

White Paper



Slider Scales in Online Surveys

By Pete Cape, Global Knowledge Director, SSI

Online surveys today are often perceived as dull, repetitive, and lacking in involvement and interactivity. Some have argued that market research is not, nor is it intended to be, a form of entertainment but is serious scientific inquiry. Others have looked, more pragmatically perhaps, at what the Internet has to offer and how these techniques might be applied to answer market research problems. The latter offers the greater potential. Market research techniques have evolved over decades to encompass the possibilities within the confines of the methodologies employed.

The Internet allows researchers to, once again, “re-think the question set,” and incorporate elements such as Adobe Flash, a multimedia platform for adding animation and interactivity. Its almost universal installation on PCs has opened up the opportunity to produce question and answer styles that, if nothing else, look engaging and offer a degree of interactivity beyond merely answering questions.

Many companies providing programming and hosting services, as well as research institutes themselves, offer Flash toolkits to replace clumsy or time-consuming traditional questioning methods. Of the tools in the toolkit, sliders as replacements for scales were among the first to be developed. Yet they remain, at least according to anecdotal evidence, the least popular.

This unpopularity probably stems from a lack of understanding of precisely how the slider is being perceived and used by the respondent along with very real concerns about loss of comparability with previous data which may have been collected via a completely different mode. As an industry, we are somewhat conservative and, in the case of scales, there is a feeling that “if it ain’t broke, don’t fix it.”

This paper attempts to answer a number of questions about the “standard” 5-point Likert Scale and how a Flash-based alternative might perform in terms of data collected, levels of engagement engendered, and satisfaction with the instrument on the part of the subject, that is the respondent.

Graphical Representation vs. Graphical Imagery

When considering using anything other than a text-based question and answer format, it is important to make a distinction between images and representations. Using images to illustrate concepts is fraught with difficulties for the researcher. People, that is respondents, find it impossible to separate the image from the concept. This of course leads to ambiguity. The researcher does not know whether the respondent is thinking of the general (the concept) or the specific (the image). A typical question that the unwary researcher might immediately seek to attach images to might be: “What type of holiday do you prefer?” Having been presented images of summer holidays (for example) the respondent could be reacting to the image and therefore a sub-type of summer holiday it represents.

Both the images below could be used to illustrate a summer holiday:

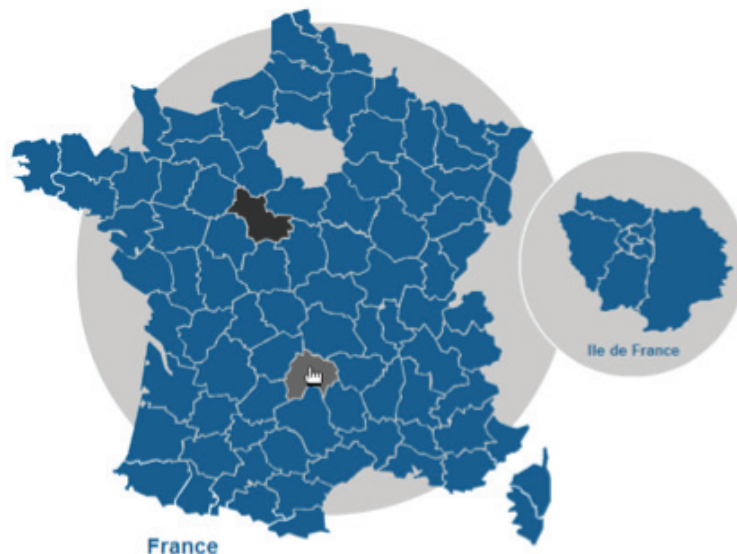


(continued)

This author for one finds one of them considerably more appealing than the other and would be more likely to choose any alternative than the crowded beach. In effect the question has become, “Which of these pictures of holidays do you prefer?” and the text answers are merely ciphers. A neutral choice of illustration is almost impossible when the question concerns an abstract concept (which indeed “summer holiday” is).

A second flawed use of graphical imagery, and one that my own company has been guilty of using in our own online demonstration questionnaire, involves the use of interactive maps. The question then becomes one of geographical knowledge, often to quite a sophisticated level.

Knowing that Toulouse is the capital of the French department Haute-Garonne is not the same as being able to pinpoint Haute-Garonne on a department map of France:



Images must always be treated with caution and with a critical eye as to what the image may represent to different people and to different cultures.

