Optimal Data Entry Designs in Mobile Web Surveys for Older Adults

Erica Olmsted-Hawala[1(✉)], Elizabeth Nichols[1], Brian Falcone[1], Ivonne J. Figueroa[2], Christopher Antoun[3], and Lin Wang[1]

[1] U.S. Census Bureau, Washington D.C., USA
{erica.l.olmsted.hawala, elizabeth.may.nichols, brian.falcone, lin.wang}@census.gov
[2] HCSC Blue Cross Blue Shield, Chicago, IL, USA
Ivonne_j_figueroa@bcbsil.com
[3] University of Maryland, College Park, MD, USA
antoun@umd.edu

Abstract. Growing numbers of people are using their mobile phones to respond to online surveys. As a result, survey designers face the challenge of displaying questions and their response options and navigation elements on small smartphone screens in a way that encourages survey completion. The purpose of the present study was to conduct a series of systematic assessments of how older adults using smartphones interact with different user-interface features in online surveys. This paper shares results of three different experiments. Experiment 1 compares different ways of displaying choose-one response options. Experiment 2 compares different ways of displaying numeric entry boxes, specifically ones used to collect currency information (e.g., prices, costs, salaries). Experiment 3 tests whether forward and backward navigational buttons on a smartphone survey should be labeled with words *(previous, next)* or simply indicated with arrow icons (<, >). Results indicate that certain features such as picker-boxes that appear at the bottom of the screen (iOS devices), fixed formatting of numeric-entry boxes, and icon navigation buttons were problematic. They either had negative impacts on performance (response times and/or accuracy) or only a small percentage of participants preferred these design features when asked to compare them to the other features.

Keywords: Mobile survey design · Mobile guidelines · Older adults
Drop-downs · Currency inputs · Mobile navigation controls

1 Introduction

More and more often people are using smartphones to interact with the electronic world [1]. In February of 2018, 77% of all U.S. adults had a smartphone and 46% of U.S. adults over the age of 65 years old had a smartphone [1]. Where in the past people may have waited until they were in front of their desktop PCs to conduct a search, fill out a form, or answer a survey, adults "on the go" are increasingly using their smartphones for such activities. While the majority of adults answer internet surveys on their PC's

J. Zhou and G. Salvendy (Eds.): ITAP 2018, LNCS 10926, pp. 335–354, 2018.
https://doi.org/10.1007/978-3-319-92034-4_26

there are some indications that responding to a survey on a smartphone is on the rise. For example, the American Community Survey, an ongoing monthly U.S. survey, has seen a steady increase in mobile respondents since 2011 when it was just under one percent through today where it is just under eight percent [2]. For private-sector surveys, almost one third of survey responses occur on mobile phones [3, 4]. It is likely that adults responding to internet surveys while on mobile phones will continue to increase over time.

At the U.S. Census Bureau, as at other survey organizations, we are interested in developing mobile web surveys that reduce measurement error while also improving the user experience. With the smaller screen real estate of the smartphone, the user interface must be adapted for the smaller space. Yet, the small amount of touchable space available on the screens of smartphones can be challenging for both survey respondents and for developers of online surveys. Evidence suggests that mobile surveys lead to lower response rates because respondents break off (i.e., don't finish the survey) as well as longer survey completion times [5, 6]. For a review on the impacts of using mobile phones to answer online survey, see [7].

An additional challenge of creating mobile web surveys is that many different age groups are now using smartphones [1]. While the older adult population is more resistant to new technology generally [8] they are using smartphones in their daily lives [9, 10] and they too need to be accommodated in the design of online surveys.

At present, there has been little empirical research on how to best design surveys on smartphones for older adults. Current literature is typically focused on the general population and not specifically for surveys [11, 12]. There is evidence that for touch screens, older adults do better with larger buttons, but this study was on kiosk-type touch screens, not on the smaller display of smartphones [13]. Within the healthcare field, the use of mobile phones by older adults to aid in managing home health has begun but is not fully tested for its effectiveness or usefulness [14]. In fact, there is some evidence that the designs of the mobile health applications cause barriers to the older adult population in terms of uptake and use [15].

It is possible that older adults may interact with smartphones differently than younger adults. For example, research shows that accurately touching a target takes longer for older adults than for younger adults, commonly referred to as the tradeoff between speed and accuracy [16, 17]. In addition, as adults age, the sensory changes with respect to touch and vision can impact what older adults are able see and touch when interacting with a small screen that contains a variety of information [18]. Literature has shown that older adults generally have reduced vision, mobility, and certain cognitive capacity such as memory, compared to younger adults [19–21].

The purpose of the present study was to conduct a series of systematic assessments on a mobile phone to determine how older adults use different user-interface designs to answer online survey questions and to identify better performing and preferred designs. The results of these assessments could be used as guidelines for developers. Our rationale was that if we develop guidelines for a mobile web survey interface that older adults can successfully complete, then younger adults would do at least as well because of their superior perceptual and motor capability. The initial impetus for the work was based on observations made while participants used mobile phones to fill out surveys during earlier, unrelated usability tests.

The rest of the paper lays out the methods we used and the specifics of each of the three experiments including hypothesis, results and conclusions.

2 Methods

In this paper we discuss results from three different experiments from a larger ongoing research study that includes multiple experiments aimed at establishing a set of mobile web survey guidelines for developers. For more information on this entire research project please see [22]. Below are highlights of methods relevant to the three experiments described in this paper.

2.1 Sample

We aimed to get a study sample of persons aged 60–75. We prescreened to include only participants who had at least 12 months of experience using a smartphone under the assumption that these participants were more typical of respondents who choose to use mobile devices to complete online surveys than those with less experience using smartphones. Additionally, we prescreened participants to include only individuals who had an education of 8th grade or more, who were fluent in English, and who had normal vision or corrected to normal with glasses or contacts. The participants were a convenience sample recruited from senior and/or community centers in and around the Washington DC metropolitan area between November 2016 and February 2017.

