methods, data, analyses | Vol. 9(2), 2015, pp. 229-260 DOI: 10.12758/mda.2015.013

Are Sliders Too Slick for Surveys? An Experiment Comparing Slider and Radio Button Scales for Smartphone, Tablet and Computer Based Surveys

Trent D. Buskirk[1], *Ted Saunders*[2] *& Joey Michaud*[2]

1 Marketing Systems Group
2 MaritzCX

Abstract

The continued rise in smartphone penetration globally afford survey researchers with an unprecedented portal into personal survey data collection from respondents who could complete surveys from virtually any place at any time. While the basic research into optimizing the survey experience and data collection on mobile devices has continued to develop, there are still fundamental gaps in our knowledge of how to optimize certain types of questions in the mobile setting. In fact, survey researchers are still trying to understand which online design principles directly translate into presentation on mobile devices and which principles have to be modified to incorporate separate methods for these devices. One such area involves the use of input styles such as sliding scales that lend themselves to more touch centric input devices such as smartphones or tablets. Operationalizing these types of scales begs the question of an optimal starting position and whether these touch centric input styles are equally preferred by respondents using less touch capable devices. While an outside starting position seems optimal for slider questions completed via computer, this solution may not be optimal for completion via mobile devices as these devices are subjected to far more space and layout constraints compared to computers. This experiment moves the mixed device survey literature forward by directly comparing outcomes from respondents who completed a collection of survey scales using their smartphone, tablet or computer. Within each device, respondents were randomly assigned to complete one of 20 possible versions of scale items determined by a combination of three experimental factors including input style, length and number formatting. Results from this study suggest more weaknesses than strengths for using slider scales to collect survey data using mobile devices and also suggest that preference for these touch centric input styles varies across devices and may not be as high as the preference for the more traditional radio button style.

Keywords: Smartphone and Tablet Surveys, Slider Scales, Radio Buttons, Experimental Design, Missing Items

1 Introduction

The continued rise in smartphone penetration globally afford survey researchers with an unprecedented portal into personal survey data collection from respondents who could complete surveys from virtually any place at any time. Indeed, over the past five years, the research in online survey data collection has extended beyond computers to include both smartphones and tablets. Buskirk (2015) describes contemporary trends in survey optimization for these mobile devices but in short, some of the current approaches not only consider how to implement well-established online survey design principles for mobile devices, but also seek to understand which and how any of these principles need to be modified for mobile devices. Mobile devices, in general, represent a type of survey mode in which potential respondents have themselves gained extensive experience using – including checking emails, using apps and browsing the web (Link & Buskirk, 2013). One might conjecture that these respondent experiences might speak to a greater sense and expectation for websites, including survey websites, to be easy to navigate, engaging and interactive.

One type of survey question scale that has been touted in the recent literature as more engaging and more interactive than the more traditional radio button variety is the slider scale. Slider scales, unlike radio buttons, enable both animation and interactivity by requiring the respondent to touch or click a slider handle and slide or drag it along a fixed axis until it reaches the desired answer choice or level. While the usual application of sliders don't go as far as gamification (Keusch & Zhang, 2014), they have been purported to afford respondents a more engaging experience (Cape, 2009; Puleston, 2011). Two related types of scales that have also been explored recently in the survey literature are visual analog scales (VAS) and graphic response scales (GRS). Unlike sliders that usually have a dragging or sliding interactivity, visual analog scales ask the respondent to place a mark for their response along an axis that is anchored by two endpoints while graphic response scales ask respondents to place a mark along an axis that has graded semantic label anchors along the continuum in addition to the two endpoint anchors (Couper et al., 2006). From a required action perspective slider scales require a dragging action

Direct correspondence to
Trent D. Buskirk, Ph.D., Vice President of Statistics and Methodology, Marketing Systems Group
E-mail: TBuskirk@m-s-g.com

Acknowledgements: We would like to thank Michael Sherwood of MaritzCX for his assistance with this project. We would also like to thank Melanie Courtright of Research Now for her assistance in coordinating data collection from the panel. Finally, we would like to thank the very helpful team at Decipher who provided assistance hosting, programming and optimizing the survey.

while both VAS/GRS and radio button scales require a clicking action. From a precision perspective, both VAS/GRS and slider scales may be preferred to radio buttons since theoretically they allow a continuum of answer choices instead of a discrete collection. The category slider represents a more discrete version of slider scales that has gained in popularity as evidenced by ease of availability in widely available pre-package survey software. Much like how graphic response scales add specific descriptors to the underlying range, category sliders add descriptors to break up the underlying continuum of satisfaction, agreement or other construct being represented by the slider. The category sliders represent the "ordinary" response categories that are typically represented by a comparable radio button scale.

The relative merits of VAS/GRS, sliders and radio button scales have been previously explored in the online survey context for computers (Couper et al., 2006) and have recently been explored for both computers and mobile devices (Toepoel & Funke, 2014). More broadly, Sikkel et al. (2014) explored the relative merits of dragging and clicking operations for category sliders, among other scale types, in the context of online surveys completed by PC and find that dragging operations increase user engagement with the survey but only when they are used sparingly. As Derham (2011) pointed out, researchers must make many choices when considering slider scales and these choices can individually and collectively impact data quality. Roster et al. (2015) posited that the considerable variability in the utility of sliders in surveys observed across research studies is in part due to the many aspects of slider construction and presentation that could be considered including among others: scale length, whether the outcome is treated as continuous or discrete, variations of graphics, use of labels and slider starting position. By far the most common starting position that has been tested in the survey literature has been left starting position (see Toepoel & Funke, 2014; Roster et al., 2015; Funke et al., 2011; Sikkel et al., 2014; Buskirk & Andrus, 2014). Petersen et al. (2013) examined sliders with a left start for scale items that had no natural neutral position and a middle start for those with a neutral position but these two starting places were not compared to other possible positions. Slider orientation was examined by Funke et al. (2011) an no discernable differences other than time were noted for vertical versus horizontal versions of the slider and the comparable radio button scale was held at fixed length. Toepoel and Funke (2014) compared sliders and radio buttons based on scales having three different lengths (5, 7 and 11 point items) and found differences between slider and radio buttons for desktop respondents for 5 and 7 point scale items and for mobile respondents for the 11 point scale items. Cape (2008) conducted an experiment comparing four versions of slider scales that varied the formatting of the slider scale but kept the starting position (left most option) and the length of scale (5 point Likert) constant. The results indicated that while different versions of the slider scale produced different response distributions, the overall mean scores across different versions of the slider scale were similar.

In this study we simultaneously compare three scale aspects for surveys items fielded across smartphones, tablets and computers. An equal number of respondents from each of these device types was recruited and then randomized to complete survey scale items whose format was determined by a combination of three experimental factors including input style, length and number formatting. This experiment moves the mixed device survey literature forward by directly comparing outcomes from respondents who completed a collection of survey scales using their smartphone, tablet or computer. The study also offers one of the more comprehensive comparisons of radio buttons to slider scales in terms of the number of simultaneous attributes of slider scale designs considered within one survey experiment.

2 Recruitment and Experimental Design

Participants for this study were recruited from Research Now's US consumer e-rewards panel which consists of nearly 2.5 million adults making it one of the largest sources of online responses in the U.S.[1] Survey invitations were sent to the panel soliciting participants to complete a short survey using either a smartphone, tablet or computer with the goal of recruiting at least 1,200 respondents from each device type which was tracked using the panelist's device user agent string (Callegaro, 2010). The overall survey consisted of up to 60 possible questions about automobile insurance satisfaction and was designed to be completed in no more than 10 minutes using a web browser. The survey was optimized for mobile devices and according to the taxonomy of Buskirk and Andrus (2012) the mobile versions would be considered active mobile browser surveys. The study fielded in the U.S. between April 4 and 11, 2014 and each respondent received an identical e-incentive that was comparable in value to other panel surveys of similar length.

