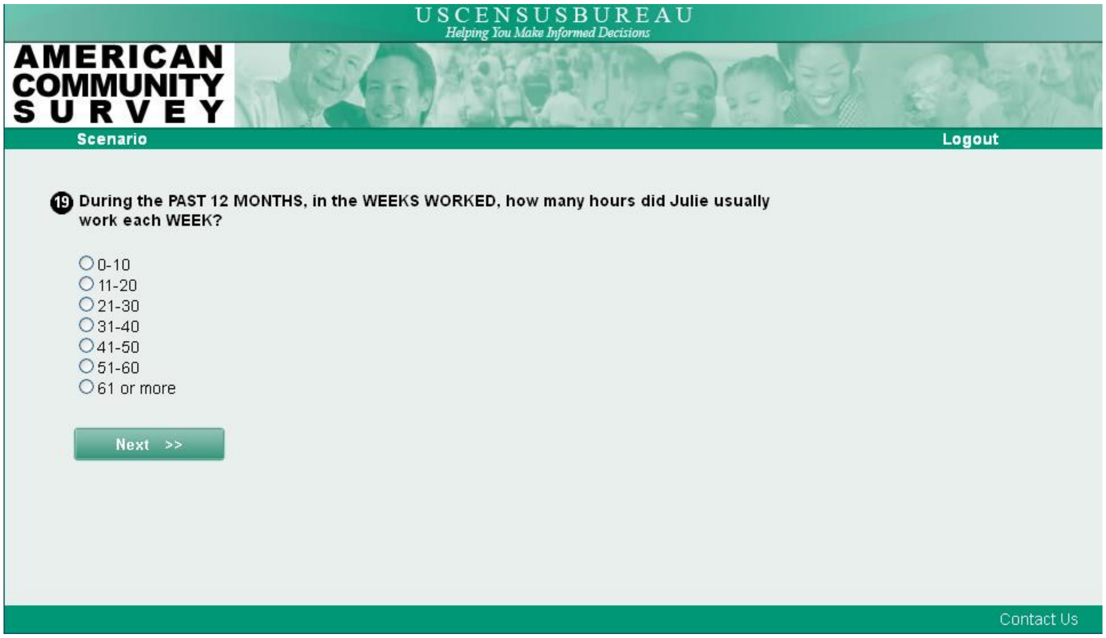
If work for cooperative, credit union, mutual insurance company, etc, then select “An employee of a PRIVATE NOT-FOR-PROFIT, tax exempt, or charitable organization”

If work for a US Federal department (Commerce, Justice, Agriculture), elected Federal official, Foreign government, then select “a Federal GOVERNMENT employee”

If work for a public state University, then select “a state GOVERNMENT employee”

If work as a private independent contractor, then select “SELF-EMPLOYED in own NOT INCORPORATED business, practice, or farm”

Hours Worked


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout
Next >>
Contact Us

19 During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did Julie usually work each WEEK?

○ 0-10
○ 11-20
○ 21-30
○ 31-40
○ 41-50
○ 51-60
○ 61 or more

Straightforward Scenario – For the past year, Julie has worked a regular 9 to 5 job with the Federal government, which works out to 40 hours per week.   She never works overtime, has not taken sick leave, and has not been on vacation this year. She does not have another job.

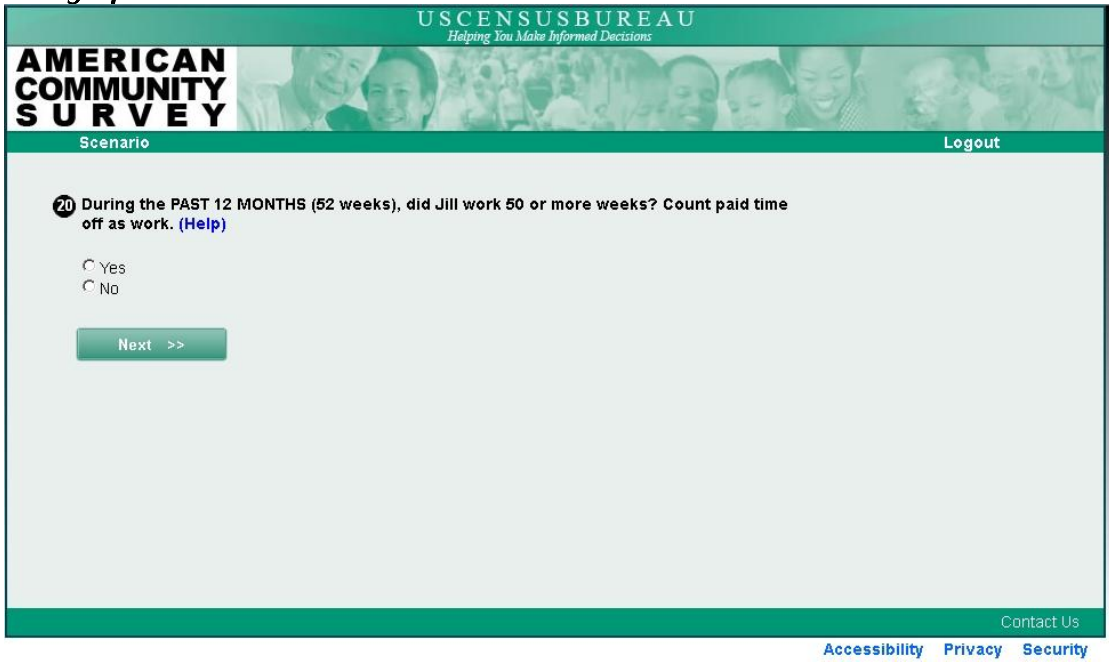
Complex Scenario – For the past year, Julie has spent 40 hours per week in the office at her government job. In the last six months Julie has been trying to start her own business in addition to her government job. So when she gets home, she spends 2 hours each weeknight working on starting her own business.

Help Information

Include extra hours usually worked, even if they are not compensated.
Select the range in which the typical hours per week fall. For example, if the person usually works 40 hours per week, select 31-40.