Experiment 1 was conducted with a pool of 30 participants, and Experiments 2 and 3 were conducted on a different pool of 32 respondents. Participants in each pool reported an average of familiarity with using the smartphone of 3 on a 5-point scale where 1 was "Not at all familiar" and 5 was "Extremely familiar." See Table 1. We consider significance to be at $p = 0.05$ or less.

2.2 Data Collection

One-on-one sessions were conducted at senior/community centers. For each session a given participant completed between 4 to 6 experiments, only some of which are the subject of this paper. Each experiment was run at a "station" with a different Census Bureau staff member (i.e., test administrator (TA)) manning the station. As participants were recruited, the first station's TA explained the purpose of the testing, had the subject sign a consent form, conducted the prescreening, and assigned the participant a unique ID number. Then the participant went to the next station where another TA worked one-on-one with the participant to complete one or two experiments. Once the participant finished the experiment(s) at one station he/she moved to the next station where a different TA worked with him/her on the next experiment. Each experiment took about 10 min to complete. At the end of the session, the participant was given $40 for their time.

Table 1. Participant demographics for 3 experiments

Experiment	Average age (Standard Error (SE))	Gender (Male/Female)	Smartphone usage [1 – Not Familiar to 5 – Extremely familiar] (SE)
Experiment 1 (n = 30)	68.8 (0.87)	10M/20F	3.96 (0.17)
Experiments 2 and 3 (n = 32)	70.5 (0.79)	7M/25F	3.56 (0.18)

The experiments were preloaded as applications (commonly referred to as apps) on Census-owned iPhone 5S. Each station had its own iPhone. The TA opened the experiment, entered the participant's unique ID and based on that ID selected the condition to administer. For each experiment, conditions were pre-assigned to IDs using a randomized (or quasi[1]-randomized) order. From there, participants were handed the iPhone with the app loaded to the correct starting location and the TA followed the individual protocol for that experiment. This included instructing participants not to talk aloud during the session, and to complete the survey to the best of their ability as though they were answering the survey at home without anyone's assistance. All three experiments were video-recorded using QuickTime with the phone plugged into a MacBook laptop.

3 Experiment 1: "Choose-One" Response Option Design

3.1 Designs Tested in the Experiment

Experiment 1 focused on "choose-one" questions. Survey designers have a number of options when designing for a question with a set number of answer choices where the user is told to choose only one answer. The most common design is a "radio button" design where the response choices are on the same screen as the question itself, as shown in Fig. 1. For that design users answer the question by touching the appropriate response choice on the screen. They can change their answer by touching another response choice. Another response-option design solution is to use an "open-text" field as shown in Fig. 2. When a respondent answers these questions, touching the open-text field brings up the character keyboard or numeric keypad and then the respondent can enter the response as shown in Fig. 3. That design is typically used when the answer is easily typed, such as a number, or when the set of answers is so large that it would be unwieldy to place them all on the screen, such as street names.

[1] Prior to collecting data, a random assignment computer algorithm was used to assign conditions for each experiment. A few of the assignments in Experiment 1 were manually manipulated so there were an equal number of participants assigned to each condition.

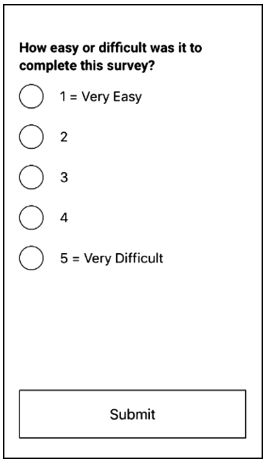

Fig. 1. Radio button design

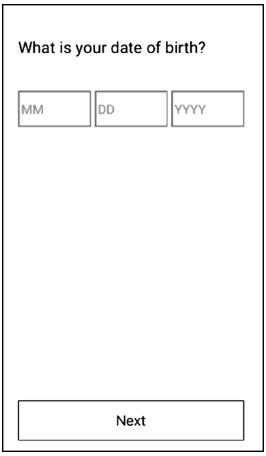

Fig. 2. Open-text field design

Fig. 3. Open-text field w/keypad

A third design solution for choose-one questions is a "dropdown" format. Dropdowns are often implemented when there is limited space on the screen or when there is a long list of response options and the response options are well known, like the list of states in the U.S. On mobile webpages, the default dropdowns display differently depending upon the operating system. For both operating systems, the user must first touch the dropdown field to see the choices, Fig. 4 shows what the screen looks like for both operating systems before the user taps the response box. Figure 5 shows what happens in the iOS when the user taps the response box: the list displays in grey at the bottom of the screen, and is called a "picker," Fig. 6 shows what the screen looks like on the Android, when what is called a "spinner" opens and displays a view more similar to a PC dropdown, with a list of choices displaying over the screen. Once a selection is made, the answer choice appears in the dropdown field and the other choices disappear as shown in Fig. 7. Dropdown designs are quite different from radio button designs. With dropdowns, the user will not know the available response choices until he or she "opens" the dropdown; with radio buttons, the user does not have to do anything to see the answers – they are already displayed on the screen.

In Experiment 1, we compared three different designs for choose-one questions using a 12-question survey and a between-subjects design. The three conditions were the iOS picker (Fig. 5); the Android spinner (Fig. 6), and a radio button/keyboard design (Figs. 1 and 2).

We hypothesized that the iOS picker design would cause more difficulties for users as compared to the other two designs because the response options (Fig. 5) appear in gray font at the bottom of the screen and are easily missed.