Because the panel provider's members generally completed surveys online or via tablet computers, we could not randomize device type to each panelist as not all participating panelists had each of these devices. Instead, we allowed device type to be a natural or native blocking variable and made all experimental randomizations within each type of device separately and independently. Specifically, once a panelist clicked on the study link they were taken to an introduction page. At this point we tracked the device type using the device's user agent string (Callegaro, 2010). After clicking start on the introduction page, each panel respondent was then

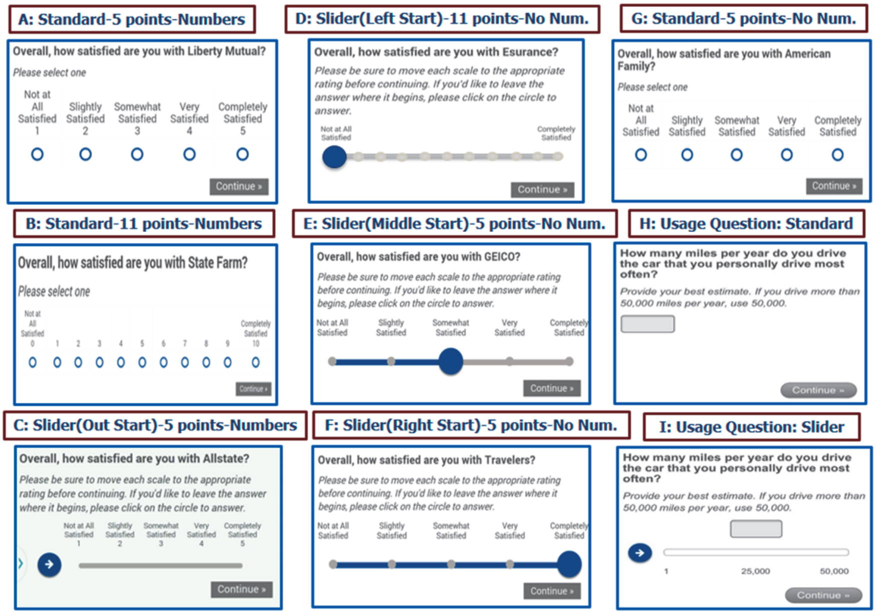
[1] Members of the e-rewards panel are recruited by invitation only from one of many participating partner loyalty programs and respondents who complete surveys while on this panel receive electronic credits that can later be redeemed for various rewards.

randomized to receive scale items for the experiment that were formatted according to one of five possible scale types including: standard radio buttons or sliders with either an outside, left, middle or right starting position as illustrated in Figure 1 A, C-F. Consistent with the recommendations made by Roster et al. (2015) we provided an additional instruction for respondents in any slider scale group to click on the slider handle if their answer was consistent with where the slider began (see Figure 1 C-F). Because this experiment was conducted within the scope of a market research study that required standard radio button scales to produce estimates, the randomization to the scale type used a 4:1 ratio within each type of device with 4 respondents being randomly assigned to standard radio buttons for every 1 randomly assigned to each type of slider scale. In addition to scale type, respondents were equally randomized to one of two scale lengths (5 point vs. 11 point) and equally randomized to one of two scale numbering formats (numbered versus not numbered). All 5-point scales were fully anchored with semantic labels and the numbered versions also included number values below each of the semantic labels (see Figure 1 E, G and A, C, respectively). All 11-point scales were end-anchored with semantic labels and the numbered versions contained number values for each possible choice ranging on the low end of 0 to the high end of 10 (see Figure 1 D and B, respectively). The slider starting position was also relative to the length of scale, so for example, middle start with 5 point scales placed the handle on option 3 and middle start with 11 point scales placed the handle on option 5.

We note that our sample is from an online data source and was not selected by probability sample and was not otherwise intended to represent the broader population of the U.S. But as others have also noted (Buskirk & Andrus, 2014; Couper et al., 2006) our intention here is to compare results across experimental factors (e.g. scale type, scale length and number formatting) as well as the blocking variable of device type. We also note that while some studies have randomized or assigned respondents to device (Peytchev & Hill, 2010; Scagnelli et al., 2012), we allowed respondents to self-select by device. In this way, the experiment is embedded in a setting that is natural to the respondent and likely more consistent with what might be found in practice with respondents completing online surveys using whatever device is available to them.

3 Survey Items and Measures

Twenty three of the 60 possible survey items were considered for this experiment. The remaining questions provided data for two other experiments, both of which have been reported elsewhere (see Buskirk, et al., 2014, Michaud, et al., 2014 and Courtright et al., 2014). The first survey item included in this experiment asked respondents to enter the total number of miles driven within the past year. If the



A: Standard-5 points-Numbers

Overall, how satisfied are you with Liberty Mutual?

Please select one

B: Standard-11 points-Numbers

Overall, how satisfied are you with State Farm?

Please select one

C: Slider(Out Start)-5 points-Numbers

Overall, how satisfied are you with Allstate?

Please be sure to move each scale to the appropriate rating before continuing. If you'd like to leave the answer where it begins, please click on the circle to answer.

D: Slider(Left Start)-11 points-No Num.

Overall, how satisfied are you with Esurance?

Please be sure to move each scale to the appropriate rating before continuing. If you'd like to leave the answer where it begins, please click on the circle to answer.

E: Slider(Middle Start)-5 points-No Num.

Overall, how satisfied are you with GEICO?

Please be sure to move each scale to the appropriate rating before continuing. If you'd like to leave the answer where it begins, please click on the circle to answer.

F: Slider(Right Start)-5 points-No Num.

Overall, how satisfied are you with Travelers?

Please be sure to move each scale to the appropriate rating before continuing. If you'd like to leave the answer where it begins, please click on the circle to answer.

G: Standard-5 points-No Num.

Overall, how satisfied are you with American Family?

Please select one

H: Usage Question: Standard

How many miles per year do you drive the car that you personally drive most often?

Provide your best estimate. If you drive more than 50,000 miles per year, use 50,000.

I: Usage Question: Slider

How many miles per year do you drive the car that you personally drive most often?

Provide your best estimate. If you drive more than 50,000 miles per year, use 50,000.

Figure 1 A-G – Visual Examples of the various factor combinations for the 22 core scale items as viewed on a smartphone. H and I refer to text input versus slider input for the single usage item (also viewed on a smartphone)

respondent was assigned to any one of the four slider scale groups, this question was presented as a slider with an outside start; otherwise, it was presented as an open-ended text box as illustrated in Figure 1 H and I, respectively. The remaining 22 questions (henceforth referred to as "core scale items" were presented over 7 separate screens and were organized into three different primary outcome measures including the: Overall Satisfaction Measure (OSM), Brand Performance Measure (BPM) and the Service Preferences Measure (SPM). The OSM was computed as the sum of three scale questions that asked respondents about their overall satisfaction with their Automobile Insurance Providers as well as how likely they were to recommend the provider to friends/colleagues and to renew their policies. The BPM was computed as the sum of ten scale items that asked respondents to rate their primary automobile insurance provider on ease of business transactions, trust, discounts, customer service, convenience, value, and accessibility using a scale that was based on anchors ranging from "Poor" to "Excellent." Finally, the SPM was computed as the sum of 9 scale items that asked respondents to rate the degree of agreement with statements about how they purchase automobile insurance, how they interact with an insurance agency, and the extent to which they want to use mobile devices for their automobile insurance needs. Each of these scale items was anchored on endpoints that ranged from “Strongly Disagree” to “Strongly Agree.” When we discuss the OSM, BPM and SPM measures throughout this paper we will add (5) or (11) to the abbreviation to refer to the number of scale points included in each of the scale items used to compute the measure. For example SPM(5)/SPM(11) refers to the service preference measure computed using scale items with 5 or 11 points, respectively. The actual values assigned to responses for 5 point scale items ranged from 1 to 5 and from 0 to 10 for 11 point scale items.

To examine both preference and consistency of reporting across scale types we also asked every respondent to answer the “overall satisfaction with their insurance provider” item (OSI) a second time at the end of the experiment using a scale presented with the opposite input style.[2] The scale numbering and length were the same across both OSI versions. After the respondent completed the second version of the OSI, they were asked “If you had the choice of how to give us your ratings, which way would you prefer?” with answer choices including “slider”, “buttons” (i.e. radio) and “no preference.” Using the two OSI items we also computed two versions of concordance. The first measure was simply a binary indicator for an exact match between the two responses (Exact Concordance). The second measure indicated concordance if the two responses differed by no more than 1 category unit up or down (±1Concordance).[3]

4 Analyses and Results

We note that for this study we are interested in comparisons across devices and across the other experimental factors as well as possible interaction effects between these factors for a series of survey related outcomes. At the extreme there could be a total of 60 unique cells, formed by crossing device (3) with scale type (5), scale length (2) and scale numbering (2), that would be compared by a model for any given outcome. Based on this extreme case, we attempted to cap the overall experiment-wise type I error rate to be at worst 30% for a given outcome by setting the *individual* type I error rate to be .005 (0.30/60). Thus for each specific survey outcome, the p-values reported in this paper are not adjusted further for multiple comparisons and we declared statistical significance for any effect or comparison if the unadjusted p-value was less than .005.

[2] All respondents initially assigned to the “radio button” scale type were additionally randomly assigned in equal proportion to one of the four slider starting positions for the purposes of the preference and consistency analysis.

[3] For example, a respondent who answered 7 for the first OSI and 8 for the second (using an 11 point scale) would not be concordant under exact concordance but would be under ±1Concordance.