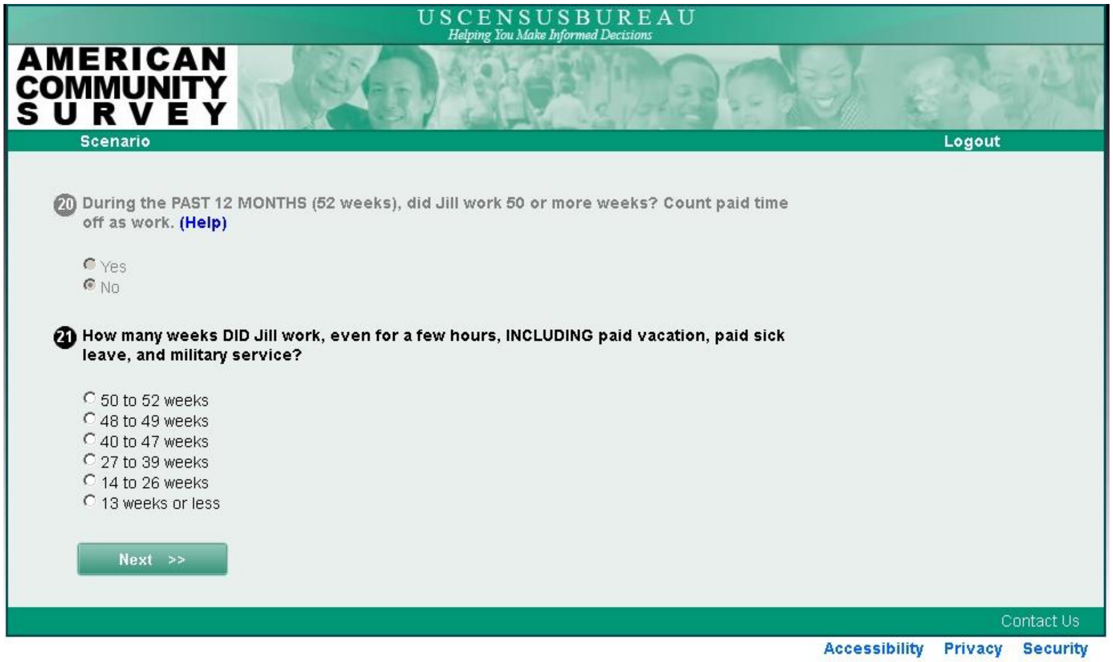
Weeks Worked

Straightforward



20 During the PAST 12 MONTHS (52 weeks), did Jill work 50 or more weeks? Count paid time off as work. (Help)

○ Yes
○ No



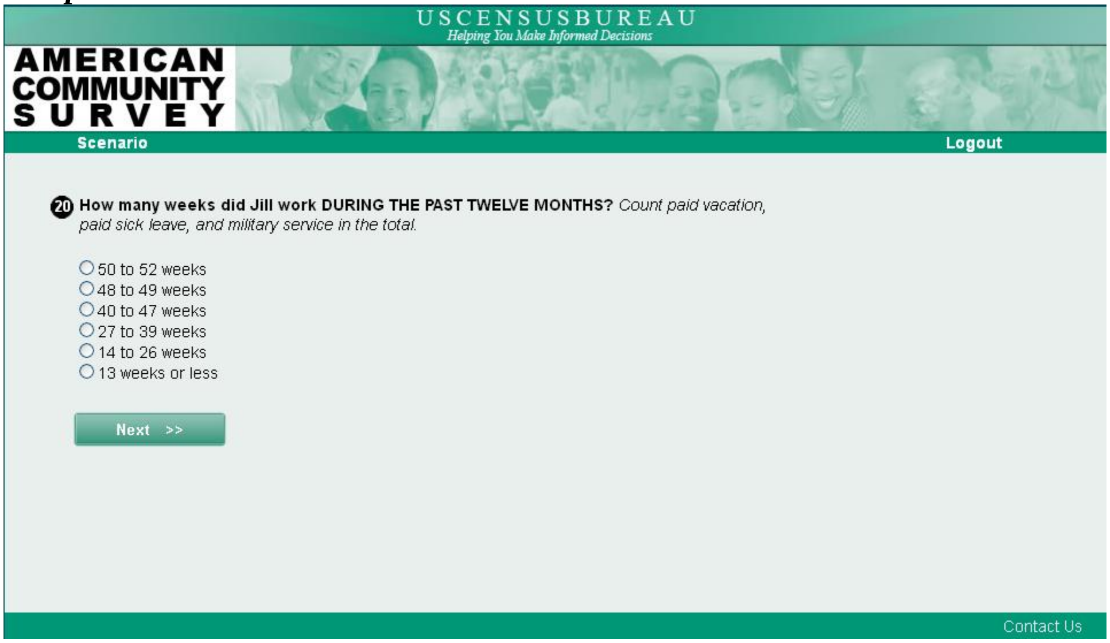
20 During the PAST 12 MONTHS (52 weeks), did Jill work 50 or more weeks? Count paid time off as work. (Help)

○ Yes
● No

21 How many weeks DID Jill work, even for a few hours, INCLUDING paid vacation, paid sick leave, and military service?

○ 50 to 52 weeks
○ 48 to 49 weeks
○ 40 to 47 weeks
○ 27 to 39 weeks
○ 14 to 26 weeks
○ 13 weeks or less

Complex


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

20 **How many weeks did Jill work DURING THE PAST TWELVE MONTHS?** *Count paid vacation, paid sick leave, and military service in the total.*

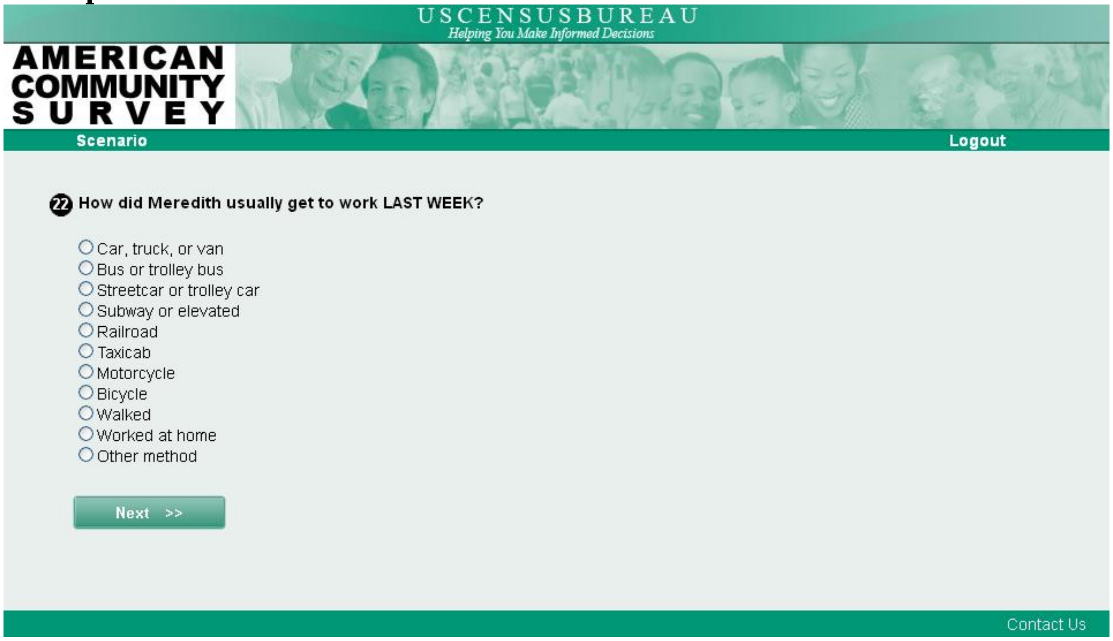
○ 50 to 52 weeks
○ 48 to 49 weeks
○ 40 to 47 weeks
○ 27 to 39 weeks
○ 14 to 26 weeks
○ 13 weeks or less

Scenario – Jill worked most of the year as a librarian but was laid off for the last three weeks of the year and did not find other employment.

Help Information

There are 52 weeks in a year.
Exclude any time off without pay for persons who had unpaid vacations, unpaid sick leave, periods of layoff or strike, periods receiving disability insurance when not actually working, etc.

Transport to Work


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Scenario
Logout

22 How did Meredith usually get to work LAST WEEK?

○ Car, truck, or van
○ Bus or trolley bus
○ Streetcar or trolley car
○ Subway or elevated
○ Railroad
○ Taxicab
○ Motorcycle
○ Bicycle
○ Walked
○ Worked at home
○ Other method

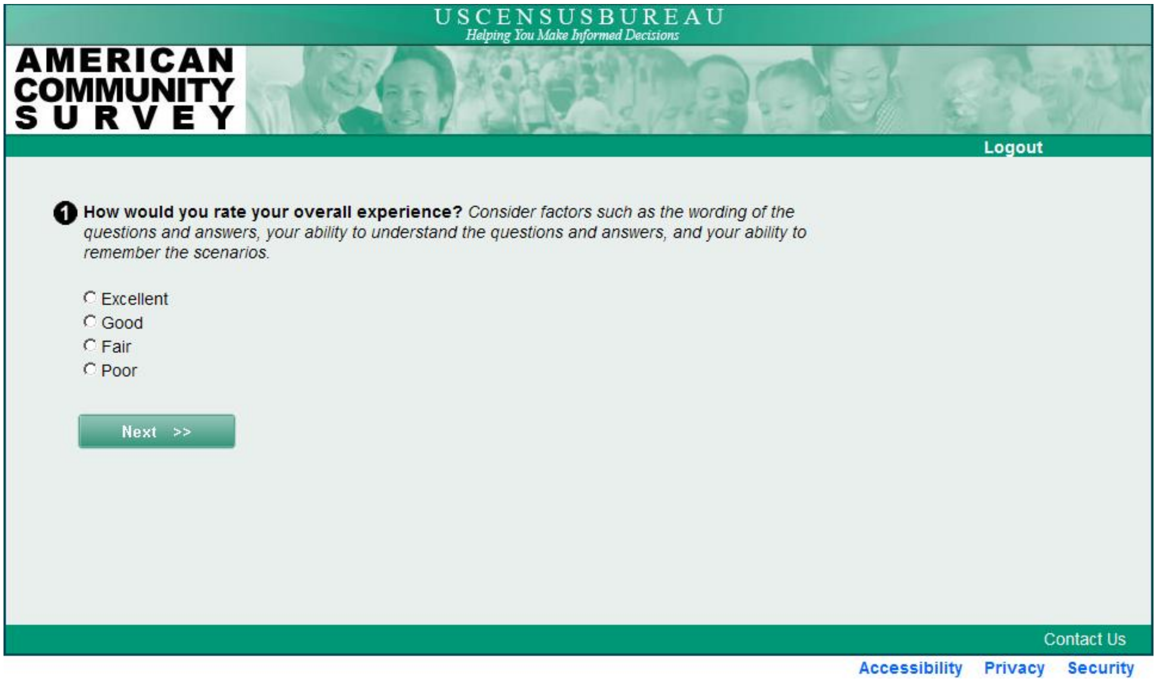
***Straightforward Scenario (UPDATED)* –** Meredith lives over an hour away from her company's office and there is no convenient public transportation near her home. Therefore, she has to drive to her car to work every day.

***Complex Scenario* –** Meredith does sales for a pharmaceutical company. There is no corporate office nearby, so her office is at her home. She makes daily sales visits to doctors' offices around the metro area. She drives a car to these appointments.