Fig. 4. Initial view

Fig. 5. iOS picker

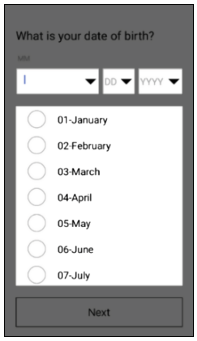

Fig. 6. Android spinner

Fig. 7. Final view

3.2 Adaptation of Survey Questions with the Three Alternative Designs

A set of 12 questions on a range of topics was selected and an app was created that displayed the 12 questions in each of the three formats described above. Participants were randomly assigned to one of the three formats, and there were 10 participants in each condition. Each participant completed the 12-question survey in the assigned condition. The 12 "choose-one" questions included some with a small number of response options (i.e., 5 or fewer) and some questions with a large number of response options (i.e., more than 5). Some questions had familiar responses in the sense that the respondent could probably predict the response options based on the survey question (e.g., question about a respondent's sex), some had ordinal responses (e.g., age categories), while other questions had response options that a respondent would probably not know prior to reading through them. We varied the response option types to be able to control for the type of question, in case particular types of "choose-one" questions performed better in one design compared to another. Table 2 provides the questions with their response option characteristics.

For the radio button/keyboard design condition, questions 2–5 and 7–12 used radio buttons. Question 6 was an open-text field that brought up a keyboard when the participant touched the field. The first question, date of birth, was also an open-text field, which when touched, brought up a keypad as shown in Fig. 4 above. We used open-text fields for those two questions because in practice, survey designers rarely, if ever, use radio buttons for states or dates. For the iOS and Android dropdown conditions, dropdowns were used for all 12 questions.

Each question was on a separate screen with forward and backward navigation buttons in a fixed location at the bottom of each screen. After completing the survey, the participant then answered a satisfaction question. The satisfaction question asked, "How easy or difficult was it to complete this survey?" with a rating scale from 1 to 5 where 1 was defined as "Very Easy" and 5 was defined as "Very difficult." Finally, the respondent interacted with a date of birth question using each of the designs – first the picker, then the spinner,

and then open-text field using the number keypad. After interacting with the date of birth question with the three designs, respondents answered a preference question that collected the respondent's design preference for that question.

Table 2. Question and question characteristic

Questions 1–6	Questions 7–12
1. Date of birth (Familiar and >5 choices)	7. Citizen of more than one country (Familiar and <=5 choices)
2. Age range (Ordered and >5 choices)	8. Fuel for heating home (Unique and >5 choices)
3. Sex (Familiar and <=5 choices)	9. Eyesight rating (Ordered and <=5 choices)
4. Marital status (Unique and <=5 choices)	10. Work status (Unique and >5 choices)
5. School level obtained (Ordered and >5 choices)	11. Opinion question (Ordered and <=5 choices)
6. State attended high school (Familiar and >5 choices)	12. Preference for reporting (Unique and <=5 choices)

3.3 Evaluation Criteria

For each condition, we measured respondent burden (operationalized as time-on-task and the number of touches on each screen of the survey); accuracy of data entries (by comparing any discrepancies between the entered data and data provided to a screening paper questionnaire administered prior to the mobile phone survey); satisfaction and preference by the responses provided within the experiment. We then compared these measures between conditions.

We modeled time to complete at the question level using a mixed model. Modeling at the question level increases the number of observations from 30 to 30 × 12 or 360 and allows us to account for different question characteristics. In the model, we controlled for the condition, and the characteristics of the question as outlined in Table 2 above, and any interaction between condition and those characteristics. To control for any participant effect because each participant would contribute up to 12 times (one time for each question), we included a random effect for the participant. We also modeled the log of time because the residuals from the first model were slightly skewed. As a check we also modeled time with controlling for the question number instead of the question characteristics.

We modeled the number of touches on the screen in the same manner, but without the log transformation.

Because of an error in the app, we did not collect data for one radio button/keyboard design participant and only partial data were saved for another participant assigned to the Android condition. In total, we had 344 observations for each model instead of the expected 360.

We had self-reported measures of sex, age range, date of birth for month and year, and education from the demographic information collected via a paper questionnaire at the beginning of the one-hour session. We compared that data, which we considered truth, to the data reported within the experimental survey. Any survey data that matched was considered accurate; and data that did not match was considered an error. Based on that assignment, we tabulated the accuracy rate for the four questions for each condition. We tabulated satisfaction scores for each of the three conditions. For these analyses, we conducted a Chi-square test of independence. And, finally we tabulated the preference data for the date of birth question. Again, because of missing data, we only collected data from 28 participants; we were missing these data from one radio button/keyboard condition and one Android condition.

3.4 Results

Respondent Burden as Measured by Time to Complete. The average time to complete a question using the iOS picker condition was nearly 21 s (Standard Error (*SE*) = 1.3), compared to 15 s (*SE* = 1.0) for the Android spinner, and 13 s (*SE* = 1.6) for the radio button/keyboard entry. Modeling time to complete, we found that questions using the iOS picker design took significantly longer on average to answer than the radio button/keyboard design ($p = 0.02$) while we did not find a difference in time to complete for questions using the Android spinner design compared to the radio button/keyboard design ($p = 0.60$). There were no significant interactions between the condition and the question characteristics. When modeling the log of time, the pattern of results was unchanged. When modeling time with the question number instead of the question characteristics, the pattern of results was unchanged.

Respondent Burden as Measured by the Number of Touches per Screen. The average number of touches per question for the iOS picker condition was 6.5 (*SE* = 0.3), compared to 3.5 (*SE* = 0.1) for the Android spinner, and 2.6 (*SE* = 0.2) for the radio button/keyboard entry. Modeling the number of touches needed to answer the question without any interactions, we found that questions using the iOS picker design required significantly more touches to answer than the radio button/text design ($p < 0.01$) and the Android spinner design took significantly more touches to select an answer than the radio button/keyboard design ($p < 0.01$). However, when interaction terms between the condition and the question characteristics were added, there was a significant interaction between the conditions and question characteristics ($p < .01$). The effect of the iOS picker design on the number of touches per question was particularly large for questions that had many response options.