4.1 Survey Break-Offs

In general the break-off rates for the experiment were moderately low across the three devices. In total, there were 1,250 computer, 1,340 tablet and 1,449 smartphone respondents who accepted our invitation to participate in the experiment and began the survey.[4] A total of 1,201 computer, 1,199 tablet and 1,198 smartphone respondents completed the experiment for respective break-off rates of 4% for computer, 11% for tablet and 17% for smartphones. While the results are not shown here we did examine break-off rates by the three experimental factors both within device and across devices and found no systematic pattern or practical differences.

4.2 Completion Times

We note that there were technical difficulties with the time tracking algorithm in the first day of fielding rendering the time stamps missing for all survey items for 369 of the 3,598 respondents across the three devices. The distribution of times to complete the single automobile usage item for the 3,229 respondents for which times were available was slightly positively skewed with extreme times observed from 20 PC respondents (2.3%) (exceeding 70 seconds), 30 Tablet respondents (2.6%) (exceeding 62 seconds) and 40 Smartphone respondents (3.4%) (exceeding 68.5 seconds). The longest time observed for this item was from a Tablet respondent who took in excess of 7,115 seconds (or just under 2 hours) to complete this question. Because of the observed skewness, we analyzed the natural log of completion times for the usage item based on a general linear model that includes device and the scale type (e.g. standard open ended text box versus slider-bar) as well as the interaction of these two factors. Based on the model we found that the completion times for the usage item (on the natural log scale) varied significantly by the device (F(2,3223)=6.55; p-value=.0015) and type of scale (F(1, 3223)=27.30; p-value<.0001). Despite the large outlying observation observed from a Tablet respondent, PC respondents had the largest geometric mean completion time for the usage item which was estimated to be about 9% longer than that from both Smartphone and Tablet respondents (p-values=.0022 and .0009, respectively) as illustrated in Table 1. No significant differences in the geometric means of completion times for the usage item were found between Tablet and Smartphone respondents

[4] The total number of survey invitations sent from the sampling provider by device type was not available as device type was determined only upon clicking continue on the initial survey introduction page. A total of 323,259 email invitations were sent to panelists yielding 21,217 opened invitations, which included 441 partial completes/break-offs, 3,598 survey completes and 1,476 panelists who did not persist past the survey intro page. An additional 12,631 panelists opened and responded to the invitation and clicked continue on the survey introduction page but did so using a device for which the quota had already been met and as such were terminated.

Table 1 Descriptive statistics for completion times (rounded and displayed to the nearest second) for the miles driven last year question by device and scale type

Device / Scale Type	n	Mean	Geometric Mean	Median	Std. Deviation	Min.	Max.
PC	884	31	21	20	145	4	3980
Tablet	1161	31	19	18	223	6	7116
Smartphone	1184	27	19	17	81	4	2124
Standard (Radio Buttons)	1613	28	18	17	185	4	7116
Slider	1616	31	21	19	133	4	3980

(p-value>.75). The geometric mean completion time for the usage item for the slider scale group was also estimated to be about 12% longer than that for the standard text box group (p-value<.0001).

The distribution of completion times for the core scale items was also positively skewed for each of the three devices. Some extreme observations[5] were observed from respondents from each of the devices including 20 of the 884 (2.3%) completing via PCs, 29 of the 1,161 (2.5%) completing using Tablets and 49 of the 1,184 (4.1%) completing via Smartphones. Basic summary statistics for the completion times by device type are given in Table 2. We note that the ranges of completion times for PC and Tablet users were generally consistent overall and across scale types, but the range for Smartphone users was quite large in comparison driven by two respondents – one who took more than 68,375 seconds (or just under 19 hours) and the other who took more than 11,873 seconds (or about 3.3 hours) to complete the questions for the experiment on their smartphones. Given the underlying skewness in the distribution, the analysis of differences in completion times across device and the three experimental conditions was conducted using a general linear model applied to the natural log of completion times. We note that the statistical comparisons of completion times for the experiment on the natural log scale were practically identical with and without these two very extreme outliers, so for posterity all analyses included these data points.

Completion times (on the natural log scale) varied significantly by both the device used for completing the survey (F(2, 3169)=27.27; p-value<.0001) and by

[5] Defined as exceeding 3 times the interquartile range plus the third quartile of completion times, recorded in seconds. Specifically, identified as completion times exceeding 352, 321 and 343 seconds for PC, Tablet and Smartphone respondents, respectively.

Table 2 Time (in seconds) to complete the core scale items (22 questions) by mode of response

Device / Scale Type	n	Mean	Geometric Mean	Median	Std. Deviation	Min.	Max.
PC	884	164	141	137	252	18	6655
Tablet	1161	134	116	110	118	17	1787
Smartphone	1184	242	124	115	2052	20	68376
Standard (Radio Buttons)	1613	202	125	118	1731	30	68376
Slider:Out	386	168	133	126	357	37	6655
Slider:Left	405	190	134	128	617	35	11873
Slider:Mid	396	149	122	114	191	17	2413
Slider:Right	429	139	118	121	132	20	2220

All analyses and statistical hypothesis tests were performed on the natural log scale so we also provide geometric means, since back-transformed means from the natural log scale estimate the geometric means from the raw, untransformed data. All times are rounded and displayed to the nearest second.

the scale type (F(4, 3169)=3.85; p-value=.0040) and these effects were additive in that no interaction between these two factors was detected. None of the other factors nor any second or higher order interactions were significantly related to the natural log of completion times (all remaining p-values >.10). The geometric mean completion time for PC respondents was estimated to be about 19% longer than that of Smartphone respondents (p-value <.0001) and estimated to be about 23% longer than that of Tablet respondents (p-value<.0001). No significant differences were found in completion times for the core scale items between Smartphone and Tablet respondents (p-value>.01). Respondents assigned to the slider left start group had the longest estimated geometric mean completion time (about 135 seconds, on average) and the geometric mean completion time for this group was estimated to be about 11% longer than that for the slider right start group (p-value=.0024). No other significant differences in completion times were found between any of the other scale types.

4.3 Missing Item Rates

While missing values were generally more of the exception than the rule for core scale items, some amount of item missingness was encountered. All in all, roughly 66% of respondents had no missing items for any of the core scale items. Among the third of respondents missing at least one core scale item, the 25th percentile of the number of items missing was 1, the median was 4, the 75th percentile was 9 and the 95th percentile of the number missing was 17. In total, 13 respondents were missing all core scale items. From the negative binomial regression model that explored the number of missing items as a function of device type and the experimental factors and higher order interactions, we determined that the variability in the number of missing items was fundamentally driven by scale type ($\chi^2(4)=2052.12$; p-value<.0001) but the impact of this factor was moderated separately by both scale length ($\chi^2(4)=36.68$; p-value<.0001) and also by device ($\chi^2(8)=43.30$; p-value<.0001).

Essentially, the slider right and middle start groups had significantly higher numbers of missing values, on average, compared to any of the other scale types and the number of missing items is practically (and statistically) consistent across the devices for each of the scale types. The main exception to this trend for device types comes from the slider right start group as shown in Figure 2 B. For this scale type we observed that Smarpthone respondents exhibited significantly higher numbers of missing items, on average, compared to PC respondents (p-value<.0001) but no significant differences were observed between Tablet or PC respondents (p-value>.02) nor between Smartphone and Tablet respondents (p-value>.05). As shown in Figure 2 A, the number of missing items, on average, was fairly consistent across the two scale lengths with the exception being found for the slider middle start group. Here respondents assigned to the 5 point scales had an average number of missing items that was about 75% larger than the 11 point scale group (p-value<.0001) which translated into about 3 additional missing items, on average. The number of missing items for respondents assigned to the 5 point version of the slider right start group exhibited about the same number of missing items than the slider middle start group (p-value>.42), but the number of missing items for the 11 point slider right start group was about twice as large as the 11 point slider middle start group (p-value<.0001). Overall, there was no difference in the number of missing items, on average for either the 5 point or 11 point versions of the slider right start groups (p-value >.30) and this scale type had the largest number of missing items on average (about 8 for the 5 point and 7 for the 11 point versions).