Help Information

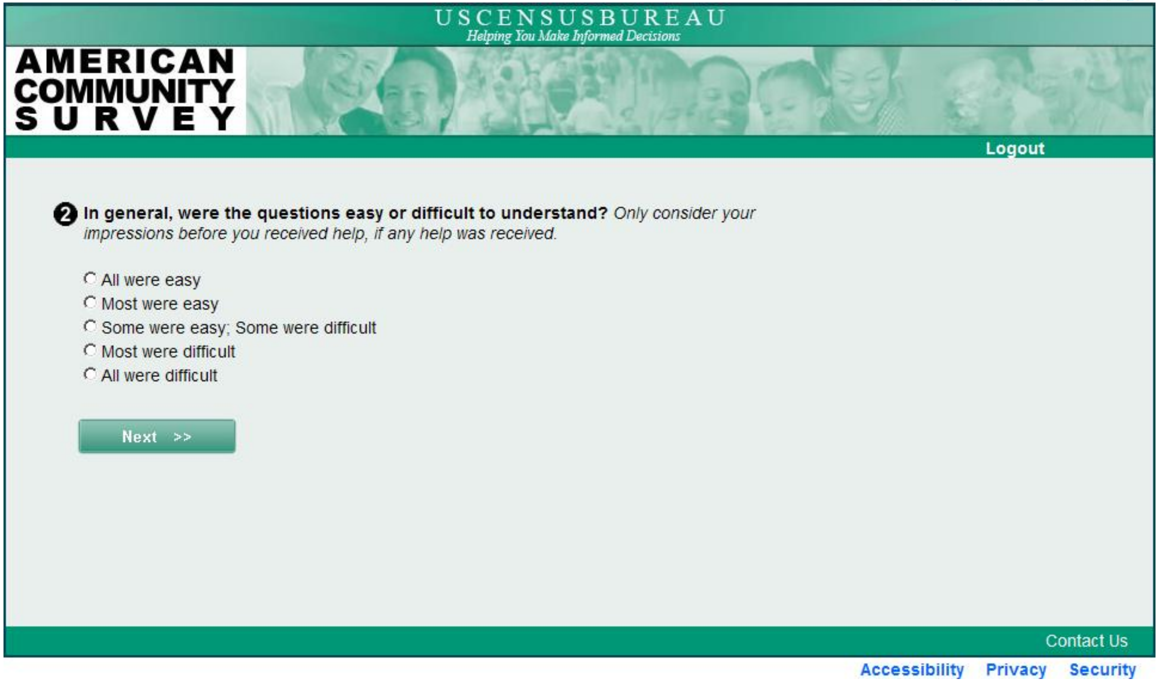
If a person works from several different locations, use their method of transportation to their primary place of work.

Appendix P. Screenshots of Satisfaction Questions for Study 3



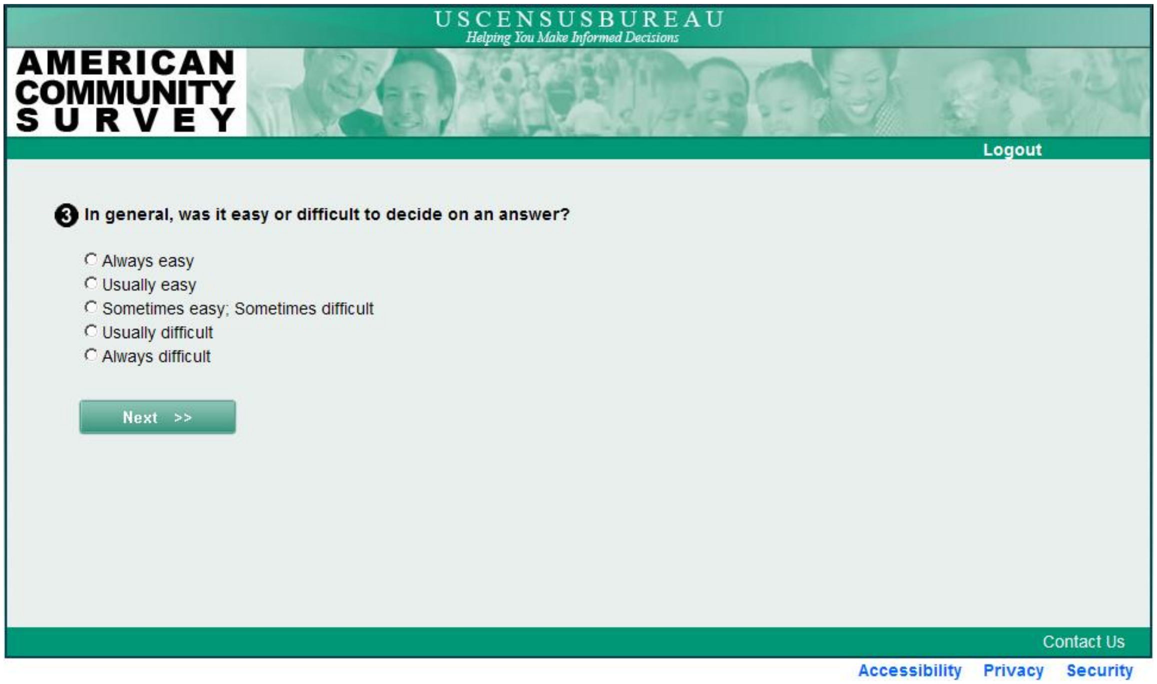
❶ **How would you rate your overall experience?** *Consider factors such as the wording of the questions and answers, your ability to understand the questions and answers, and your ability to remember the scenarios.*

○ Excellent
○ Good
○ Fair
○ Poor



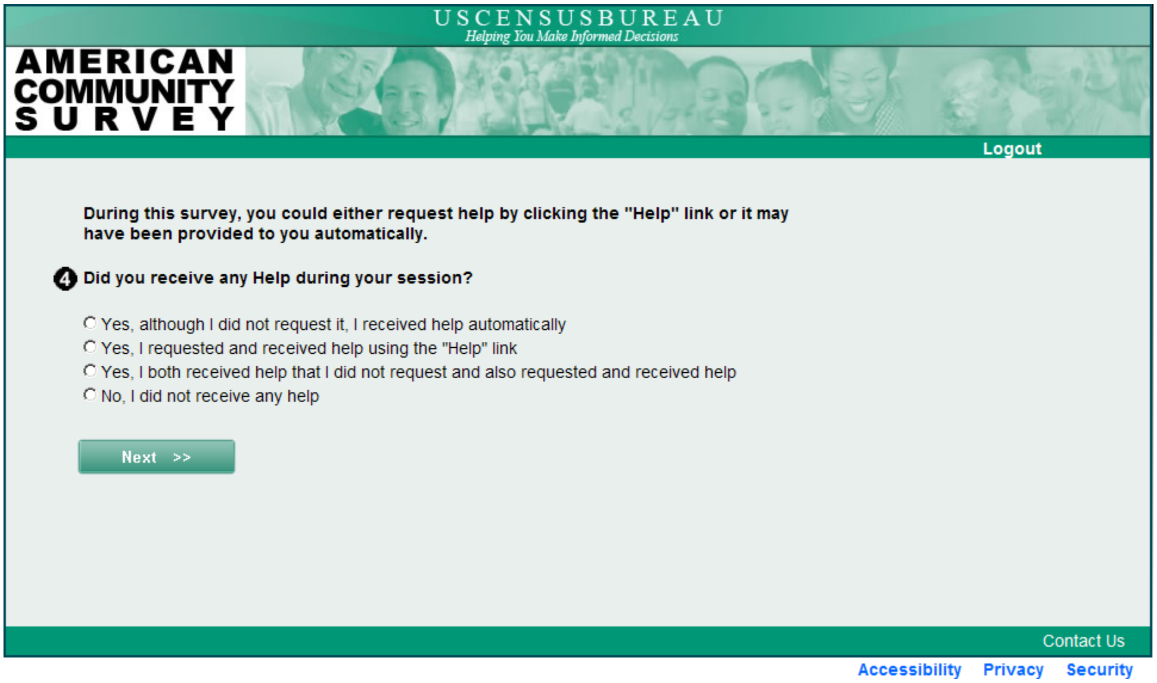
❷ **In general, were the questions easy or difficult to understand?** *Only consider your impressions before you received help, if any help was received.*

○ All were easy
○ Most were easy
○ Some were easy; Some were difficult
○ Most were difficult
○ All were difficult


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

❸ In general, was it easy or difficult to decide on an answer?