Accuracy of Responses by Condition. We found no significant differences in accurate reporting by condition. The accuracy rate for all 28 participants was 100% for each condition for the sex question and the age range question. For date of birth, the accuracy rate was 100% for the radio button/keyboard design; 89% for the iOS design and 78% for the Android design. The iOS and Android conditions had 100% accuracy for education, but the radio button design's accuracy rate was 67% for that field. Even with these

differences, there was no significant difference in accuracy rates for date of birth (χ^2 = 2.3, p = 0.3, n = 27) or education (χ^2 =7.1, p = 0.1, n = 28) by condition.

Satisfaction Scores by Condition and Response Option Preference. Satisfaction was measured on a 5-point scale where 1 was very easy and 5 was very difficult. The average satisfaction score was 1.3 (SE = 0.3) for the iOS picker; 1.1 (SE = 0.1) for the Android spinner; and 1.0 (SE = 0) for the radio button/keyboard design. We found no differences in satisfaction scores by condition (χ^2 = 3.98, p = 0.4) and with the exception of one participant who rated the iOS picker as difficult, the participants found the designs easy to use. However, once participants were able to use each of the designs, they overwhelmingly preferred the keypad design for the date of birth question, with 22 of the 28 participants selecting only that design as their preferred response option design. Their preference was not based on the design they used during the main portion of the experiment (χ^2 = 3.1, p = 0.8).

4 Experiment 2: Layout of Currency Fields

4.1 Designs Tested in the Experiment

In Experiment 2 we investigate alignment and formatting of currency fields on mobile devices. Surveys that ask for monetary information and online banking apps vary in both of these aspects. Part of this research was inspired by what we had seen when respondents were answering questions that included monetary amounts on the American Community Survey. During usability studies participants attempted to add in the dollar sign and decimal place even though it already appeared on the screen.

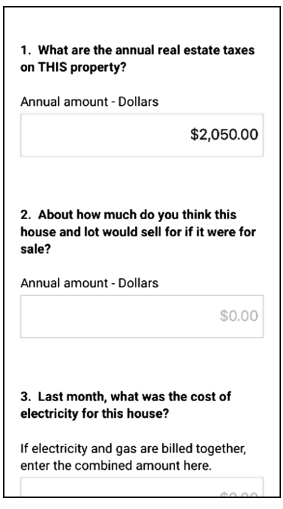

Fig. 8. Right alignment

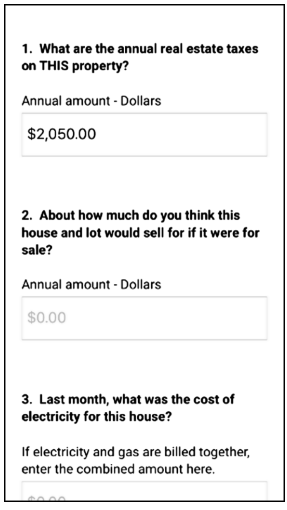

Fig. 9. Left alignment

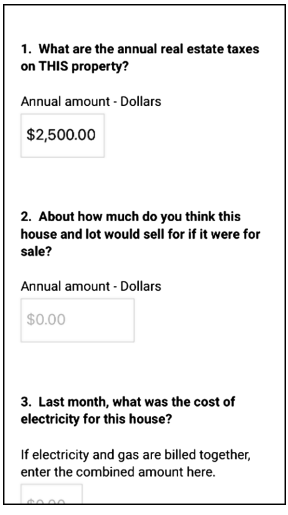

Fig. 10. Center alignment

One aspect of response options for currency data is 'alignment' – that is, where in the response field the numbers appear once the respondent begins entering the numbers. For this study, we chose three variations in alignment to test. First is what we call right alignment, where currency data are entered into a response field with the numbers coming in on the right, like the numbers on most calculator displays. See Fig. 8. Second is what we call left alignment where the numbers representing currency amounts are treated more like text, coming in from the left. See Fig. 9. Third is what we called center alignment where the field itself gets longer or shorter based on the number of digits the respondent enters, such as on apps like Cash© or Paypal©. See Fig. 10 for an example of center alignment. Our hypothesis was that left-alignment would not perform as well as the other alignment types because in earlier usability studies we had noticed users miss the cents display when currency was left-aligned.

Another aspect of the response option for currency data is "formatting" by which we mean the way the currency cues (e.g., dollar sign and cents, including the decimal point) are displayed on the screen, either fixed and always present on the screen or where the application itself is programmed to react to the users' data entry.

We examined three different alternatives for formatting of currency fields. First is what we called the fixed formatting, when the dollar and cents symbols are fixed in place and always present in the field. See Fig. 11. Second is what we call post-entry formatting where formatting occurs only after the user has entered the number in the field; this is indicated when the respondent taps "Done" on the keypad. At that point, the program rounds to the nearest dollar and enters (.00) and ($) into the field. In Fig. 12, the amount in the field at the top of the screen is what was shown after the participant selected "Done," and the amount in the field at the bottom is what was shown as the participant touches the numbers on the keypad, prior to selecting "Done."

Fig. 11. Fixed ($) and (.00) permanently on screen

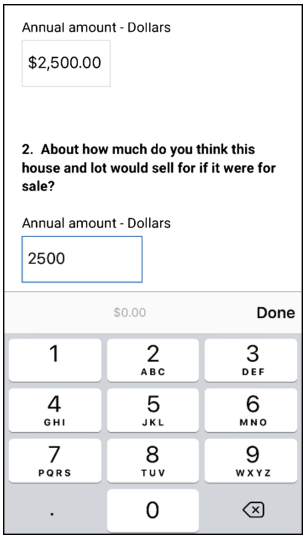

Fig. 12. Post-entry formatting

And finally the third formatting type is what we called automatic formatting where a dollar sign and/or the cents symbol automatically appear in real time as the user enters the currency numbers. (Note, there is no figure image example of this as it would have to be a video.)