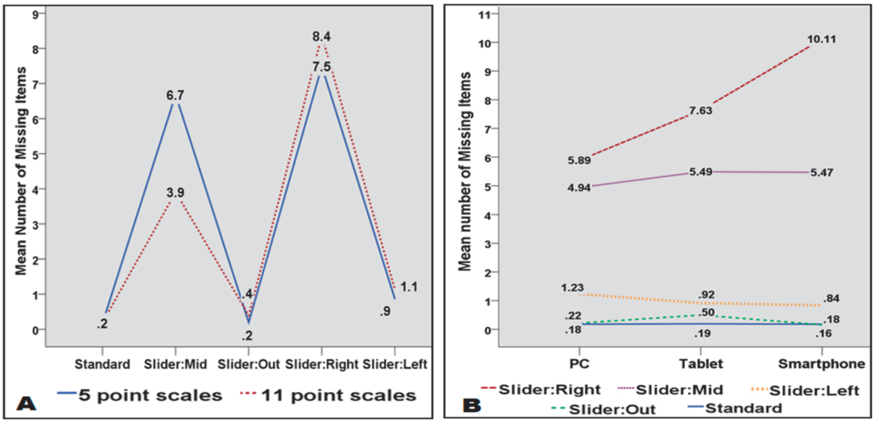


Figure 2 A: Mean number of missing items by scale type and scale length;
B: Mean number of missing items by device type and scale type

4.4 Survey Outcomes

4.4.1 Miles driven in the past year

Respondents entered the number of miles they drove within the past year using either a slider scale or an open ended text box. One aspect of sliders that differs from open ended texts, especially for numeric data is that sliders give the respondents a clear sense of the range with labels marking the beginning and ending points of the slider as illustrated in Figure 1 I. Open ended text boxes, on the other hand, can also provide respondents a sense of the range if explicit instructions are included as illustrated in Figure 1 H. Because the slider endpoints are more explicit we expected that more respondents in the slider group would enter values corresponding to the upper or lower endpoints compared to the text group. However, we found no significant differences between the slider and text groups for either the rate of respondents reporting the highest option (i.e. 50,000) or the lowest option (i.e. 0) (both p-values>.47). On the other hand, there were significant differences noted in the proportions of respondents reporting the highest option across devices (p-value<.003) with more Smartphone respondents reporting the maximum allowable amount compared to either PC or Tablet respondents as shown in the rightmost section of Table 3.

To compensate for the positive skewness observed in the miles driven distribution, we analyzed the relationship between the natural logarithm of 1 plus the miles driven and input style, device type and the interaction of these two factors

Table 3 Summary statistics for the number of miles driven in the past year by device type and entry method. Overall statistics for the entire sample are given in the bottom right-hand corner

Input Method	Device type	n	Mean	Std. deviation	Median	Geometric mean	% at 0	% at 50K
Text entry (standard)	PC	602	11327.13	7425.48	10000	8668.09	0.17%	0.50%
	Tablet	595	13442.34	8332.71	12000	10637.59	0%	0.67%
	Smartphone	587	14639.12	9752.80	12000	10192.01	0.67%	2.35%
	Statistics for text entry (across all devices)	1784	13122.36	8652.05	12000	9788.66	0.28%	1.17%
Slider entry	PC	597	11630.87	7368.88	10000	9383.08	0.17%	1.01%
	Tablet	600	13826.69	8708.15	12000	10876.36	0.33%	1.33%
	Smartphone	603	15544.23	9935.90	12500	11373.86	0.50%	2.16%
	Statistics for slider entry (across all devices)	1800	13673.79	8879.16	12000	10512.80	0.33%	1.50%
Summary statistics for devices (across entry types)	PC	1199	11478.37	7395.83	10000	9017.01	0.17%	0.75%
	Tablet	1195	13635.32	8521.88	12000	10756.81	0.17%	1.00%
	Smartphone	1190	15097.76	9852.27	12000	10774.68	0.58%	2.25%
Overall summary statistics for all respondents		3584	13399.30	8769.96	12000	10145.89	0.31%	1.33%

using a general linear model.[6] As was the case for completion times for the experiment, differences in the natural log of miles driven (plus one) varied significantly across device type (F(2, 3578)=11.38; p-value<.0001) but not by the style of input (F(1, 3578)=4.06; p-value>.04) nor by the interaction of device and input style (F(2, 3578)=.54; p-value>0.50). In particular, the geometric mean for the miles driven (plus one) for PC respondents was estimated to be approximately 16% less than that of either Tablet or Smartphone respondents who reported geometric means of roughly 10,757 and 10,775 miles driven within the past year, respectively (p-values<.0001).

4.4.2 High, middle and low option selection patterns for core scale items

Before examining specific substantive outcomes, we first explored general response patterns classified as the selection of "high", "middle" and "low" box options for each of the core scale items. On the five point scale we declared that the respondent selected a: "high option" if their response was either a 4 or 5; a "middle option" if their response was a 3 and a "low option" if their response was either a 1 or 2. For the 11 point scale, "high" options were defined as responses between 7 and 10; "middle" as responses between 4 and 6 and "low" for responses between 0 and 3. We created three separate models to examine the relationship between the selection rates of high, middle and low response options for core scale items and the three experimental factors, device type and all higher order interactions. To adequately compensate for observed over-dispersion for each of these three rates, we used negative binomial regression models with an offset equal to the natural log of the number of core scale items answered.

High option selection rates

High option selection rates varied significantly across scale type (χ^2(4)=147.72; p-value<.0001) and scale length (χ^2(1)=20.07; p-value<.0001) and by the interaction of these two effects (χ^2(4)=30.23; p-value<.0001). Neither the main effects of scale numbering nor device type nor any of the other interaction effects were found to be significant (all p-values >.12). In general we found that respondents in the slider middle and right start groups had significantly higher and lower, respectively, high option selection rates compared to other scale type groups as depicted by the solid red lines in Differences between scale type groups for the 11 point scale items

[6] We added 1 to all reported miles to avoid irregularities in the natural logarithmic transformation applied to the rather small number of zeroes that were reported for miles driven (Yamamura, 1999).

were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. Specifically, among respondents assigned to 5-point scales in the slider middle start group selected higher response options at rates that were, on average, nearly 30% more than that those of respondents in the slider left start, slider outside start and standard scale groups (all p-values<.0001). In contrast, respondents assigned to the slider right start group had estimated high option selection rates that were, on average, about 15% less than those of the slider left start, slider outside start and standard scale groups and about 34% less than those of the slider middle start group (all p-values<.0011). The differences observed for the 5 point scale items were generally consistent for the 11 point scale items, as shown in the right panel of Differences between scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3, although the magnitude of differences was less and the number of significant differences fewer[7].

Middle option selection rates

The middle option selection rates varied significantly across device type ($\chi^2(2)=13.74$; p-value=.0010), scale type ($\chi^2(4)=483.73$; p-value<.0001), scale length ($\chi^2(1)=8.18$; p-value=.0042) and the interaction of scale type and length ($\chi^2(4)=175.90$; p-value<.0001). Neither the main effect of scale numbering nor any of the other interaction effects from the full model were found to be significant (all p-values>.08). As indicted in Table 4, PC respondents selected middle response options about 14% less often than those for respondents completing by Smartphone, but no other significant differences across devices were noted (p-values>.015). As for scale type differences, generally respondents from the middle slider start group exhibited far lower middle option selection rates compared to any other scale type as depicted by the long-dashed green line in the left and right panels of Differences between scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. More specifically, for the 5 point scale items, respondents in the middle slider start group had middle options selection rates that were, on average, about 85% less than those of respondents from the slider left, outside and right start as well as the standard scale groups (all p-values<.0001). Differences between

[7] More specifically, respondents from either the right or outside slider start groups had high option selection rates that were, on average, about 10% less than those of the slider left start and standard scale groups and approximately 20% less than those of the slider middle start group (all p-values<.003) and no other significant differences between scale types were found for the 11 point scales (all p-values>.026).

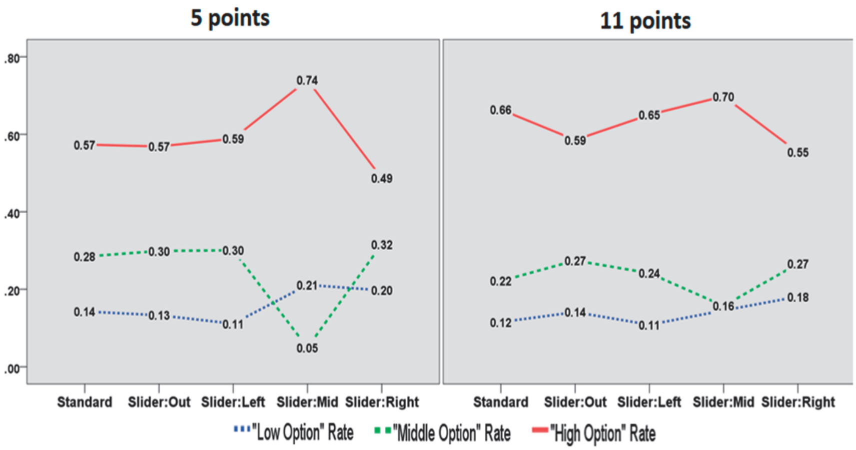


Figure 3 Low, middle and high option selection rates by type and length of scales for core scale items answered

scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less[8].