Graphical representation is different and is often more intuitive for the respondent. Ranking questions, for example, work much better in a graphical form than in a textual form. It is a very simple task for people to sort items into a rank order, simpler even than trying to pick the number 1 item from a list of 12 over the telephone. Other graphical representations—virtual shelf tests, “page turning” magazines, dragging and dropping—generally are all advances in terms of efficiency, interactivity, and, most probably, data accuracy.

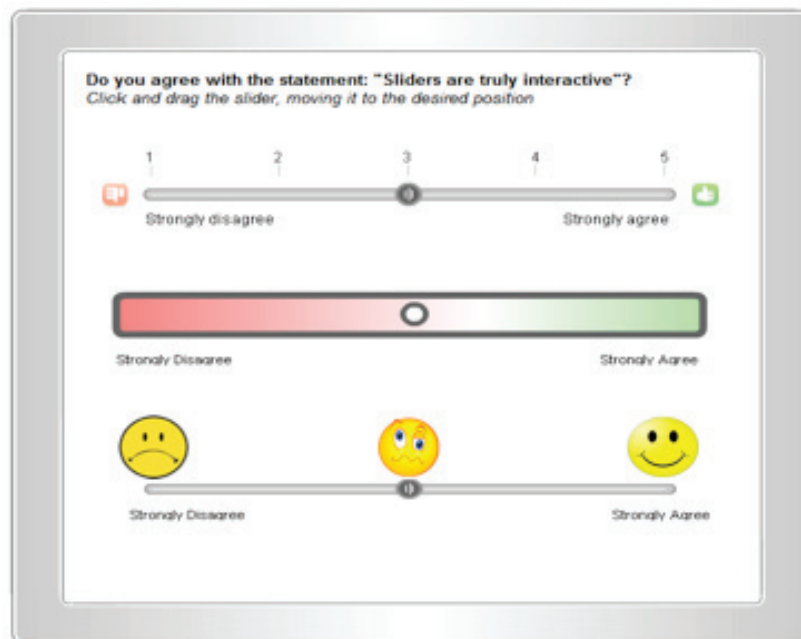
Scales

Scales fall into a space somewhere between an image and a representation. Scales are not real in that you cannot touch, smell or feel them; they are imaginary constructs made real by repeated use within a specific cultural context. Humans have no need for fixed scales in their own behavior. We can each

(continued)

assess a range of products on a range of criteria and quickly come to a purchase decision based on the performance of the product on each criteria and the relative importance to us of each criteria. We can do all this without going through some formal scoring and assessing procedure. The problem for researchers is that this human behavior is very hard to self-describe and categorize, hence our need to formalize the process, however unreal it may be.

The act of writing down the scale, particularly in the case of online research, brings in the dimension of imagery. In the early days of online research, the use of illustrative images was often mooted for Likert Scales. The image below shows examples of illustrated sliders.



Notice on the first slider the use of thumbs up/thumbs down. In the context of the slider it is supposed to represent agreement. Yet thumbs up/thumbs down in many cultures actually signifies approval. Approval, of course, is not the same as agreement. Imagine that the statement being rated was something unpleasant—an unpalatable but true political fact, for example. Would some respondents be somehow (even unconsciously) reluctant to give this fact the thumbs up? The same is true for the smiley scale. Does agreeing with something actually always make you happy? Other criticisms of the smiley scale might be “inappropriateness” or even “infantilism” depending on the subject of the research. Red and green signifies stop and go, but, again—it is not always the case that one would want to stop something you disagreed with.

No doubt Rinsus Likert would have been horrified, back in 1932, if it had been suggested that he replace his simple agree/disagree scale with smiling faces! But, would he have wanted to present his scale as a continuum had it been possible to collect and analyze the resultant data to the sort of degree that is possible nowadays? The true potential of the slider may well be in the granularity of detail that it can collect rather than the “fun” that can be had moving the slider and seeing things change.

(continued)

Likert Scales

There are any number of issues with Likert Scales and the way in which market researchers have used them. Our experiment concerns two key questions about the traditional Likert Scale:

1. **Equivalency of ratings:** If item A is rated “agree slightly” and item B is also rated “agree slightly” does the respondent agree to both items to the same degree? In short, is the Likert Scale too blunt an instrument to detect the subtle differences between items?
2. **The spaces between:** Is there some level of agreement between “agreeing slightly” and “agreeing strongly?” Is there some level of agreement below “agreeing slightly?” Do we force a person into stating something that is not their true opinion because we offer too few alternatives?