Our hypothesis was that the fixed formatted design would cause more problems for users. The rationale for this was that we had noticed in earlier usability testing studies that respondents often fail to notice the static dollar sign and decimal place for cents and consequently attempted to add that data in. The alternate designs that we tested used some form of automatic or real time formatting that we hypothesized respondents would notice more readily.

4.2 Adaptation of Survey Questions with Alternative Alignment and Formatting of Monetary Fields

The two design elements (alignment and formatting) each had three versions, and we fully crossed them in a 3-by-3 design We chose questions for this experiment that are asked in the American Community Survey.

The five questions included the following:

- Cost of real estate property taxes;
- Cost of the lot and house;
- Cost of electricity for the previous month;
- Annual cost of water and sewer for the house and
- Cost of the gross annual income.

Using a within-subjects design, participants were presented with all nine conditions in counterbalanced and randomized order. See Table 3.

Each condition required that the participant enter currency for five questions (for a total of 45 trials). Each condition had the same five questions and for each condition the five questions appeared on the same screen. This required the participant to scroll to answer all five questions. For example, Fig. 8 above shows the first three questions in Condition 5 (e.g., right alignment with post-entry formatting of the ($) and (.00)). Figure 11 on the other hand shows the first three questions in Condition 1 (e.g., left alignment, fixed formatting). The fourth and fifth question require scrolling and so are below the fold of the screen.

Table 3. Experiment 2 properties of each condition

Condition	Alignment	Format of ($) and (.00)	# of questions on screen
1	Left	Fixed	5
2	Right	Fixed	5
3	Center	Fixed	5
4	Left	Post entry	5
5	Right	Post entry	5
6	Center	Post entry	5
7	Left	Automatic	5
8	Right	Automatic	5
9	Center	Automatic	5

For each condition participants were provided identical pieces of paper, mocked up as an actual bill, with the exact amounts to be entered for each question. The amounts were formatted with common features such as commas, cents, and dollar signs. Regardless of condition, the keyboard that popped open allowed users to manually enter the numbers and period but not the dollar sign or commas.

4.3 Evaluation Criteria

Following each condition, participants rated the ease of entering currency using a 5-point Likert scale where 1 was "very easy" and 5 was "very difficult." Finally, the participant was asked which of the three different types of formatting (fixed, post-entry, auto-formatting) they preferred. We examined the difference in satisfaction, accuracy, and respondent burden as measured by time-on-task, accuracy, satisfaction and subjective preference for participants on the currency data tasks for each condition.

A repeated measures ANOVA (2X2) was conducted in SAS©. There were two Generalized Linear Models. We modeled the log of time controlling for currency formatting (fixed, post-entry, auto-formatting) and currency alignment (left, right, center). We modeled total accuracy controlling for currency formatting (fixed, post-entry, auto-formatting) and currency alignment (left, right, center).[2]

Chi-square tests were used to determine differences in difficulty ratings between the nine conditions and differences in preference between currency formatting (fixed, post-entry, and auto-formatting).

4.4 Results

Time (Efficiency). Modeling the log time to complete, we found no main effect of currency formatting (fixed, post-entry, auto-formatting) on time per page ($F(2, 30) = .85$, $p > .05$) and no main effect of currency alignment type (left, right, or center) on time per page ($F(2, 30) = .17$, $p > .05$).

Accuracy (Effectiveness). A repeated measures model was used to determine if total accuracy per page is influenced by currency formatting and currency alignment. Each condition had five questions and a score of 1 was given to the correct responses. To calculate the variable total accuracy, we calculated a sum value for the five questions per page. A perfect score per page would have a score of five. Total accuracy for fixed entry was $M = 4.06$, $SE = .17$; total accuracy for post-entry was $M = 4.57$, $SE = .17$; and total accuracy for auto-formatting was $M = 4.26$, $SE = .17$.

Comparison tests (Tukey) reveal a significant difference between fixed entry and post-entry groups only (difference between means = .51, $p < .05$). This suggests that there was a difference in total accuracy between the groups—entering numeric data for post-entry resulted in higher accuracy compared to fixed formatting. There was no difference between fixed entry and auto-formatting or between post-entry and auto-formatting.

[2] We checked for significant interactions and found none, so we use a main effects model.

Results reveal a main effect of currency alignment type (left, right, or center) on total accuracy per page ($F(2, 30) = 3.99$, $p < .05$). Total accuracy per page for left alignment was $M = 4.46$, $SE = .17$, total accuracy per page for right alignment was $M = 4.05$, $SE = .17$, and total accuracy per page for center alignment was $M = 4.41$, $SE = .17$. Comparison tests (Tukey) reveal a significant difference between left and right alignment (difference between means = .41, $p < .05$) only but no significant difference between left and center alignment, or right and center alignment.

Difficulty Rating (Satisfaction). Two chi-square tests were used to determine an optimal currency formatting (fixed, post-entry, auto-formatting) and currency alignment type (left, right, center). Chi-square results reveal no difference in satisfaction ratings between currency formatting ($\chi^2(4) = .67$, $p > .05$). Chi-square results also reveal no difference in satisfaction ratings between currency alignment types ($\chi^2(4) = 1.6$, $p > .05$). There were no ratings lower than 3, suggesting that overall, participants did not find the task too difficult.

Preference. A chi-square test was conducted to determine if there was a significant difference in participant's subjective preference between currency formatting only. The chi-square was not significant ($\chi^2(2) = 5.6$, $p = .06$). Four participants preferred fixed entry compared to 14 for post-entry and 12 for auto-formatting.