○ Always easy
○ Usually easy
○ Sometimes easy; Sometimes difficult
○ Usually difficult
○ Always difficult

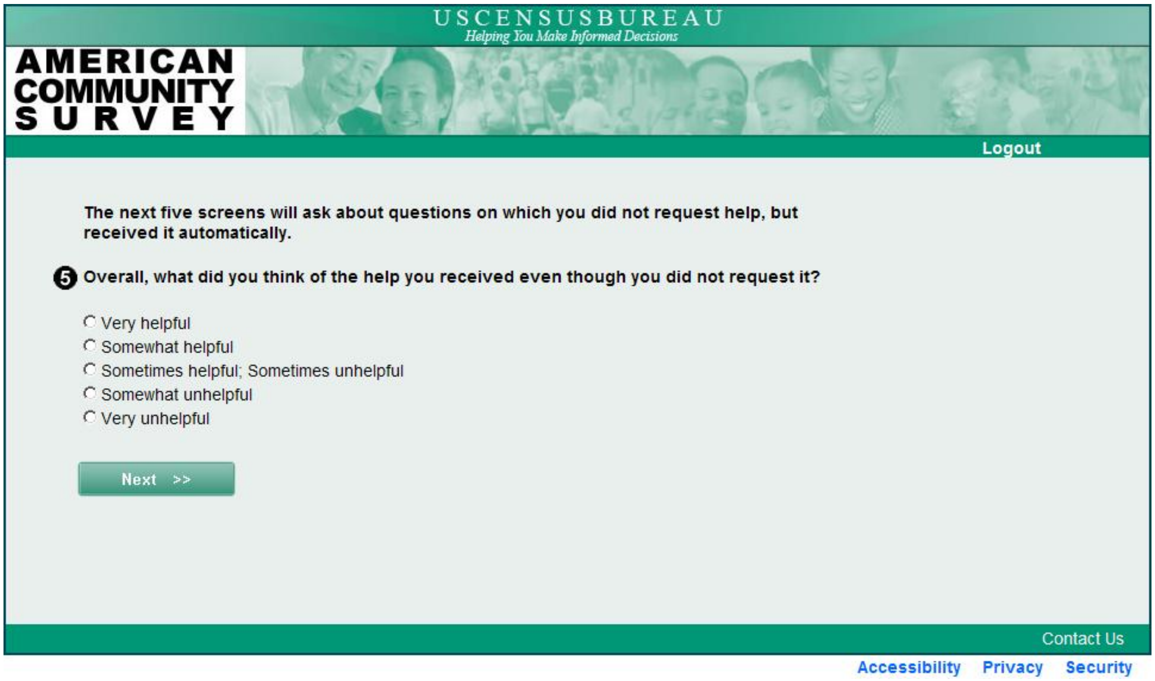

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

During this survey, you could either request help by clicking the "Help" link or it may have been provided to you automatically.

❹ Did you receive any Help during your session?

○ Yes, although I did not request it, I received help automatically
○ Yes, I requested and received help using the "Help" link
○ Yes, I both received help that I did not request and also requested and received help
○ No, I did not receive any help

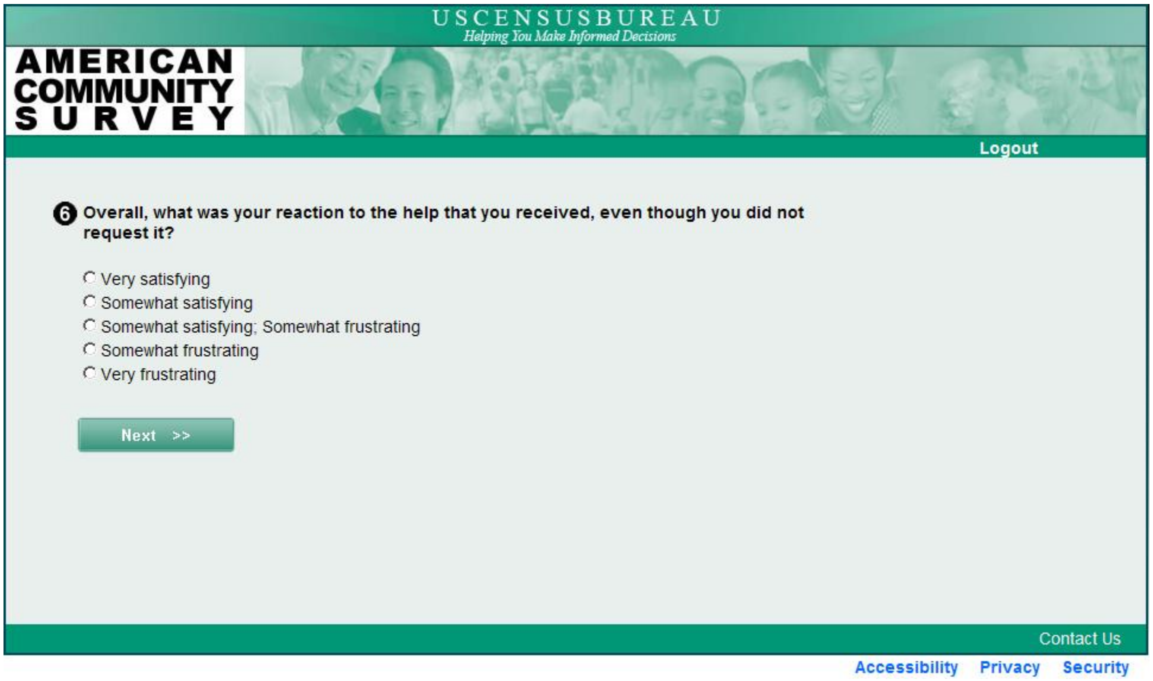
- Participants who selected the first response option received questions 5-9; 12-14
- Participants who selected the second response option received questions 10-14
- Participants who selected the third response option received questions 5-14
- Participants who selected the fourth option received questions 12-14



The next five screens will ask about questions on which you did not request help, but received it automatically.

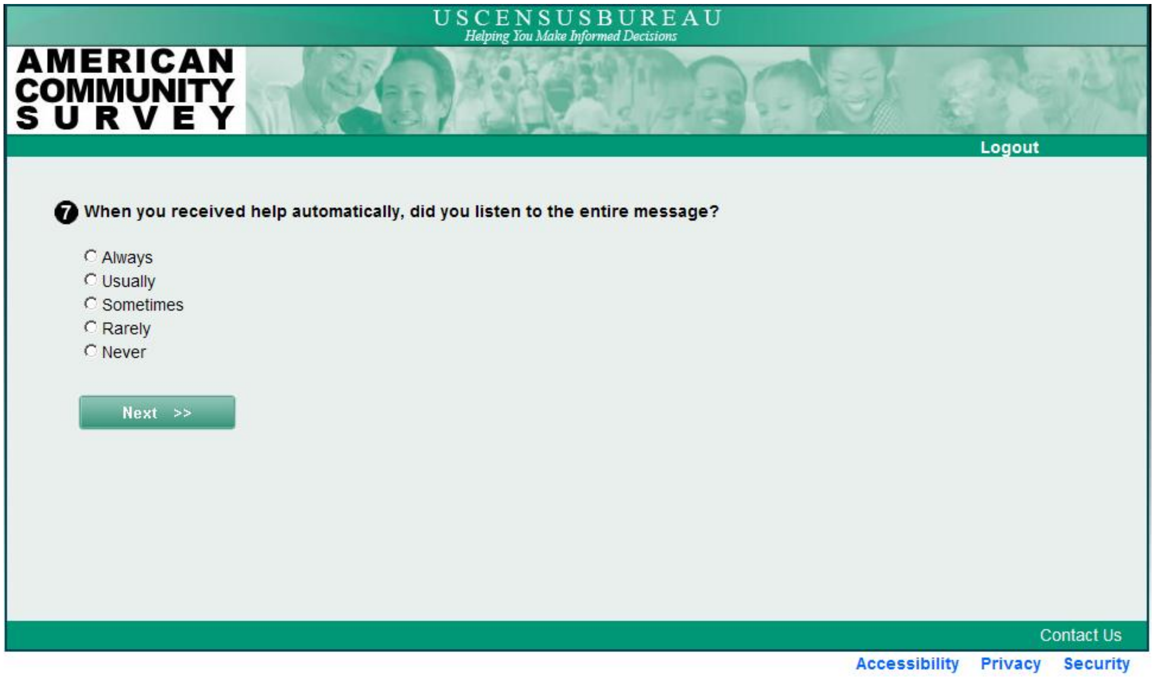
5 Overall, what did you think of the help you received even though you did not request it?

○ Very helpful
○ Somewhat helpful
○ Sometimes helpful; Sometimes unhelpful
○ Somewhat unhelpful
○ Very unhelpful



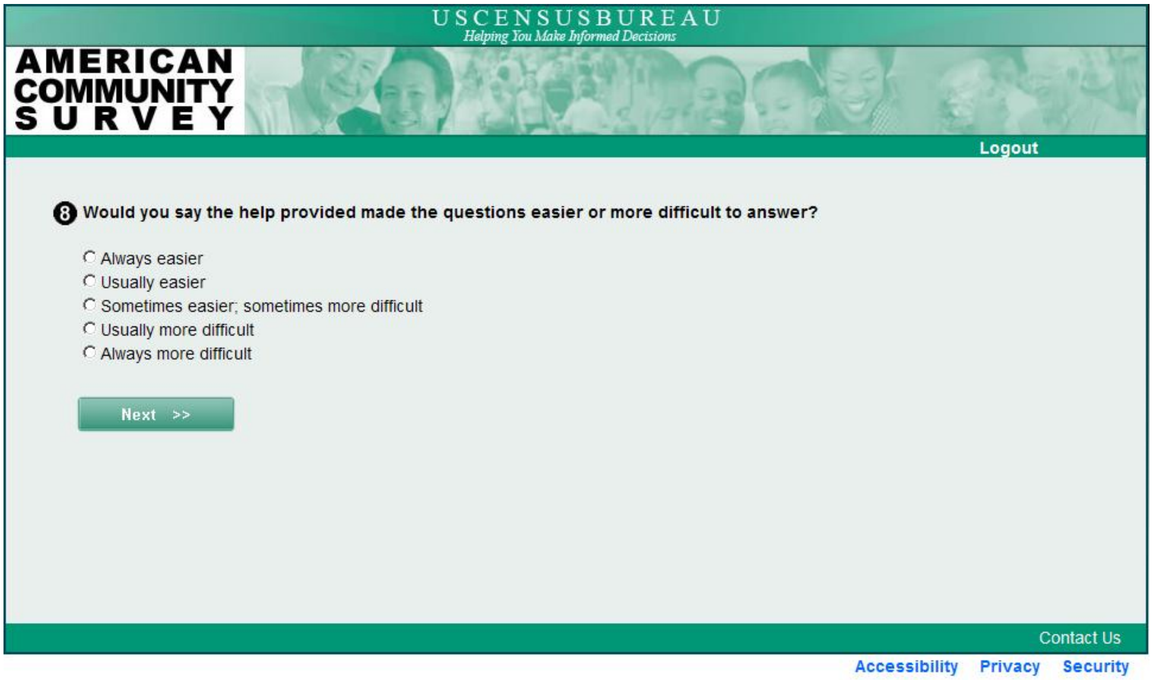
6 Overall, what was your reaction to the help that you received, even though you did not request it?