Low Option Selection Rates

The low option selection rates varied significantly across device type ($\chi^2(2)=36.92$; p-value<.0001), scale type ($\chi^2(4)=98.59$; p-value<.0001) and scale length ($\chi^2(1)=16.32$; p-value<.0001) as well as the interaction between scale type and length ($\chi^2(4)=17.98$; p-value=.0012). Neither the main effect of scale numbering nor any of the other interaction effects were found to be significant (all p-values≥.02). As shown in Table 4, PC respondent had low option selection rates that were, on average, about 16% and 30% higher than those of Tablet and Smartphone respondents, respectively (p-values<.0003). With respect to differences in the low option selection rates across scale types, we found that generally respondents in the left slider start group had significantly lower rates while respondents in the middle and right slider start groups had significantly higher rates compared to other scale types as depicted by the blue short-dashed lines in Differences between scale type groups

[8] In particular, respondents in the middle start group selected middle response options at rates that were, on average, about 35% less than those for either the slider left start or standard scale groups (p-values<.0001) and about 45% less than those for either the slider outside or right start groups (p-values<.0001). The middle option selection rates for the standard scale group were also about 20% lower, on average, than those of either the slider outside or right start groups (p-values<.0007).

Table 4 Selection of “Low”, “Middle” or “High” options across core scale items by device

	Type of device		
	PC (n=1200)	Tablet (n=1192)	Smartphone (n=1193)
Option selection rate	Mean (std. error)	Mean (std. error)	Mean (std. error)
„Low Option“	0.162 (0.004) †,❖	0.135 (0.004) ‡	0.126 (0.004)
„Middle Option“	0.224 (0.005) n.s.,❖	0.255 (0.006) n.s.	0.259 (0.006)
„High Option“	0.614 (0.006) n.s., n.s.	0.610 (0.007) n.s.	0.615 (0.007)

† indicates PC user rate is significantly different from Tablet user rate (α=.005)
❖ indicates PC user rate is significantly different from Smartphone user rate (α=.005)
‡ indicates Tablet user rate significantly different from Smartphone user rate (α=.005)
n.s. indicates corresponding comparison is not statistically significant

for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. Among respondents randomly assigned to 5 point scales, the slider left start group had low option selection rates that were, on average, at least 40% less than those for the slider bar middle or right start groups (p-values<.0001) and 23% less than those for the standard scale group (p-value=.0001). Respondents in both the slider middle and right start groups had low option selection rates that were, on average, at least 40% higher than those of the slider outside start group (p-values<.0001) and at least 29% higher than those of the standard scale group (p-values <.0008). The pattern of differences across scale types for the 11 point scale items was generally consistent with the findings for the 5 point items, although the overall magnitude of differences was generally lower and the number of significant differences fewer.[9]

[9] No significant differences were noted between the slider right, middle and outside start groups (all p-values>.08) nor between the slider left start and standard scale groups (p-value=.230). Respondents assigned to either the slider left start or standard scale groups had low option selection rates that were, on average, at least 20% less than those of either the slider middle or outside start groups (p-values<.0011) and at least 30% less than those of the slider right start group (p-values<.0001).

4.4.3 Satisfaction, brand performance and service preference measures

To explore how the patterns in response option selections might translate into differences in the actual substantive measures of interest, we also examined the relationship between the OSM, BPM and SPM measures and device type, scale type, and scale numbering along with all possible higher order interactions using general linear models computed separately for each measure at each scale length. Normality assumptions were investigated for each of these scales across the experimental conditions and no major issues were detected. The overall reliability for both the five and 11 point scale versions of the OSM and BPM measures, as measured by Chrombach's alpha, exceeded .90 with very little practical variability across the devices. Lower reliability measures were observed for both the SPM(5) and SPM(11) measures (.67 and .72, respectively) but again, very little practical differences in the reliability statistics were observed across the devices.[10]

Due to space considerations we now provide an overall summary of the separate models followed by more specific details for the analyses pertaining to the Brand Performance Measure (BPM). Additional information about any of the models can be obtained upon request from the lead author.

The profile plots for the overall means for the OSM, BPM and SPM outcome measures by scale type and device are displayed separately by scale length in Figure 4. Generally the OSM(5), OSM(11) and BPM(5) measures varied significantly across both scale type and device as main effects. As displayed in Figure 4 A, B and D, PC respondents reported, on average, higher values of these measures compared to Smartphone and Tablet respondents. Moreover, respondents in the slider right start group reported significantly lower measures, on average, than those in the slider middle start group, but both of these groups had significantly lower measures, on average, compared to those for the slider left and outside start and standard scale groups. Similar patterns in differences across scale types were also observed for the BPM(11), SPM(5) and SPM(11) outcome measures, but the but the magnitude and direction of the differences was impacted by th e specific combination of scale type and device (e.g. significant interaction between scale type and device in the models for these outcomes) as depicted in Figure 4 C, E and F. Overall, the findings for both the 5 and 11 point versions of the three scale measures were generally consistent with those reported for the middle and high response selection rate analyses.

[10] Lower reliability for the SPM is likely related to the inclusion of at least two items that asked respondents about service preferences that were in direct contrast to one another – namely one item asked whether or not a respondent preferred to work with the insurance agent directly and another question asked whether they would prefer to interact with the insurance company directly without going through an agent.

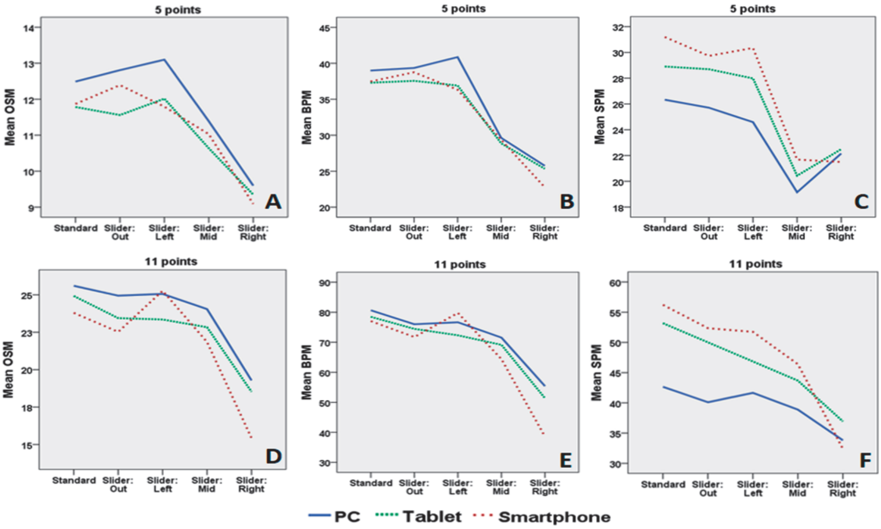


Figure 4 Mean values for the OSM (A (5 point) and D (11 point)), BPM (B (5 point) and E (11 point)) and SPM (C (5 point) and F (11 point)) measures for each scale type and device by scale length

The brand performance measure (BPM)

We found that BPM(5) values varied significantly by scale type (F(4, 1718)=115.17; p-value<.0001) and marginally significantly by device type (F(2, 1718)=5.22; p-value=.0055). None of the other main effects nor any of their interactions were found to be significant (all p-values >.22). As suggested by mean profile plot provided in Figure 4 B, on average PC respondents had BPM(5) values that were estimated to be about 2 scale units higher than those for Smartphone respondents (p-value=.0027) and no significant differences were detected between any other pairs of devices (p-values>.01). Estimated differences in BPM(5) values between scale types were notably larger than those across devices. The average BPM(5) value for the slider right start group was estimated to be roughly 14 units lower than the slider left and outside start and the standard scale groups, about 5 units lower than the slider middle start group (all p-values <.0001). The average BPM(5) value for the middle start group was also estimated to be about 9 points lower than the slider left and outside start groups as well as the standard scale group (all p-values<.0001) and no other significant differences between pairs of scale types were noted.

Differences in scale type for BPM(11), while generally consistent with those found for BPM(5), were moderated by the device used to complete the survey. In particular, we found that BPM(11) values varied significantly by device (F(2, 1726)=8.58; p-value=.0002) and type of scale (F(4,1726)=102.71; p-value<.0001) but also by the interaction of device and scale type (F(8, 1726)=3.26; p-value=.0011). Generally speaking, BPM(11) values were higher for PC respondents followed by Tablet, and then Smartphone respondents on all scale types except the slider left start group which was higher for Smartphone respondents on average, as indicated in Figure 4 E. As for scale types, the slider right start group had significantly lower BPM(11) values, on average, compared to any of the other scale types, but these differences varied in magnitude depending on the type of device. For example, for Smartphone respondents, the slider right start group had an estimated BPM(11) average value that was about 41 units lower than the slider left start group, 38 units lower than the standard scale groups and 33 units lower than the slider outside start groups. The differences in these groups for Tablet users was estimated to be 27, 21 and 23 units, respectively and for PC respondents 25, 21 and 20 units, respectively (all p-values <.0001).