5 Experiment 3: Forward and Backward Navigation Buttons

5.1 Designs Tested in the Experiment

Experiment 3 focused on navigation buttons. Forward and backward navigation buttons are a necessity in the design of online mobile surveys. These buttons are what allow respondents to move to the next page and progress through a survey or move back to a previous page to fix a mistake on a question they have already answered. Due to the importance of forward and backward navigation on the successful completion of a mobile web survey, it is imperative that the function of these buttons is clear. Due to the limited screen size on mobile devices, buttons are often labeled with icons rather than text labels because they can be smaller and take up less space. However, this practice has the potential to make the function of these buttons ambiguous to populations not familiar with them.

In a study by [23], they tested the success of novice computer users in initially learning to use an end user application program on a desktop computer over the course of two 90 min sessions separated by one week to test knowledge retention. The interfaces for this application implemented buttons labeled with only icons, only text, or a combination of icons with text. The icon-only labeled buttons performed the worst out of the three interfaces in all performance measures in the first session. However, by the end of the second session, the icon-only was not significantly different from the other groups. This research suggests that buttons labeled with icons rather than text will not be understood by total novices.

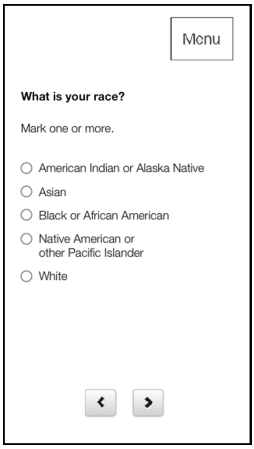

Fig. 13. Icon navigation button label

Fig. 14. Text navigation button label

Further, in a study by [24], they tested modern icons from mobile phones with both younger (age 20–37) and older adults (age 65+) and found that older adults have more problems using existing mobile device icons. It was also found that text labels help both young and old adults to initially use icons. They suggest on the basis of their findings that mobile device icons should be labelled at least initially, especially for older adults.

This was a between-subjects design with a single experimental factor, navigation button labels. This factor had two levels:

- Level 1: Labelled with text
- Level 2: Labelled with icons

Two different versions of a short five question survey were developed where the forward and backward navigation buttons were labeled using one of these two methods. In the text labelled condition, the forward button was labelled with "Next" and the backward navigation was labelled with "Previous". In the icon labelled condition, the forward button was labelled with ">" and the backward button was labelled with "<". Sixteen participants completed the survey in the text labelled condition and sixteen participants completed it in the icon labelled condition. See Figs. 13 and 14 for examples of both labeling conditions.

5.2 Adaptation of Survey Questions

The survey questions were based on real questions that are used in government surveys. Four of the questions were yes/no questions and included the following:

- Have you completed a secondary (high) school diploma or equivalent?
- Last week were you employed for pay at a job or business?
- During the past 12 months, did you take any work related training, such as workshops or seminars?
- Do you have a currently active professional certificate or a state or industry license?

The fifth was a question on race and can be seen in Figs. 13 and 14.

5.3 Evaluation Criteria

The app collected behavioral measures, which included trial navigation response times, optimal navigation deviations, and difficulty ratings. Trial navigation response times was the length of time it took participants to find and tap the navigation buttons. The time was recorded starting from the point that the participant tapped a response option for each survey question and ended when they tapped any navigation button or link. In this way we were able to isolate the time spent navigating and not the time spend interpreting and answering the question. Deviations from an optimal navigation path were recorded as any buttons tapped that were not the forward navigation button. The purpose of this was to identify whether participants had difficulty interpreting the button labels to move forward. Finally, difficulty ratings were recorded at the very end of the survey with a short 5-point rating scale for participants to rate the difficulty of completing the short survey from "very easy" to "very difficult." At the very end of the session, participants were shown both design conditions printed on a piece of paper and asked to choose which one they would prefer to assess overall preference.

It is assumed that any differences in trial navigation response times between conditions resulting from a lack of understanding of the navigational icon labeling would disappear or decrease after the first trial due to a learning effect. Therefore, an independent samples t-test was conducted for the response times for the first trial only and another t-test was conducted for the average of the remaining four trials (after learning occurs) response times. It is well known that response time data is susceptible to skewness due to the fact that it is bounded at zero to the left side but not on the right. There can be lapses in attention or distractions which can result in large outliers that may reduce the power of hypothesis tests of response time means between conditions [25]. To address this, we applied a log transform of the response time data before conducting these analyses. Additionally, three extreme outliers were identified in the first trial after visual inspection of the raw response time data. Video recordings from the sessions were reviewed and it was confirmed that the three outliers in the first trial had selected a response option, which started the timer, and then began to speak with the TA about the content of the question instead of immediately trying to navigate to the next question. These extreme values were excluded from the first trial t-test because we were able to confirm that they resulted from human error.

Optimal navigation deviations were expected to be rare so this was collapsed across all trials and whether a deviation occurred at all at any point during the survey was simply coded as 1 or 0. Due to the low expected values, the assumptions for a Chi-squared test could not be met and a Fisher's exact test was conducted instead.

A Chi-squared test was conducted for satisfaction ratings and overall preference.

5.4 Results

First Trial Completion Time (Efficiency). An independent sample unequal variances t-test was conducted to identify any significant differences between labeling conditions on the time it took to navigate forward after the first survey question. The results of this test found that the icon condition (M = 4.16, SE = 1.17) was significantly slower to navigate forward than the text condition (M = 1.53, SE = 0.16) after selecting a response option when participants saw the navigation button for the first time; $t(16.42) = 2.43$, $p < .05$. See Fig. 15.