Once we understand whether there is a need to replace the traditional Likert Scale, we then need to assess a number of representations of Likert Scales using sliders.

For the first part of the experiment, a random sub-sample of respondents was presented with a traditional 5-point Likert Scale containing 4 items. After completing the exercise, respondents were asked to what extent the instrument had allowed them to accurately give their true opinion. (The irony of using a fixed scale to collect this information given the subject matter of the paper is not lost on the author.)

Subsequently, they were presented with the items again, showing first all those that they agreed strongly with, then those they agreed slightly with, and so on. Respondents were offered the opportunity to re-score each item using 5 points above and/or below the original stated answer (as appropriate), approaching but not reaching the adjacent level(s) on the Likert Scale.

A set of statements about personal values was used in the experiment:

Statement 1: The highest standards of morals and ethics is the most important thing in life.

Statement 2: Acquiring wealth or material possessions is what will make me happiest.

Statement 3: I would always put family before friends.

Statement 4: Under all circumstances you should obey the law.

The response to the traditional instrument was generally positive as can be seen in Table 1.

Table 1: Percent Able to Give Opinion Completely or Very Accurately Using Traditional 5-Point Likert Scale

	UK	USA	Germany	China
Completely accurately	36%	41%	18%	21%
Completely or very accurately	68%	71%	57%	46%

(continued)

Based on the response to this question, one might expect the majority of people would not take advantage of the opportunity to re-score the items. However, a large number of people did elect to change their rating as can be seen in Table 2.

Table 2: Percent Electing to Change Stated Opinion

	UK	USA	Germany	China
Statement 1	60%	52%	47%	46%
Statement 2	45%	51%	50%	52%
Statement 3	42%	36%	37%	51%
Statement 4	54%	54%	51%	42%

It is interesting to see which responses were more or less likely to encourage re-ratings. Taking just statement 1 and looking across the whole dataset, we can see that the highest volume of re-scoring comes from those originally electing a rating of “slightly” (irrespective of whether this was agreeing or disagreeing). Some 70% of those disagreeing slightly and 75% of those agreeing slightly chose to move their answer either towards disagreeing/agreeing strongly or neither/nor. This clearly demonstrates a weakness in the Likert Scale in capturing responses that are neither strong nor neutral. The true meaning of agree or disagree slightly must be some continuum between a strongly held opinion and no opinion at all.

Examination of the neither agree/nor disagree option is fascinating. On statement 1, 42% of respondents chose to move their response towards either agreeing or disagreeing. This clearly implies that there is a level of agreement/disagreement below “slightly” that is better described as neutrality rather than “slightly.”

The strong reactions are not immune to modification but we saw very different reactions depending on which end of the scale the original answer was given. Half of those who originally said they disagreed strongly with statement 1 subsequently moved their answers towards disagree slightly. In contrast “only” 23% of those who had said they agreed strongly with the statement subsequently modified their opinions towards slightly agree. If it is the case that strong agreement is a position easier to hold than strong disagreement, it could be argued that these two verbal labels are not opposites.

From these results we can clearly see that respondents have a finer definition of agreement/disagreement than the instrument currently allows us to collect. An examination of responses at an individual level should enable us to see whether the respondent is recalibrating the entire scale to reflect their own personal view of what constitutes “strong” opinions and all points in between or if the respondent is taking the opportunity to show the subtle differences in opinions held across all items. If the respondent is doing the latter, then the rescoring exercise could, in effect, produce a ranking of items according to strength of agreement, at the individual level. This level of data is almost impossible to obtain from traditional scales when so many items are so often given the same score.

(continued)

If we examine all those who gave the same rating to all 4 items, we see a mixture of rescoring and separation of items. Of the 44 cases of zero variance, just over half (55%) did not change their scoring at all. Five people did change their scores and yet still kept each of the items at the same rating. Interestingly, 2 people “crossed-over”—one who had originally marked all 4 as disagree strongly (score 1) moved to a score of 1.83 for all items; the other who had originally scored all items “disagree slightly” (score 2) moved them all to 1.33. A large minority (34%) took the opportunity to make a distinction between all or some of the items.

We also had 22 cases where all 4 items had been given a different score. Of these, only 2 people elected to keep the original scores. The rest changed their scores to better reflect their own opinion.

The evidence suggests that the traditional Likert Scale is a somewhat blunt instrument and that there is not an equivalency between items that have been assessed as the same. We have also seen that the Likert Scale is perceived as a reasonably accurate instrument to collect opinions. In contrast, respondents who were given the slider scale to use (irrespective of its design) reported higher levels of satisfaction with this instrument as a means of capturing their true opinion.