The slider right start group also had significantly lower BPM(11) values, on average, compared to those for the slider middle start groups, but the magnitude of the estimated differences varied from 26 units for Smartphone respondents to 18 units for Tablet respondents to 16 units for PC respondents (all p-values <.0001). Significant differences were also noted for BPM(11) values between the slider middle start and standard group across the three devices and between the slider middle and left start groups for Smartphone respondents. The degree of these differences varied across the devices.[11]

4.4.4 Imputed versions of survey measures using slider starting position

The pattern of differences in the OSM, BPM and SPM measures across both scale type and device is generally consistent with the overall missing item patterns for the core scale items – namely more missing items for the middle and right slider positions with the degree varying by device type. For negatively skewed scale items, it seems reasonable that sliders with a right or middle starting position might have indicated the respondents' desired answer choices more consistently, and as such, respondents might not have realized a need to do anything more to register these choices but to click the "continue" button. To better understand whether some of the differences observed in the three outcome measures could be explained or

[11] The slider middle start group also produced significantly lower BPM(11) values, on average, compared to the standard scale group across all three devices with the magnitude of the difference varying from 12 units for Smartphone respondents (p-value<.0001) and 9 units for both Tablet (p-value=.0002) and PC (p-value=.0014) respondents. Finally, the slider middle start group was found to be about 14 points lower, on average, compared to the slider left start group among Smartphone respondents (p-value<.0001).

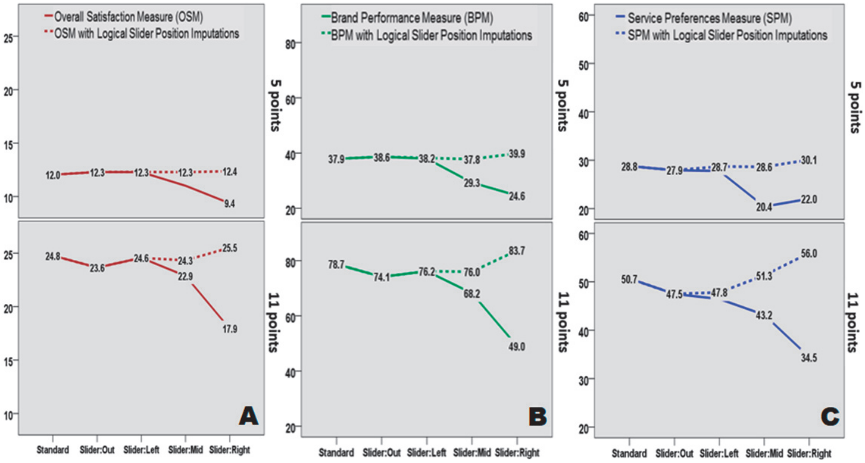


Figure 5 Summary statistics for the three survey outcome measures (A: OSM; B: SPM; and C: BPM) by scale length and scale type (solid lines) and their imputed versions (dashed lines)

adjusted for the impact of item missingness, we imputed the response value that corresponded to the slider's starting position whenever a respondent had a missing item for that scale item. The means for the recomputed "imputed" versions of the OSM, BPM and SPM measures (plotted as dashed lines) are displayed along with those from the original versions (plotted as solid lines) in Figure 5 A, B and C, respectively. For simplicity of display, these plots and analyses aggregated scale measures across device type.

What becomes quickly apparent from Figure 5 for each of the three measures across both the 5 and 11 point scale items is the considerably lower scale values of both the middle and the right slider start groups across for respondents for which scale measures could be computed. These figures represent the key findings of the last section with respect to scale type. What is also apparent is that the imputed versions for each of the three outcome measures generally fall more in line across the scale types. More specifically, for the OSM scale, there were no practical differences across scale types using the imputed versions for both the 5 and 11 point scales as seen in Figure 5 A (top and bottom, respectively). The imputed 5 point version of the BPM still had significant differences between the slider right start group and all other scale types but these differences were practically negligible; moreover no differences were detected between the slider middle start group and any of the other scale types, except for the slider right start group. A similar pattern was found for the 11 point BPM version as well except that the imputed version was significantly higher for the slider right start group compared to all the other scale types, but the magnitude of the overall differences has been attenuated. Finally for the SPM there are still significant differences between the right slider start scale type and the other scale types for both 5 and 11 point versions but the differences for the 5 point version are now practically negligible. The 11 point versions for the slider middle and slider right groups are still significantly different from the other groups, but the direction has also been reversed and the magnitude has decreased.

4.4.5 Preference for slider scales

To better understand preference and consistency rates (in the next subsection), the scale type factor was separated into two variables – scale input style (e.g. slider or radio button) and slider start position (e.g. outside, left, middle and right). Scale input style specifies the order in which the two versions of the OSI were presented – if scale input style is "slider" then the first OSI (and all other core scale items) was presented on the slider scale using a start position dictated by the slider position variable and the second OSI was presented using radio buttons and vice versa for the "radio buttons" input style. Preference rates for the slider input style were analyzed based on 2,649 (74%) respondents who declared a definitive preference for one of the two input styles using a logistic regression model that included device type, scale length, scale input style and slider position and all higher order interactions among these factors. The scale numbering factor was not included in this analysis to avoid possible sample size issues in the logistic regression model that incorporated the additional scale input style factor and was based only on those respondents who indicated a preference for one of the two input styles.[12]

Preference rates for sliders scales across device and scale input style are given in left side of Table 5 and in total, of the 2,649 respondents included in the analysis, 43% expressed a preference for slider scales. These preference rates varied significantly by both survey scale input style ($\chi^2(1)=319.73$; p-value<.0001) and device type ($\chi^2(2)=202.54$; p-value<.0001) but the differences across device were moderated by both the scale input style ($\chi^2(2)=15.69$; p-value<.0005) and the slider starting position ($\chi^2(3)=20.10$; p-value<.0003). None of the other main effects or higher order interactions were significant (all p-values >.04). The odds for preferring slider scales versus radio buttons across devices showed the same general pattern but were generally larger among respondents who completed the core scale items using slider scales compared to radio button scales. Smartphone and Tablet respondents completing the core scale items using slider scales had significantly

[12] We examined slider preferences across the levels of the scale numbering factor as well as separately by device type, slider start position groups and levels of the scale length factor and found no significant differences in the slider preference rates (all p-values>.05). Thus we suspect that pooling across scale numbering would likely have little impact on the substantive findings from the model.

Table 5 Preference rates for the slider input style based on 2,649 respondents who declared a definitive preference for one of the two input styles. Left: Preference for sliders by device type and input style; Right: Slider bar preferences by slider starting position

	Input style used to complete core scale items				Statistics for each device type		Slider input preference by slider start position for the slider version of the overall satisfaction item		
	Radio buttons		Sliders						
Device type	n	Prefer sliders (%)	n	Prefer sliders (%)	n	Prefer sliders (%)	Slider starting position	n	Prefer sliders (%)
PC	416	12.26	383	31.33	799	21.40	Outside	668	35.33
Tablet	455	26.81	443	71.33	898	48.78	Left	636	47.64
Smartphone	459	33.12	493	78.70	952	56.72	Mid	659	47.34
Statistics for each input style	1330	24.44	1319	62.47	2649	43.37	Right	686	43.44

higher odds of preferring a slider scales than PC respondents completing core scale items using sliders (p-values<.0001) with the odds of preferring sliders for Smartphone and Tablet respondents being an estimated 8.8 and 5.9 times the odds for PC respondents, respectively. Among the PC respondents assigned to complete core scale items using slider scales, we note that just less than one-third actually preferred sliders, but nearly three quarters of Smartphone and Tablet respondents assigned to slider scales for the core scale items expressed a preference for sliders over radio buttons (left side of Table 5). There was also no significant difference in the odds for preferring the slider input style to radio buttons between Smartphone and Tablet respondents completing survey scale items using sliders (p-value>.01). Among those assigned to the radio buttons survey input style significant differences in the odds of preferring slider versus radio buttons were also observed between PC respondents and both Smartphone and Tablet respondents (p-values<.0001) but not between Smartphone and Tablet respondents (p-value>.045). In particular, the odds for preferring slider scales for Smartphone and Tablet respondents were estimated to be 3.5 and 2.6 times that of PC respondents, respectively.

Differences in the odds for preferring slider input to radio button input were also observed between the different starting positions for the slider scales as indicated in the right side of Table 5. In particular the odds for preferring slider input styles among respondents with a left or middle starting slider scale were estimated to be about 1.7 times those for respondents using a slider scale with an outside start (both p-values=.0001). No significant differences in the odds of preferring slider input styles were observed among respondents completing survey items using slider scales with a left, middle or right starting position (all p-values >.04).