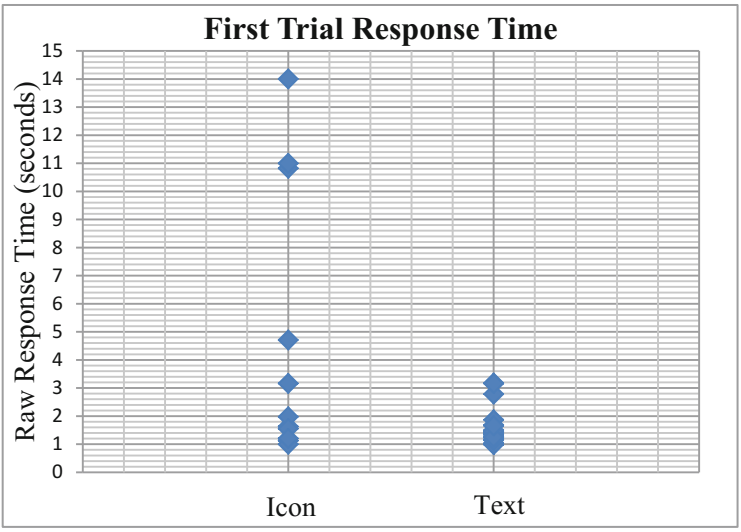

Fig. 15. Scatterplot split by labeling conditions of the raw response times for the first survey question.

Average Trial Completion Time (Efficiency). An independent sample t-test was conducted to identify any significant differences between labeling conditions for the mean time it took to navigate forward after a response option was selected. The results of this t-test found that there was not a significant difference between the icon (M = 1.76, SE = 0.27) and text (M = 1.36, SE = 0.11) conditions in the average time it took to navigate forward after selecting a response option; $t(30) = -0.05$, $p > .05$. See Fig. 16.

Optimal Navigation Deviations (Effectiveness). A Fisher's exact test was conducted to identify whether the number of participants who tapped an incorrect button to navigate forward differed significantly between label conditions. The results of the Fisher's exact test found that there was not a significant difference between groups ($p > .05$) for optimal navigation deviations. There were no deviations at all in the text-labeled group and there were two deviations in the icon-labeled group.

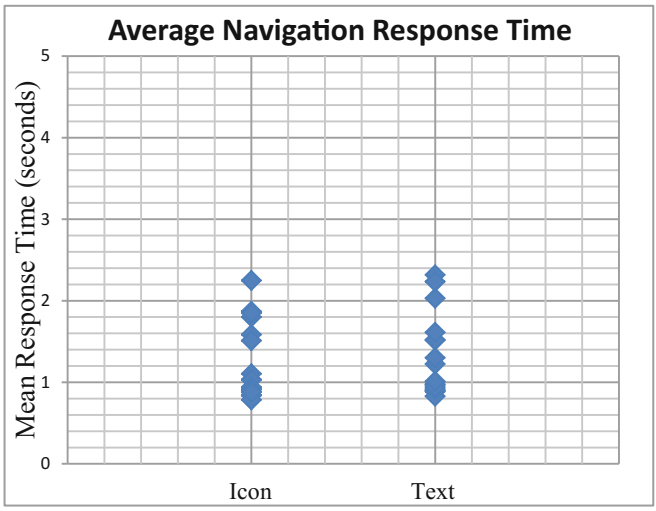

Fig. 16. Scatterplot split by labeling condition showing the mean navigation response times excluding the first trial.

Satisfaction (Satisfaction). A Chi-squared test was conducted to determine whether difficulty ratings were significantly different between labeling conditions. The results of this test did not yield any significant differences ($\chi^2(2) = 1.04$, $p > .05$). Difficulty ratings were virtually identical between groups with almost all participants reporting a rating of 1 (very easy).

Overall Preference (Satisfaction). A Chi-squared test was conducted to determine whether there was a significant difference between proportion of participants that preferred one labeling design of the other. The participants were shown both designs and were asked to choose which one they would prefer to use in a survey or both/neither. The results of the Chi-squared test found that a significant number of the older adults preferred the text labeled navigation buttons ($\chi^2(2) = 19.75$, $p < .001$). Almost 70% of the 32 participants preferred the text labeled navigation buttons compared with just over 20% that preferred the icons.

6 Overall Discussion and Implications for Future Research

The basis of this research was to learn more about how to design mobile surveys for older adults. For the first experiment, the data suggest that the iOS picker took longer and was preferred less than the other designs, which supports the hypothesis. We observed a significant increase in respondent burden as measured by time-on-task and by number of touches to the screen for the iOS picker design compared with a radio button/keyboard design. There were a high number of touches per question on questions with many response options when using the iOS picker. This finding matched our observation that the wheel at the bottom of the screen went fast and a lot of participant

manipulation was needed to select from long lists such as months, days, years, and states. When asked to compare three different designs, participants overwhelming selected the keypad entry design as the preferred mode to enter date of birth. Thus we recommend designers avoid using the default iOS picker design for response options and instead opt for the Android spinner style or the radio button/keyboard design. To accomplish this for iOS systems, developers will need to implement additional programming to override the default iOS design. For date of birth, a common survey question, a keypad is preferred over dropdowns for this user group.

In the second experiment we observed that left-aligned currency resulted in higher accuracy which did not match the hypothesis. The data also suggest that currency formatting was least effective when it appeared as fixed format, which is in line with our hypotheses. Currently at the Census Bureau the fixed formatting is used in many online surveys because auto-formatting will not work if the respondent has JavaScript turned off. This will be something that should be investigated as more tools are developed and programming for web-based surveys evolve.

Finally, in the third experiment the hypothesis that the navigation button labeled with text words would outperform the button labeled with only an icon was supported by the data. We recommend always labeling the forward and backward navigation buttons using text rather than icons for older adults. It was found that older populations may not be as familiar with the functionality associated with common internet UI elements such as the forward and backward arrow icons. The icon labeled buttons were simply ambiguous to some participants upon their first encounter with them as was seen by the longer response times on the first survey question and errors that only occurred when an icon was present versus a text label.