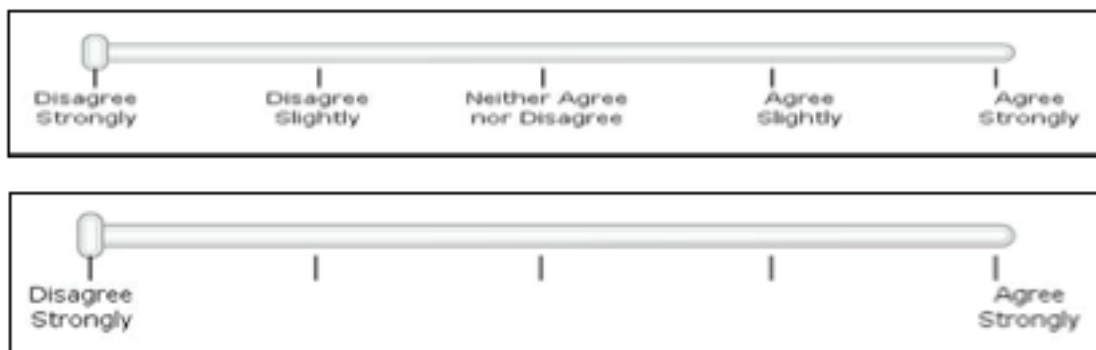
Table 3: Percent Able to Give Opinion Completely or Very Accurately Using Slider Likert Scale

	UK	USA	Germany	China
Completely accurately—slider	43%	54%	32%	21%
Completely/very accurately—slider	72%	82%	68%	47%

Slider Likert Scale

The next part of our experiment concerns what the impact on the data will be in varying the design of the slider. We used four different designs. All were text anchored at the ends of the scale and all eschewed the use of visual, illustrative elements.

The first slider has all the traditional points labelled:



(continued)



The final slider showed a numerical score between 1 and 5 to denote precisely where on the scale the slider was positioned:



In all instances, the slider was positioned initially at the “disagree strongly” point. Error checking code was employed in the case of the slider not being used to confirm that it was the respondent’s intention to disagree strongly.

The starting position of the slider may have some biasing effect on the response, similar to the order of response options in traditional Likert Scales. An examination of this however was not part of our experiment.

Our first observation of the data is that respondents use the mark points where they are available. Interestingly, they also use the numerical equivalents when the score box is displayed. This could be because they have prior experience of seeing numbered Likert Scales.

Table 4: Numbers (grossed to equal bases) Marking Specific Points—Statement 1

	Disagree Strongly	Disagree Slightly	Neither/Nor	Agree Slightly	Agree Strongly
Traditional	12	40	105	231	201
Re-Scored Traditional	6	12	61	57	154
Labeled Slider	4	25	77	125	129
Mark Slider	1	17	92	88	114
Blank Slider	1	8	27	21	97
Slider with Score	4	10	51	47	124

(continued)

Each slider design produces a different data distribution. This would lead us to suggest that design is important and caution must be exercised. We would certainly advise against overly visual/illustrative sliders and recommend consistency both within a questionnaire and across waves of data collection.

Wherein lies the truth for Statement 1? This, of course, we cannot say. Market researchers use a number of descriptive statistics to place meaning on the data and to put it into usable context.

When the results for Statement 1 are summarized to agree/disagree, few researchers would have difficulty saying that the data, however collected (Table 5) is essentially telling the same story.

Table 5: Netted % Disagree and Agree—Statement 1

	Disagree	Agree
Traditional	9%	73%
Re-Scored Traditional	6%	73%
Labeled Slider	8%	70%
Mark Slider	6%	65%
Blank Slider	6%	77%
Slider with Score	6%	73%

In terms of the mean score (and it should be acceptable to produce a mean score based on slider data) the answers are again similar.

Table 6: Mean Scores Across All Methods—Statement 2

	UK	USA	Germany	China
Traditional	2.27	2.32	2.50	3.79
Re-Scored Traditional	2.60	2.28	2.50	3.87
Labeled Slider	2.56	2.64	2.74	3.89
Mark Slider	2.33	2.35	2.46	3.61
Blank Slider	2.32	2.39	2.29	3.62
Slider with Score	2.34	2.44	2.33	3.56

(continued)



We can clearly see the similarity between the UK, the USA and Germany (who all disagree somewhat) and the difference compared with China (who tend to agree somewhat). This is true irrespective of the data collection instrument. When we consider that the statement is “Acquiring wealth or material possessions is what will make me happiest,” the answers also make sense in context.