4.4.6 Consistency of responses across slider and radio button scales

From the 3,190 respondents for which concordance measures could be calculated, the exact concordance rate was 68.2% and the ±1concordance rate was 94.2%.[13] Concordance rates using both measures are given in Table 6 by device type and the experimental factors. From the logistic regression model relating exact concordance to device type, slider input style, scale length and slider position and scale numbering we found that these rates varied significantly by device ($\chi^2(2)=20.516$; p-value<.0001) and by scale length ($\chi^2(1)=175.811$; p-value<.0001). The exact concordance rates were not statistically different by scale input style, slider position or scale numbering and none of the higher order interactions between these and other effects were significant (all p-values>.024). The odds for exact concordance for PC respondents were approximately 1.6 times those for Smartphone respondents (p-value<.0001) and about 1.4 times those for Tablet respondents (p-value=.0012). No significant differences were noted for the odds for exact concordance between Smartphone and Tablet respondents (p-value=.2250). The odds for exact concordance for respondents assigned to the 5 point version of the OSI were estimated to be about 3.1 times those for respondents assigned to the 11 point version of the OSI (p-value<.0001) and these differences were consistent across device types.

[13] There were 210 respondents who did not answer the first Overall Satisfaction Item (OSI) and another 198 who did not answer the second OSI version. A majority of these missing items come from the slider right starting position group compared to the other starting positions and from respondents completing the survey by smartphone compared to other devices.

Table 6 Observed concordance rates between the overall satisfaction item presented as part of the main survey and again in an alternate format at the end of the survey. The value for the two items matched exactly for the exact concordance rates and matched up to 1 scale unit up or down for the second concordance measure

		Concordance rate between the two versions of the overall satisfaction item	
Group / Experimental factor	n	Exact	±1Concordance
Device type			
PC	1122	73.26%	97.06%
Tablet	1053	66.57%	93.92%
Smartphone	1015	64.24%	91.33%
Slider start position			
Outside	884	67.99%	93.21%
Left	882	67.57%	94.10%
Mid	790	70.51%	94.68%
Right	634	66.40%	95.11%
Scale length			
5 items	1576	80.27%	99.43%
11 items	1614	56.38%	89.10%
Scale numbering			
Numbered	1621	69.96%	95.56%
Not numbered	1569	66.35%	92.80%
Input style for core scale items			
Radio buttons	1610	68.63%	94.53%
Sliders	1580	67.72%	93.86%

5 Discussion

Several studies have found that slider scales, while engaging, can take longer to complete than comparable traditional radio button scales (Sikkel et al., 2014; Roster et al., 2015; Husser & Fernandez, 2013; Funke et al., 2011, among others). However in many of these studies, radio button completion times were compared to sliders with a left starting position. Our results for the completion times for the single continuous item “number of miles driven in the past year” were consistent with these studies in that the slider group had completion times that were longer, on average, compared to the group which entered their responses directly into an open-ended text box. Our results, for sliders with a left start also echo the findings from prior research in direction but the differences we observed were not statistically significant[14]. However, our findings for the other slider start positions, including most notably sliders with a right or middle starting position were in the opposite direction in that we found completion times for respondents in these two groups to be shorter than those for the standard scales, albeit not statistically significantly different. This opposing result could be directly related to the fact that we observed higher missing items from respondents from both the middle and right starting slider scale groups. In some cases, respondents in the right slider start group who were highly satisfied with their insurance provider might have taken much less time to answer the satisfaction questions simply because their responses corresponded to the slider starting position. As such respondents may not have taken the time to click on each item, but instead hit the next button for the survey to continue, resulting in missing data.

Throughout this paper we have presented empirical evidence showing that the slider starting position can greatly affect the amount of missing items and could impact measurement. As Funke et al. (2011) note "if the handle is placed at the position of a valid answer, intentional response and non-response cannot be distinguished." One starting position that would avoid this issue is outside or off of the slider itself. However, this choice requires more space for the overall slider graphic. While making the slider handle smaller to create more room for the actual slider bar itself might work for mouse interfaces, it might be less optimal for interfaces that rely on finger taps. In our study we also found that respondents completing scale items using an outside starting slider were the *least likely* to prefer slider scales compared to any other starting position[15].

Another option to remedy the missing item issue might require respondents to move the slider away from its starting position and then back to the response category to register the response. Such a requirement would however increase the num-

[14] We note had our study used the same Type I error rate for declaring significance as used in both of these studies (α=.05), then we would have also declared differences in completion times to be significantly lower for the left slider start group compared to the radio button group. Moreover, our results were based on the Geometric mean (natural logarithm transformed completion times) rather than the arithmetic mean and our analyses did not eliminate any outliers.

[15] Certainly a plausible factor in preference, or lack thereof, for slider scales with an outside start could be related to poor operationalization of this type of slider (slider handle doesn't appear in its entirety on the screen or isn't responsive to respondents actions). However, we believe this factor should contribute as most minimally given that we made every effort possible in the programming phase to ensure that this specific slider scale would be optimized for all three devices including positioning and sizing the slider handle in such a way that it would appear wholly on the screen and not interfere with the legibility of the scale point labels and numbers as displayed in Figure 1: C.

ber of taps required to complete the question from one to two for the slider scales compared to what is required for the radio button scale (Buskirk, 2015b). Such an approach was used by Sellers (2013) who compared slider bars scales with middle, left and right starts to radio buttons. They found that with a forced choice requirement, respondents in the right slider group reported higher right choice options and respondents in the left choice group reported more lower choice options compared to respondents in other groups. Contrary to the method employed by Sellers, we did not force respondents to confirm answer choices for which the slider was neither moved nor clicked and we observed that respondents in the middle and right slider start groups tended to select these answer categories significantly *less often* than any other scale group. Respondents in the right start slider scale group who registered answers for scale items moved the slider away from the starting position but ultimately did not move it back. This pattern was generally consistent across the three devices and both scale lengths; however, the pattern was much stronger with the shorter version of the scale. More specifically, the high option selection rates for those assigned to 5 point scales with middle slider scales were 25% higher than those from any other scale group. Respondents seeing 5 point scale items in the right slider group selected higher categories at rates that were between 8 to 50% *less* than those of any of the other scale groups. We also found that respondents in the middle slider start group also chose lower end options more often than any other scale type except the right slider start group. This finding replicates the pattern observed by Petersen et al. (2013) who reported higher amounts of “2s” and “4s” being selected on a five point slider scale that had a middle start compared to other non-slider presentations. The similarity in the percentage of respondents in the left starting slider and radio button groups choosing higher options for the core scale items echoes what Cape (2009) found in a study comparing left starting sliders with different labelling options to more traditional radio buttons. Specifically, Cape (2009) found that while distributional differences were noted for survey outcomes across different scale types, the “box top” or percentage agreeing with a statement, were nearly identical across the scale types. However, in our study we also saw contrasting results between the radio button group and both the middle slider group where, respondents had significantly higher “box top” rates, and the right slider group, where respondents exhibited significantly lower “box top rates.”

In addition to differences in response options and survey outcome measures, we also found differences in preferences for the slider scales. Such differences in preference rates by scale input style might reflect more of a conditioning effect in that respondents may likely prefer what they are comfortable with rather than something new. We expected that some respondents with radio button survey input style would, for example, express higher preferences for radio buttons when faced with a choice between those and a new slider version, and conversely for slider input styles. Indeed others have found somewhat similar results in experiments that simply asked satisfaction with sliders/radio buttons at the end of the survey experience without requiring respondents to choose between alternate methods of input. For example, While Cape (2008) found that compared to respondents using more traditional Likert scales, respondents who were presented questions using slider scales reported higher levels of satisfaction with it as an instrument to capture their true opinions. In our study we certainly saw evidence of a conditioning effect for preference as well in that those who were presented slider bar questions in the main experiment and then asked to complete an item using radio buttons generally expressed interest in sliders. However, they did not express this interest as consistently as those who completed standard scales in the experiment and then completed one additional slider item did for standard radio buttons (76% of respondents in the radio button version expressed interest for radio buttons compared to 63% of respondents in a slider group expressed interest for sliders. ($\chi^2(1)=53.11$; p-value<.0001). We also found that generally, the preference for sliders increased from PC to Tablet to Smartphone respondents but the degree of differences across devices was still influenced with the input style to which respondents were assigned. More work is needed to better understand whether preferences for sliders might be higher among PC respondents who have touchscreen monitors compared to mouse only input.