○ Very satisfying
○ Somewhat satisfying
○ Somewhat satisfying; Somewhat frustrating
○ Somewhat frustrating
○ Very frustrating



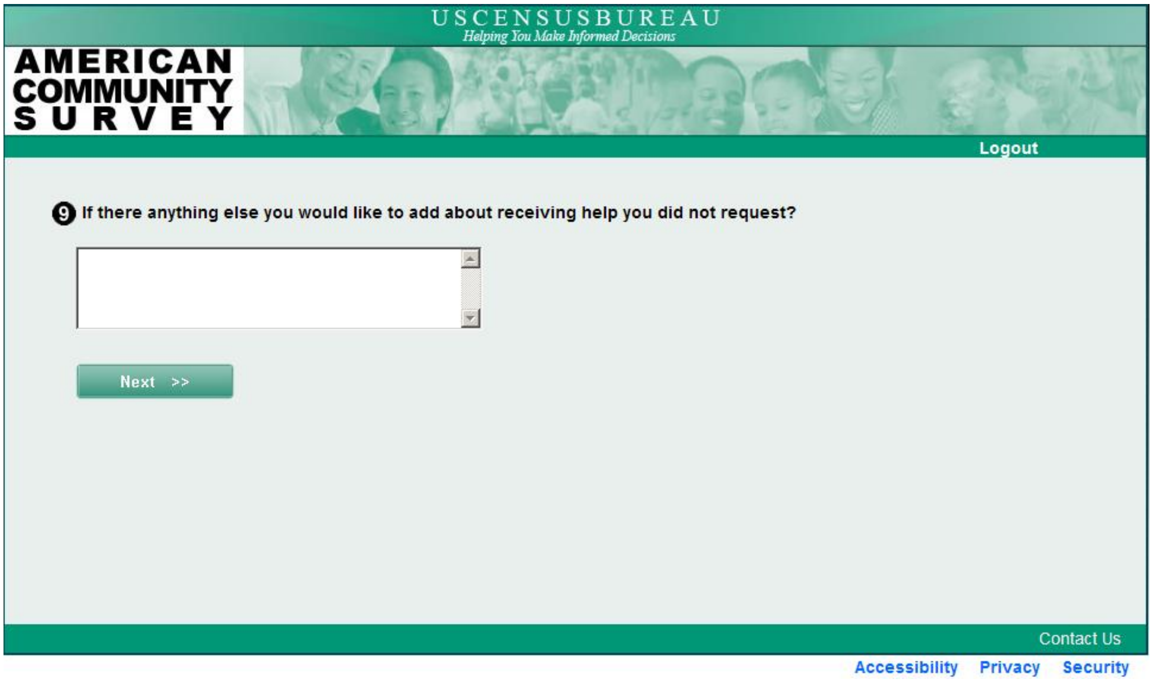
7 When you received help automatically, did you listen to the entire message?

○ Always
○ Usually
○ Sometimes
○ Rarely
○ Never



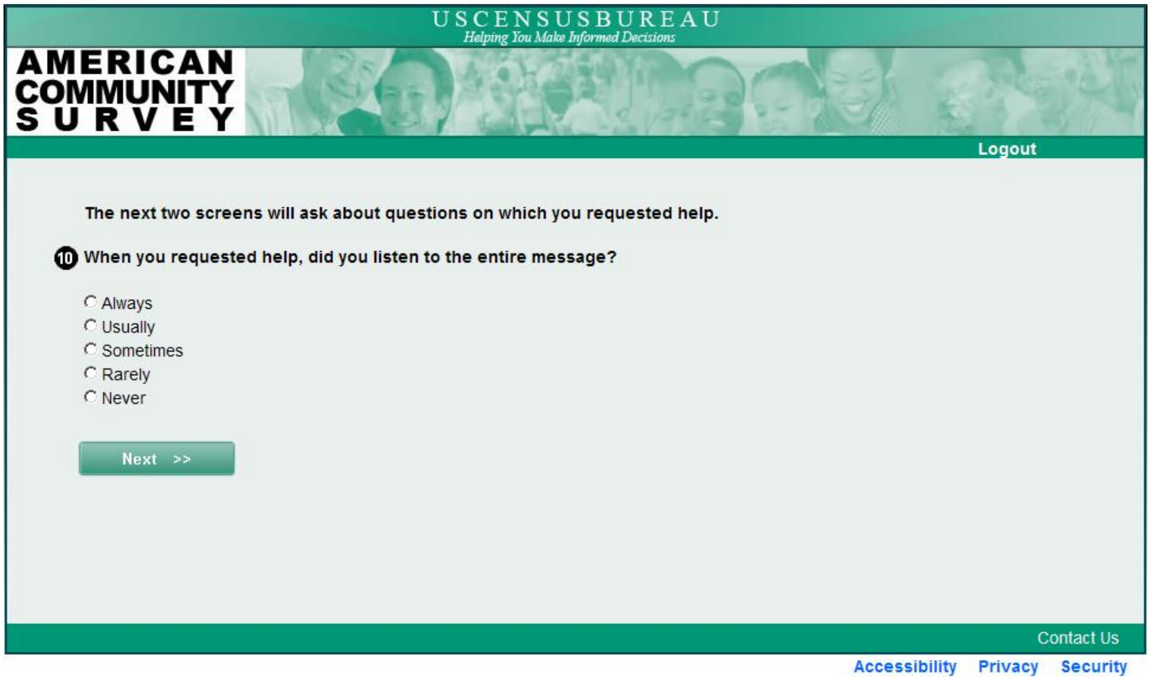
8 Would you say the help provided made the questions easier or more difficult to answer?

○ Always easier
○ Usually easier
○ Sometimes easier; sometimes more difficult
○ Usually more difficult
○ Always more difficult


USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
Contact Us
Accessibility
Privacy
Security

❾ If there anything else you would like to add about receiving help you did not request?

Next >>

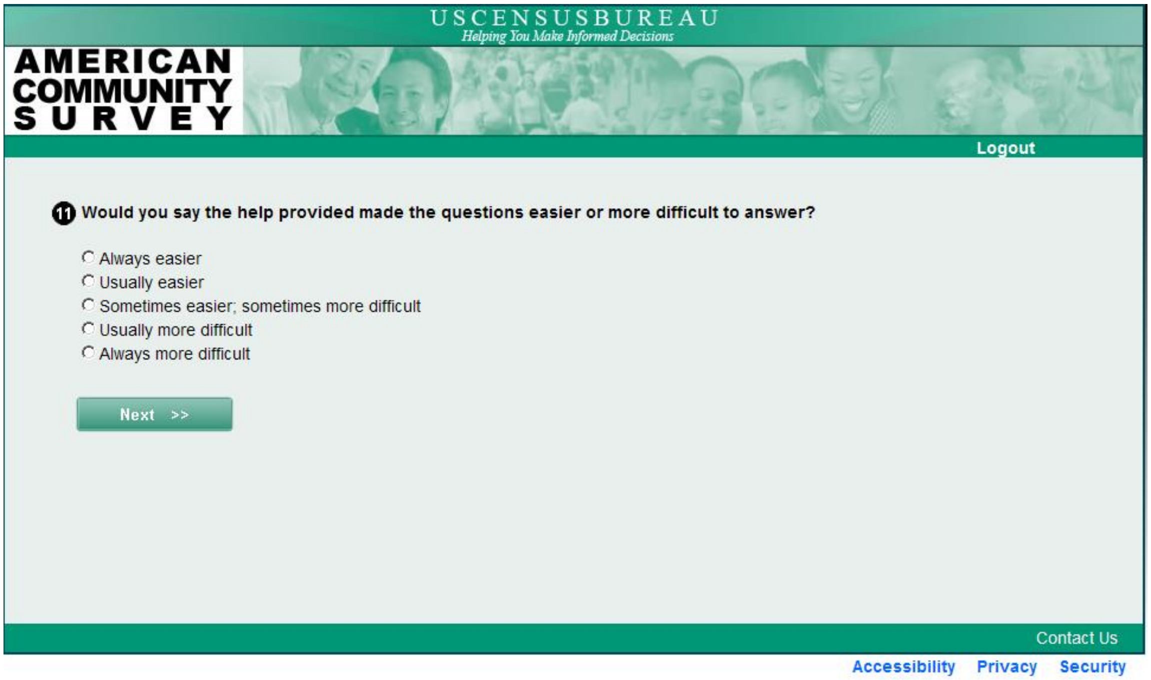

USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout
Contact Us
Accessibility
Privacy
Security

The next two screens will ask about questions on which you requested help.