What does change is the size of the “Top Box,” which will tend to reduce. While all items may not reduce by the same amount, we can imagine a similar rank ordering of items resulting from either method. The reduction in Top Box scores ought to be a one-off data change, but this may not make it acceptable to clients. The whole concept of “Top Two Box” (apart from simple agree/disagree netting) disappears as soon as a continuum is introduced. With a continuum of data, it is possible to start using descriptors such as quartiles and deciles, although these are both hard to convey and difficult to use in any practical sense.

Finally, as was pointed out in relation to the re-scoring exercise, the sliders data (at the individual level) is much better at differentiating opinions across the items. For many applications, this represents a clear advantage over the traditional methodology.

Engagement

As researchers, we are urged nowadays to make our surveys more engaging and interactive. Whether or not market research is part of the entertainment business or whether the attention span of the average respondent is contracting is outside the scope of this paper. However we do recognize that some of the things we do are boring, certainly repetitive and, to be frank, often overly long. Flash alternatives are often positioned as more engaging and more interactive.

The Flash sliders do not improve the length of the survey. The 4-statement battery took, on average, 20 seconds longer to complete using the slider. Part of this additional time may be due to unfamiliarity with the instrument and may reduce over time. It also might not be a bad thing to add some sliders to grid questions. We know from feedback from many studies on respondent attitudes that these types of questions are among the least popular and that attention levels to the questions are not always as high as researchers would like.

We asked a series of questions at the end of the survey to gauge respondents’ feedback about the interview itself. Data in Table 7 compares those who were not exposed to a slider (labelled “standard”) with those who were.

(continued)

Table 7: Qualitative Feedback on Survey

Interesting Survey	UK	USA	Germany	China
Agree—Traditional	93%	90%	90%	63%
Agree—Sliders	97%	97%	94%	71%

Enjoyable Survey	UK	USA	Germany	China
Agree—Traditional	92%	89%	83%	73%
Agree—Sliders	97%	96%	89%	81%

Will Take Next Survey	UK	USA	Germany	China
Agree—Traditional	99%	98%	98%	98%
Agree—Sliders	99%	99%	99%	97%

The results do show an increase in interest in the survey and enjoyment in taking it, which may well be linked, when there are sliders rather than traditional Likert Scales. The sliders do little to increase engagement overall when measured by willingness to take the next survey.

Summary

The traditional Likert Scale is, in many ways, a rather crude instrument. It can be shown that a respondent’s true opinion can lie in the spaces between the allowable answers. It may even be the case that the true opinion lies beyond the traditional end points.

That said, however, the scale is not “broken.” Respondents feel that it does reflect their opinion well. To what extent this is driven by familiarity is a matter of conjecture.

Certainly, it implies an equality between similarly coded items that is simply not the case. It also overstates the strength of feeling on issues. Researchers attempting to assess qualitatively the true meaning of the level of and strength of feeling of “strongly agreeing,” for example, will be mistaken.

(continued)



Slider scales do not produce means or levels of agree/disagree that are markedly different from traditional scales. This implies that sliders can be substituted for traditional scales without too much data inconsistency. Any researcher analyzing top box scores needs to be aware that these will decrease if the scale type is changed.

Where the slider scale is superior to the traditional scale is in its ability to show the subtle difference between items. If the research design calls for this feature, then sliders would provide the tool without the need for secondary ranking questions designed to separate equally rated items.

It should be noted that our findings also have implications for those who would look for poor quality respondents by examining straightlining behavior in grids. Where it is possible or reasonable for a single person to agree (or disagree) with all items, any apparent straightlining may be the fault of the instrument rather than of the respondent. The instrument simply does not allow the respondent to state his true differences in opinion between each item in the grid.

Care must be taken in the design of the slider itself to avoid bias associated with visual imagery. The slider scale does not dramatically improve levels of engagement and (at least at first) takes longer to do. Finally, in our quest for “engagement” we must not forget that slider scales will not make long, dull, repetitive grids interesting.

Technical Notes

SSI conducted 4,261 online interviews in August 2008 among panelists belonging to our proprietary SurveySpot and OpinionWorld panels; 862 interviews were conducted in the UK, 854 in the USA, 1,450 in Germany and 1,095 in China.

This paper was first presented at the CASRO Panel Conference in 2009.