In contrast to some of the literature on healthcare and older adults that show the use of apps and mobile phones to cause barriers to older adults, [15], this work shows that older adults are able to answer survey questions on a mobile phone and that some designs outperform others. This work is in line with the general conclusion that found that older adults are interested and able to learn how to use mobile phones in their daily lives and that improvements to the design will aid in their performance [26]. Designers should take these recommendations into consideration when optimizing survey response choices for mobile phone and older adults. Future work should look at comparing how young and middle aged adults perform on these same tasks to see if there are any differences.

Acknowledgements. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. We would like to thank Andrew Roberts and Joanne Pascale for their reviews of an earlier draft of this paper.

References

1. Pew Research Center.: Mobile Fact Sheet (2018). http://www.pewinternet.org/fact-sheet/mobile/
2. Horwitz, R.: Personal Communication. U.S. Census Bureau (2018)

3. Cunningham, J., Neighbors, C., Bertholet, N., Hendershot, C.: Use of mobile devices to answer online surveys: implications for research. BMC Res. Notes **6**, 258 (2013). https://doi.org/10.1186/1756-0500-6-258
4. Pew Research Center: Tips for Creating Web Surveys for Completion on a Mobile Device (2015). http://www.pewresearch.org/files/2015/06/2015-06-11_tips-for-web-surveys-on-mobile.pdf
5. Antoun, C., Couper, M., Conrad, F.: Effects of mobile versus PC web on survey response quality: a crossover experiment in a probability web panel. Public Opin. Q. **81**(S1), 280–306 (2017)
6. de Bruijne, M., Wijnant, A.: Comparing survey results obtained via mobile devices and computers: an experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. Soc. Sci. Comput. Rev. **31**(4), 482–504 (2013)
7. Couper, M.P., Antoun, C., Mavletova, A.: Mobile web surveys: a total survey error perspective. In: Biemer, P., et al. (eds.) Total Survey Error in Practice, pp. 133–154. Wiley, New York (2017)
8. Zhou, J., Rau, P., Slavendy, G.: Use and design of handheld computers for older adults: a review and appraisal. Int. J. Hum. Comput. Interact. **28**(12), 799–826 (2012)
9. Parker, S., Jessel, S., Richardson, J., Reid, M.: Older adults are mobile too! Identifying the barriers and facilitators to older adults' use of mHealth for pain management. MC Geriatr. **13**(43) (2013)
10. Zhou, J., Rau, P.-L.P., Salvendy, G.: A qualitative study of older adults' acceptance of new functions on smart phones and tablets. In: Rau, P.L.Patrick (ed.) CCD 2013. LNCS, vol. 8023, pp. 525–534. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39143-9_59
11. Hoober, S.: Design for Fingers and Thumbs Instead of Touch. UX Matters (2013). https://www.uxmatters.com/mt/archives/2013/11/design-for-fingers-and-thumbs-instead-of-touch.php
12. Hoober, S.: Common Misconceptions About Touch. UX Matters (2013). https://www.uxmatters.com/mt/archives/2013/03/common-misconceptions-about-touch.php
13. Jin, Z.X., Plocher, T., Kiff, L.: Touch screen user interfaces for older adults: button size and spacing. In: Stephanidis, C. (ed.) UAHCI 2007. LNCS, vol. 4554, pp. 933–941. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73279-2_104
14. Joe, J., Demiris, G.: Older adults and mobile phones for health: a review. J. Biomed. Inf. **46**, 947–953 (2013)
15. Fletcher, J., Jensen, R.: Mobile health: barriers to mobile phone use in the aging population. Online J. Nurs. Inf. (OJNI) **19**(3) (2015). http://www.himss.org/ojni
16. Goggin, N., Meeuwsen, H.: Age-related differences in the control of spatial aiming movements. Res. Q. Exerc. Sport **63**(4), 366–372 (1992)
17. Ketcham, C., Seidler, R., van Gemmert, A., Stelmach, G.: Age-related kinematic differences as influenced by task difficulty, target size, and movement amplitude. J. Gerontol. Psychol. Sci. **57B**(1), 54–64 (2002)
18. Wallace, S., Graham, C., Saraceno, A.: Older adults' use of technology. Perspect. Gerontol. **18**(2), 50–59 (2013)
19. Craik, F.I.M., Salthouse, T.A.: The Handbook of Aging and Cognition. Lawrence Erlbaum Associates, Mahwah (2000)
20. Fisk, A.D., Rogers, W.A.: Handbook of human Factors and the Older Adult. Academic Press, San Diego (1997)
21. Salthouse, T.: When does age-related cognitive decline begin? Neurobiol. Aging **30**(4), 507–514 (2009)

22. Wang, L., Antoun, C., Sanders, R., Nichols, E., Olmsted-Hawala, E., Falcone, B., Figueroa, I., Katz, J.: Experimentation for developing evidence-based UI standards of mobile survey questionnaires. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2998–3004. ACM Press, Colorado (2017)
23. Wiedenbeck, S.: The use of icons and labels in an end user application program: an empirical study of learning and retention. Behav. Inf. Technol. **18**(2), 68–82 (1999)
24. Leung, R., McGrenere, J., Graf, P.: Age-related differences in the initial usability of mobile device icons. Behav. Inf. Technol. **30**(5), 629–642 (2011). https://www.learntechlib.org/p/52032/
25. Whelan, R.: Effective analysis of reaction time data. Psychol. Rec. **58**(3) (2008). Article 9
26. Wright, P., Bartram, C., Rogers, N., Emslie, H., Evans, J., Wilson, B., Belt, S.: Text entry on handheld computers by older users. Ergonomics **43**(6), 702–716 (2000)