❿ When you requested help, did you listen to the entire message?

○ Always
○ Usually
○ Sometimes
○ Rarely
○ Never

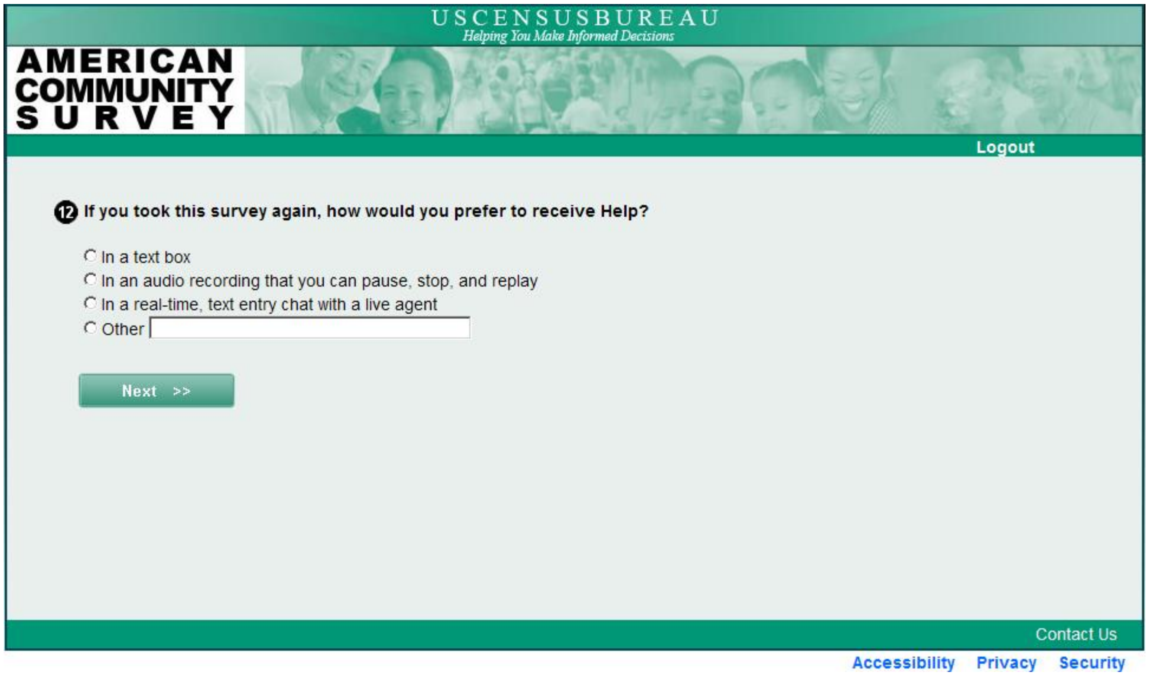
Next >>


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

⓫ Would you say the help provided made the questions easier or more difficult to answer?

○ Always easier
○ Usually easier
○ Sometimes easier; sometimes more difficult
○ Usually more difficult
○ Always more difficult

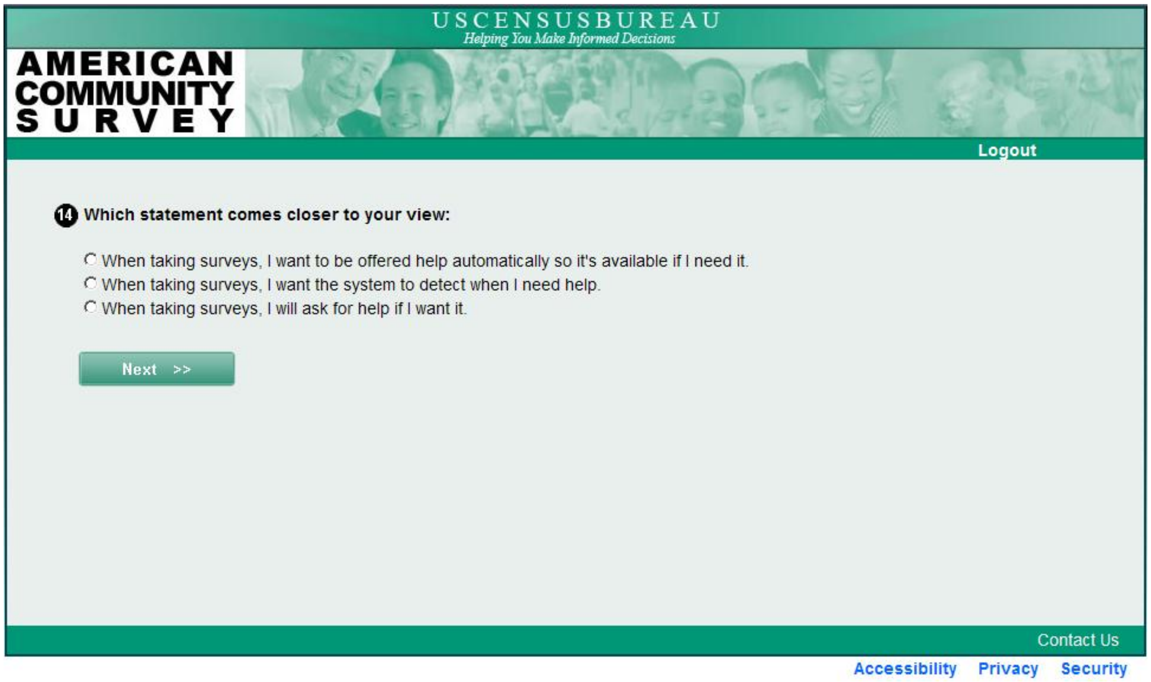
Next >>
Contact Us
Accessibility
Privacy
Security


U S C E N S U S B U R E A U
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

⓬ If you took this survey again, how would you prefer to receive Help?

○ In a text box
○ In an audio recording that you can pause, stop, and replay
○ In a real-time, text entry chat with a live agent
○ Other ______

Next >>
Contact Us
Accessibility
Privacy
Security



USCENSUSBUREAU
Helping You Make Informed Decisions
AMERICAN COMMUNITY SURVEY
Logout

⓮ **Which statement comes closer to your view:**

○ When taking surveys, I want to be offered help automatically so it's available if I need it.
○ When taking surveys, I want the system to detect when I need help.
○ When taking surveys, I will ask for help if I want it.

Appendix Q. Protocol for Study 3

Date_______________________;Participant#________; Experimenter:_______________

General Introduction: Measuring Question Difficulty on the American Community Survey Internet Instrument

Thank you for your time today. My name is Rachel Horwitz and I am a student here in the Joint Program of Survey Methodology and am also employed at the US Census Bureau in the American Community Survey Data Collection area. I will be working with you today. If you have a cell phone, please turn it off or put it in vibrate. In order to help us improve our surveys, we turn to people like you to find out if our questions make sense and are fairly easy to understand and answer. We have found that the best way to do that is to actually conduct the survey with people and see how it works for them. So you will be helping us test a questionnaire from one of our surveys. Since it is difficult to recruit participants with a wide range of experience to come in to take these studies, we will be asking you to read a scenario for each question and to answer the question based on the information you have read, not your own personal experience. I did not create the survey, so please share both your positive and negative reactions to it. The entire session should last 30-45 minutes. Your comments and feedback will be given to the developers of the survey and may be used to improve it.

First, I would like to ask you to read and sign this consent form. It explains the purpose of today's session and informs you of your rights as a participant. It also tells you that we would like to record the session, with your permission. Only those of us connected with the project will review the recording and any other data collected during the session, and it will be used solely for research purposes. We may also use clips from the recording to illustrate key points about the survey to the Web design team.

Hand the participant the consent form; give time to read and sign; sign own name and date if you have not already done so.

Start the tape.
While you are completing the survey, we will record the movements of your eyes with our eye-tracking monitor to get a record of where you are looking on the screen and we will record your mouse movements to see how you are interacting with the survey.
Now I am going to calibrate your eyes for the eye-tracking.

Do Calibration
Now that we have your eyes calibrated, we are ready to begin. Please respond to the survey questions online. Please read each scenario carefully and then answer the questions as they apply to the scenario. To view the scenario, click the "Scenario" link at the top of the question page (*show screenshot of where the link is located*).