In summary, we found consistent patterns in missing item rates and lower, middle and higher response option selections for the respondents in the middle and right slider start groups compared to any of the other slider scale or radio button groups. These trends were generally consistent across devices, and were slightly more pronounced for 5 point compared to 11 point scales. Moreover, these differences were seemingly not impacted by whether scales were additionally numbered or not. The higher missing rates and lower levels of selecting higher categories across the scale items resulted in stark differences in three main survey outcome measures. While the slider start position based imputation resulted in fewer significant differences and practically small differences, it did not fully compensated for the item missingness – especially for the 11 point scale items. For each of the three survey measures, the imputed 11 point version produced overall scale measures for right and middle slider start groups that trended well above the general pattern for the remaining scale types and could give the indication that satisfaction was much higher than reality might suggest. Clearly, without the imputation, the right and middle start slider types generated measures of satisfaction that are likely to be too low. More work is needed to understand if such an approach can be applied uniformly for sliders with missing values or if it should be applied more judiciously. The outcome measures and more specifically, the individual items were generally expected to have a negative skew based on historical trends for similar customer satisfaction/loyalty items. Thus, many of the expected responses were in the upper region of the scales and the direction of item missingness and overall differences in measures tracked very closely to the expected response pattern. More work is needed to see if comparable results might be obtained for the middle and left slider start groups using scale items with an expected positive skew.

We note that our study has some clear limitations. Our consistent null findings for the scale numbering factor might be related to the fact that the numbering was added to scales that always included semantic labels. The labels, especially for the 5 point scales, might have been sufficient to overshadow any additional impact that numbering could have provided. For the 11 point scales we expected the numbering to have a more pronounced effect since these scales were only labeled at the two anchor points. The difference in scale labeling pattern across the two scale lengths might also confound differences observed for the scale length factor, but we note that the method used to label the 5 point and 11 point scales is generally consistent with typical uses in practice. We also note that while we were able to experimentally randomize respondents to receive different input styles and slider starting positions, scale lengths and scale numbering we had to embed the overall experiment within each of the three devices. Panel expectations and device ownership within the panel sourcing our sample precluded randomizing panelists to device type. Hence, the device used to complete the survey was taken as a natural blocking variable. In light of this, as one might expect, we found some natural differences in the ages of respondents using each type of device with PC respondents being older than tablet respondents and Tablet respondents being older than Smartphone respondents, on average. Differences in other demographic variables that were correlated with age were also found to vary similarly across the three devices and were consistent with other studies that also allowed respondents to self-select their device (see Baker-Prewitt & Miller, 2013 for example). So in sum, when interpreting the device specific comparisons and effects reported in this study one has to consider that they could represent not only device but also the cluster of demographic variables related to the usage of that device.

While sliders may offer more engagement for respondents they come at a cost when thinking about implementing them across many device types with differing space and hardware constraints. And no matter how engaging sliders can be compared to radio buttons, missing items still persist and can certainly be a function of starting position as well as the underlying distribution being estimated. Preference for sliders tends to skew towards those using mobile devices to complete surveys, but this preference doesn't overwhelm previous experience with radio buttons. Even though sliders might be more preferred by smartphone respondents, they also add to the completion times, overall. And given that many studies have consistently shown that surveys tend to take longer on smartphones compared to PCs (Buskirk, 2015b; Wells et al., 2014), it's hard to know whether the positive impact sliders have on engagement would outweigh or be nullified by the negative impact of longer surveys. More work is needed to understand just how slick a slider needs to be to hit this sweet spot.

References

Baker-Prewitt, J. & Miller, J. (2013). What Happens to Data Quality When Respondents Use a Mobile Device for a Survey Designed for a PC. Paper presented at the 2013 CASRO Online Research Conference, San Francisco, March, 2013. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F-6CEF475/Paper_-_Jamie_Baker-Prewitt_-_Burke.pdf

Buskirk, T. D. (2015). The Rise of Mobile Devices: From Smartphones to Smart Surveys. *The Survey Statistician, 72*, 25-35. Available at: http://isi-iass.org/home/wp-content/uploads/N72.pdf

Buskirk, T. D. (2015b). Going Mobile with Survey Research: Design, Data Collection, Sampling and Recruitment Considerations for Smartphone and Tablet Based Surveys. Shortcourse presented at the Journal of Official Statistics Anniversary Conference, 2015. Stockholm, Sweden. Available at: http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/JOS-2015/buskirk-FINAL-participant-JOS2015ShortCourseBuskirkJUNE2015.pdf

Buskirk, T. D. & Andrus, C. (2012). Smart surveys for smart phones: Exploring various approaches for conducing online mobile surveys via smartphones. *Survey Practice, 5.* Available at: http://surveypractice.wordpress.com/2012/02/21/smart-surveys-for-smart-phones/

Buskirk, T. D. & Andrus, C. (2014). Making mobile browser surveys smarter: Results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods*, 26, 322-342.

Buskirk, T. D., Michaud, J., & Saunders, T. (2014). Swipe, Snap & Chat: Mobile Survey Data Collection Using Touch Question Types and Mobile OS Features. Paper presented at the 39th Annual Conference of the Midwest Association of Public Opinion Research, November 21-22, 2014, Chicago, Il. Available at: http://www.mapor.org/confdocs/absandpaps/2014/1C1_Buskirk_slides.pdf

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*. Available at: http://surveypractice.wordpress.com/2010/12/08/devicerespondent-has-used/

Cape, P. (2009). Slider Scales in Online Surveys. Paper presented at the 2009 CASRO Panel Conference, Feb. 2-3, 2009 New Orleans. Retrieved on August 31, 2015 from: http://www.surveysampling.com/ssi-media/Corporate/white_papers/SSI-Sliders-White-Pape.image

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, *24*(2), 227-245.

Courtright, M. Saunders, T. & Tice, J. (2014). Innovation in Web Data Collection: How 'Smart' Can I Make My Web Survey? Paper presented at the CASRO Technology and Innovation Event, May, 2014, Chicago. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/97E56036-D4ED-4552-8A5F-E0A75899AEA8/2T1.1_-_T_Saunders_-_Maritz_-_M_Courtright_-_Research_Now_-_J_Tice_-_Decipher.pdf

Derham, P. A. J. (2011). Using preferred, understood or effective scales? How scale presentations effect online survey data collection. *Australasian Journal of Market & Social Research, 19*(2), 13-26.

Dobronte, A. (2012, August 21). Likert scales vs. slider Scales in commercial market research. Retrieved June 27, 2015, from https://www.checkmarket.com/2012/08/likert_v_sliderscales/

Funke, F, Reips, U.-D., & Thomas, R. K. (2011). Sliders for the Smart:Type of Rating Scale on the Web Interacts With Educational Level. *Social Science Computer Review*, 29(2), 221-231.

Husser, J. A. & Fernandez, K. E. (2013). To click, type, or drag? Evaluating speed of survey data input methods. *Survey Practice, 6*(2), 1-7.

Keusch, F. & Zhang, C. (2014). A review of Issues in Gamified Survey Design. Paper presented at the 2014 Midwest Association of Public Opinion Research Conference, November 21-22, 2014, Chicago. Available at: http://www.mapor.org/confdocs/absandpaps/2014/4A2_Keusch_slides.pdf

Link, M. W. & Buskirk, T. D. (2012). The role of new technologies in powering, augmenting, or replacing traditional surveys. Short-course presented at the annual meeting of the American Association for Public Opinion Research, Orlando, FL.

Michaud, J., Buskirk, T. D., & Saunders, T. (2014). You CAN Touch This: An Experiment to Compare Computer and Mobile Surveys Using Touch Friendly Question Types." Paper presented at the 69th Annual American Association of Public Opinion Research Concerece, May 15-18, 2014, Anaheim, CA.

Peterson, G., Mechling, J., LaFrance, J., Swinehart, J., & Ham, G. (2013). Solving the unintentional mobile challenge. Paper presented at the CASRO Online Research Conference, March, 2013, San Francisco, CA. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F6CEF475/Paper_-_Gregg_Peterson_-_Market_Strategies_International.pdf

Puleston, J. (2011, March 14). Sliders: A user guide. Retrieved June 27, 2015, from http://question-science.blogspot.com/2011/02/slider-how-to-use-them.html

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring Slider vs. Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice, 11*(1), Article D1. Accessed on August 30, 2015 from: http://jrp.icaap.org/index.php/jrp/article/view/509/413.

Sikkel, D., Steenbergen, R., & Gras, S. (2014). Clicking vs. dragging: Different uses of the mouse and their implications for online surveys. *Public Opinion Quarterly, 78*, 177-190.

Sellers, R. (2013). How sliders bias survey data. *Alert!, 53*(3), 56-57.

Toepoel, V. & Funke, F. (2014). Investigating Response Quality in Mobile and Desktop Surveys: A Comparison of Radio Buttons, Visual Analogue Scales and Slider Scales. Paper presented at the 2014 American Association of Public Opinion Research Conference. Anaheim, CA, May, 2014.

Wells, T., Bailey, J., & Link, M. W. (2014). Comparison of Smartphone and Online Computer Survey Administration. *Soc. Sci. Comput. Rev. 32*(2), 238-255.
DOI=10.1177/0894439313505829. http://dx.doi.org/10.1177/0894439313505829