Do you have any questions?

Start the eye-tracking software: Tobii Studio. The survey opens in Internet Explorer when Tobii is started. Tell the respondent to answer as if they were home alone. Leave room.

Overall Probe: Make a note if a person left a page with a blank answer, asked a question to the researcher, or displayed signs of confusion (hovers, regressions, using the mouse as a marker).

What was your overall impression of the survey?

Were there any questions you found to be difficult or challenging to answer?

> If yes, show the respondents the relevant questions again and ask them to explain what was confusing to them and what they were thinking about while answering the question.

For any questions where the participant confirmed were difficult or they were unsure of:

> *If Help was not selected,* Did you consider selecting the Help option?
>
> > *If 'Yes'*, Why did you decide not to select Help?
> > *If 'No'*, Why didn't you consider selecting Help?
>
> *If Help was selected,* Was the Help useful in answering the question?

For any participant that did not answer the satisfaction questions accurately, ask them the questions they should have received.

Thank you again for your participation today. It is greatly appreciated.

Appendix R. Diagnosing the Effect of each Participant on Mean Accuracy Measures

We were concerned that our mean accuracy calculations did not account for variations in individual accuracy and on how many questions participants actually received help, which could result in a handful of participants driving the results. To determine whether this was actually happening, we individually removed each participant from the analysis and recalculated the means. Figures R.1 through R.3 show the change in the mean as a result of removing each participant from the analysis.

Figure R.1. Change in Mean Accuracy on Questions Associated with Straightforward and Complex Scenarios after Removing each Participant

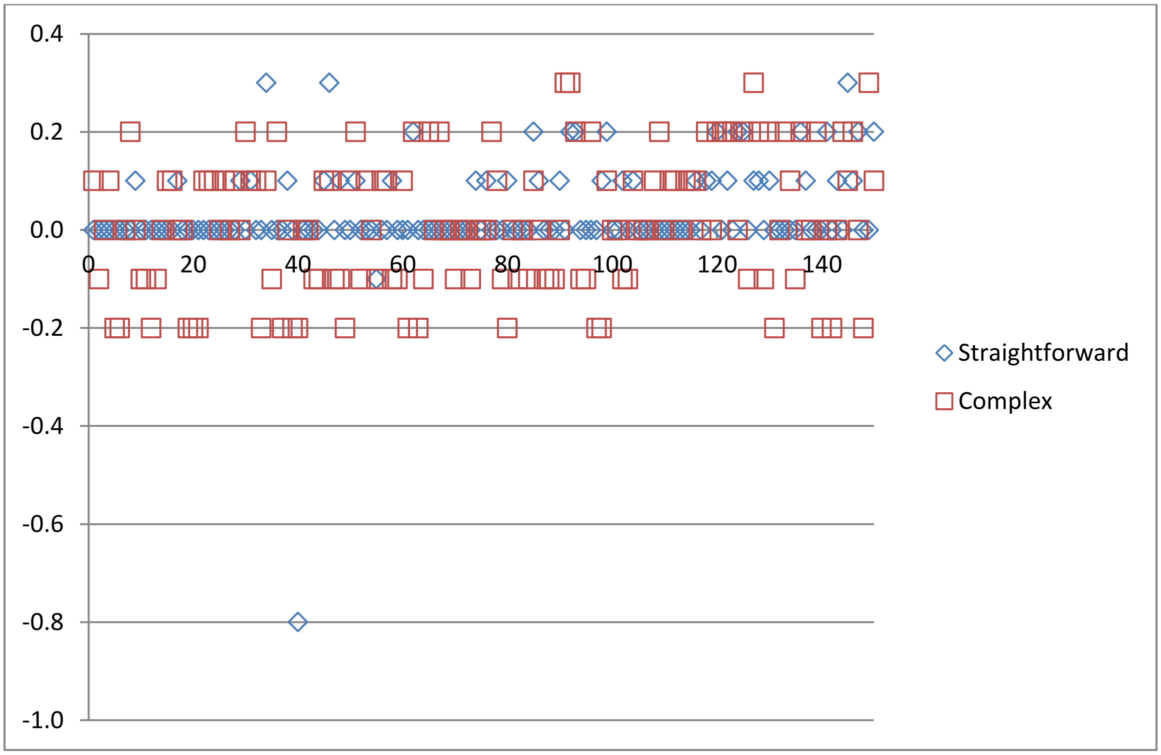


Figure R.2. Change in Mean Accuracy when Help was and was not Received after Removing each Participant

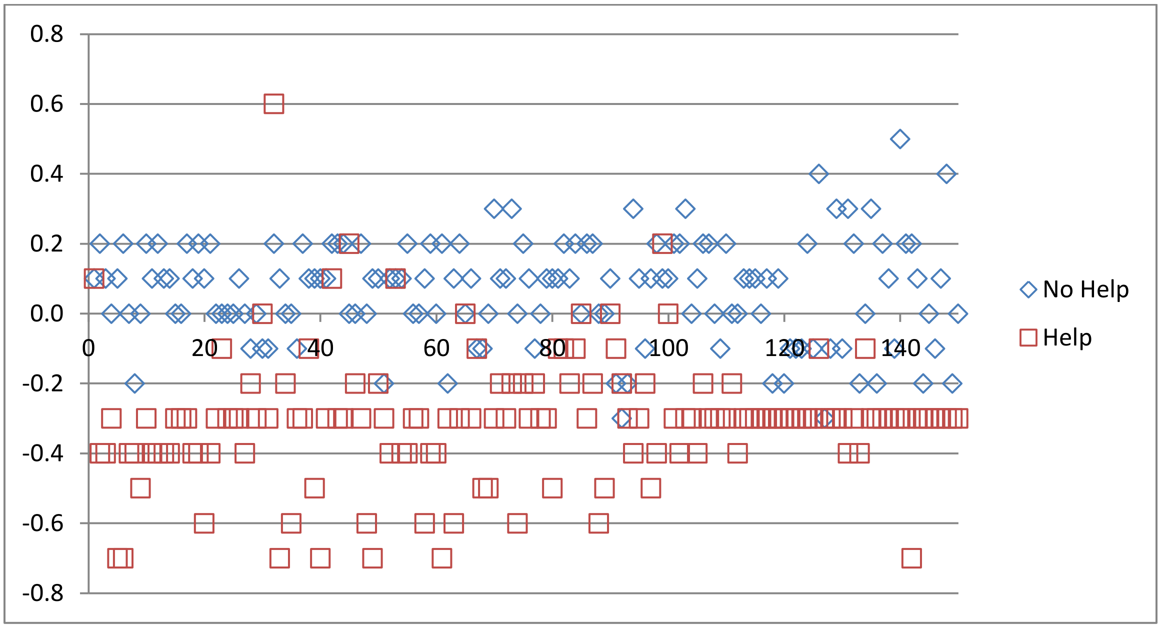
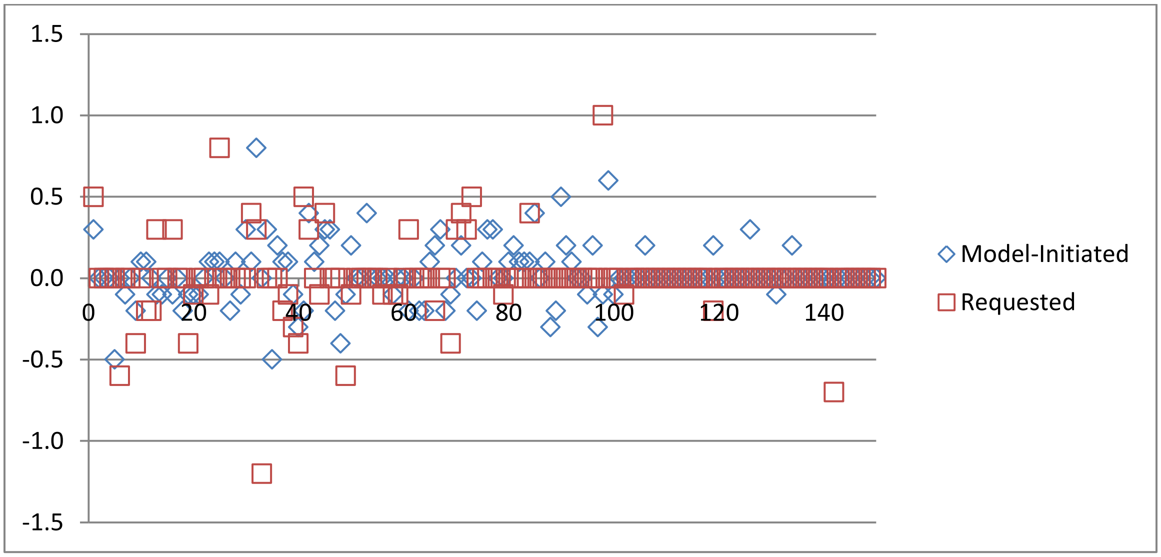


Figure R.3. Change in Mean Accuracy when Help was Requested and Model-Initiated after Removing each Participant



We can see from these figures there is little variation in the mean across observations, suggesting that no single participant is driving the results of the analyses in Section 4.4.2.

Bibliography

Archer, N.P., Wollersheim, J.P., & Yuan, Y. (1996). Investigation of voice and text output modes with abstraction in a computer interface. *Interacting with Computers* 8(4), 323-345.

Arroyo E., Selker, T., & Wei, W. (2006). Usability tool for analysis of web designs using mouse tracks. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 484-489). ACM.

Atkinson, R., Mayer, R., & Merrill, M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, 30(1), 117-139.

Ayers, P. & Sweller, J. (2005). The split-attention effect in multimedia learning. *The Cambridge Handbook of multimedia learning*, (pp. 135-146). Cambridge University Press.

Baddeley, A.D & Hitch, G.J. (1974). Working memory. *The psychology of learning and motivation*, 8, 47-89.

Barlow, S., Converse, S., Kahler, S., Lester, J., & Stone, B. (1997). Animated pedagogical agents and problem solving effectiveness: a large scale empirical evaluation. In *Artificial Intelligence in Education, 1997: Knowledge and Media in Learning Systems: Proceedings of AI-ED 97, World Conference on Artificial Intelligence in Education, Kobe, Japan* (Vol. 39, p. 23). Ios Press Inc.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research. A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, 55(3), 331-346.

Binnick, Y., Westbury, C., & Servan-Schreiber, D. (1989). Case Histories and Shorter Communications. *Behavioral Research Therapy*, 27, 303-306.

Brennan, S.E. & Williams, M. (1995). "The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers." *Journal of Memory and Language*, 34(3), 383-398.

Conrad, F.G., Schober, M.F, & Coiner, T (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology*, 21(2), 165-187.

Conrad, F.G., Couper, M., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics*, 22(2), 245-269.

Conrad, F.G., Couper, M., Tourangeau, R., & Peytchev, A. (2005). Impact of progress feedback on task completion: first impressions matter. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1921-1924). ACM.

Crawford, S.D., Couper, M.P., & Lamias, M.J. (2001). Web surveys: perceptions of burden. *Social Science Computer Review,* 19(2), 146.

Dai, J., Li, Z., & Rocke, D. (2006). Hierarchical logistic regression modeling with SAS GLIMMIX. In *Proceedings of Western Users of SAS Software Conference*. Irvine, CA.

Duchowski, A. (2007). *Eye Tracking Methodology: Theory and Practice* (Vol. 373). Springer-Verlag.

Ehlen, P., Schober, M. F., & Conrad, F. G. (2005). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Processes,* 44(3), 245-265.

Eichenwald, K. (November 2, 1986). "Hi, voter. This is your president." *New York Times, Section 3*, 19.

Fang, X., Xu, S., Brzezinski, J., & Chan, S. (2006). A study of the feasibility and effectiveness of dual-modal information presentations. *International Journal of Human-Computer Interaction,* 20(1), 3-17.

Fazio, R. H. (1990) A practical guide to the use of response latency in social psychological research. In Review of Personality and Social Psychology, vol. 11, *Research Methods in Personality and Social Research* (eds C. Hendrick and M. S. Clark), pp. 74–97. Newbury Park: Sage.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: new insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly,* 72(5), 892-913.

Graesser, A., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question understanding aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly,* 70(1), 3-22.

Graham, S. & Barker, G. (1990). The down side of help: an attributional-developmental analysis of helping behavior as a low-ability cue. *Journal of Educational Psychology,* 82(1), 7-14.

Guo, Q. & Agichtein, E. (2008). Exploring mouse movements for inferring query intent. In *Proceedings of SIGIR Conference on Research and Development in Information Retrieval*. Singapore.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review.* 21(3), 360-373.

Hijikata, Y. (2004). Implicit user profiling for on demand relevance feedback. In *Proceedings of the 9th international conference on Intelligent user interfaces* Funchal, Portugal: ACM.

Hogg, A., & Miller, J. (2003). “Watch out for dropouts.” *cited April 20*, 2005.

Huang, J., White, R., & Dumais, S. (2011). No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the 2011 annual conference on Human factors in computing systems*. Vancouver, BC.

Katunobu, I., Satoru, H., & Hozumi, T. (1992). Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding n-best sentence hypotheses. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference*). Nice, France: IEEE.

Johnston, M. (2007). Automating the survey interview with dynamic multimodal interfaces. *Envisioning the Survey Interview of the Future* (pp. 137-160).

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13(4), 351-371.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1),159-174.

Laurel, B. (1990). Interface agents: metaphors with character. *The Art of Human-Computer Interface Design*, (pp. 355-365). Addison-Wesley.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Los Angeles, CA: ACM.

Leiva, L. A. (2011, May). MouseHints: easing task switching in parallel browsing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. Vancouver, BC: ACM.

Lind, L. H., Schober, M. F., & Conrad, F. G. (2001). Clarifying question meaning in a web-based survey. In *Proceedings of the American Statistical Association*. Alexandria, VA: AMA.

Liversedge, S. & Findlay, J. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*. 4(1), 6-14.

MacElroy, B. (2000). Variables influencing dropout rates in Web-based surveys. *Quirk's Marketing Research Review, Vol. July.(http://www. quirks. com/articles/article. asp.*

Manfreda, K. L., & Vehovar, V. (2002). Survey design features influencing response rates in web surveys. In *Proceedings from the International Conference on Improving Surveys*. Copenhagen, Denmark.

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, 95(2), 419-425.

Miller, C. (2004). Human Computer Etiquette: Managing expectations with intentional agents. *Communications of the ACM,* 47(4), 31-34.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits in our capacity for processing information. *Psychological Review*, 63(2), 81-97.

Mishra, P. & Hershey, K. (2004). Etiquette and the design of educational technology. *Communications of the ACM,* 47(4), 45-49.

Moreno, R., Mayer, R.E., Spires, H.A., & Lester, J.C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, 19(2), 177-213.

Mueller, F., & Lockerd, A. (2001, March). Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI'01 extended abstracts on Human factors in computing systems* (pp. 279-280). ACM.

Nass, C. (2004). Etiquette and equality: exhibitions and expectations of computer politeness. *Communications of the ACM,* 47(4), 35-37.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology,* 29(5), 1093-1110.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004, March). The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. San Antonio, TX: ACM.

Penney, C. G. (1989). Modality effects and the structure of short-term memory. *Memory and Cognition*, 17(4), 398-422.

Person, N. K., D'Mello, S., & Olney, A. (2007). Toward Socially Intelligent Interviewing Systems. *Envisioning the Survey Interview of the Future, 542*, 195.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly,* 73(1), 74-97.

Rayner, K. (1992). *Eye movements and visual cognition: Scene perception and reading*. Springer-Verlag.

Rayner, K. (1983). *Eye movements in reading: perceptual and language processes*. Academic Press Professional, Inc.

Redline, C., Tourangeau, R., Couper, M., Conrad, F., & Ye, C. (2009). The effects of grouping response options in factual questions with many options. In *Annual Conference of the Federal Committee on Statistical Methodology. Available at: http://www. fcsm. gov/09papers/Redline_IX-B.pdf.*

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Chicago, IL, US: Center for the Study of Language and Information; New York, NY: Cambridge University Press.

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *CHI'08 extended abstracts on Human factors in computing systems* (pp. 2997-3002). ACM.

Schober, M. F. & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *The Public Opinion Quarterly*, 61(4), 576-602.

Schober, M., Conrad, F., Ehlen, P., & Fricker, S. (2003). How web surveys differ from other kinds of user interfaces. In *Proceedings of the American Statistical Association*. Alexandria, VA: ASA.

Schober, M. F. & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes*, 38(3), 287-308.

Shneiderman, B. (1987). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley.

Smith, V.L. & Clark, H.H. (1993). On the course of answering questions. *Journal of Memory and Language*, 33,25-38.

Sproull, L., Subramani, M., Kiesler, S., Walker, B., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, 11(2), 97-124.

Suessbrick, A., Schober, M. F., & Conrad, F. G. (2005). When do respondent misconceptions lead to survey response error? In *Proceedings of the American Statistical Association*. Alexandria, VA: American Statistical Association.

Tancreto, J.G., Davis, M.C., & Zelenak, M.F. (2012). Design of the American Community Survey Internet Instrument. U.S. Census Bureau. Available at: http://www.census.gov/ acs/www/Downloads/library/2012/2012_Tancreto_02.pdf

Tatler, B. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 1-17.

Tourangeau, R., Conrad, F., Arens, Z., Fricker, S., Lee, S., & Smith, E. (2006). Everyday concepts and classification errors: judgments of disability and residence. *Journal of Official Statistics*, 22(3), 385-418.

Tourangeau, R., Couper, M., & Steiger, D. (2003). Humanizing self-administered surveys: experiments on social presence in web and IVR surveys. *Computers in Human Behavior*, 19(1), 1-24.

Tourangeau, R., Rips L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* New York*:* Cambridge University Press.