Did You Misclick? Reversing 5-Point Satisfaction Scales Causes Unintended Responses

Martin Pielot
Google
Munich, Germany
mpielot@google.com

Mario Callegaro
Google
London, UK
callegaro@google.com

ABSTRACT

When fielding satisfaction questions, survey platforms offer the option to randomly reverse the response options. In this paper, we provide evidence that the use of this option leads to biased results. In Study 1, we show that reversing vertically oriented response options leads to significantly lower satisfaction ratings – from 90 to 82 percent in our case. Study 2 had survey respondents verify their response and found that on a reversed scale, the very-dissatisfied option was selected unintentionally in about half of the cases. The cause, shown by Study 3, is that survey respondents expect the positive option at the top and do not always pay sufficient attention to the question, combined with the similar spelling of satisfied and dissatisfied. To prevent unintentional responses from biasing the results, we recommend keeping the positive option at the top in vertically-oriented scales with visually-similar endpoint labels.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**;

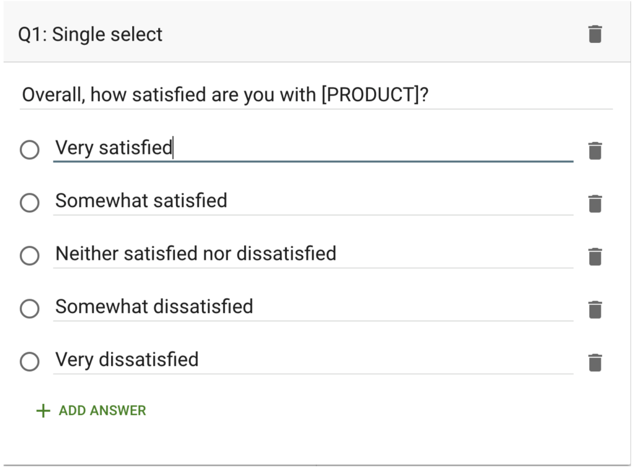
KEYWORDS

usability, satisfaction, satisfaction surveys, response bias, survey response option order

ACM Reference Format:
Martin Pielot and Mario Callegaro. 2024. Did You Misclick? Reversing 5-Point Satisfaction Scales Causes Unintended Responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613904.3642397

1 INTRODUCTION

Product teams frequently rely on customer feedback metrics as a leading indicator of future performance (Morgan and Rego [22], de Haan et al. [7]). One of the staple metrics is asking for the satisfaction of users with a product. One canonical form to elicit this sentiment is a bipolar-ordered response scale (Smyth et al., [26]) that offers 5 response options, ranging from *very satisfied* to *very dissatisfied* [see Figure 1].



Overall, how satisfied are you with [PRODUCT]?

○ Very satisfied
○ Somewhat satisfied
○ Neither satisfied nor dissatisfied
○ Somewhat dissatisfied
○ Very dissatisfied

Figure 1: Example of a canonical satisfaction survey being entered into a survey tool.

The responses to such scales drive strategy. A drop of a few percent points in satisfaction may trigger a serious re-prioritization of road maps and redirect the spending of millions of dollars. Thus, it is critical to capture user sentiment in a valid and reliable way.

When designing a survey, the researcher needs to decide how to display the response scale. As more people take surveys on their mobile phones, we see a shift from horizontally- to vertically laid-out response options (DeCastellarnau [8]).

A secondary decision is how to order the response options. Advanced survey platforms (e.g. Qualtrics[1], QuestionPro[2] and SurveyMonkey[3]) contain an option to randomly reverse response options called flip choices or flip option or randomly reversed [see Figure 2]. That means, half of the respondents will be shown the survey with the response option *very satisfied* on top, the other half will see the response options in reverse, that is, with the *very dissatisfied* response option being on top. One rationale for enabling this option is to counterbalance order effects as some survey platforms advise (e.g. QuestionPro[4]).

In this paper, we will share evidence that the use of randomly reversing response options introduces a significant fraction of invalid results to the survey.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642397

[1] https://qualtrics.com/
[2] https://questionpro.com/
[3] https://surveymonkey.com
[4] https://www.questionpro.com/blog/alternate-flip-options-scale/

We conducted 3 survey studies with a total of about 8,000 useful responses, finding that:

- When response options are reversed, the *very dissatisfied* option is picked 10 times more frequently (5% instead of 0.5%), resulting into an 8% point drop in top-2-box satisfaction (82% instead of 90%).
- When respondents who rated their satisfaction on a reversed scale were asked to verify their selection, about half of the respondents who had picked the *very dissatisfied* option indicated that they had wanted to select *very satisfied* instead. Our study is the first to explicitly ask respondents to verify their response.
- Open-ended feedback indicates that the reasons are that (1) the participants rushed their response, (2) they expected the positive response option on top, and (3) since *dissatisfied* and *satisfied* are spelled almost identically, they had not noticed the scale reversal.

On the basis of these findings, we recommend to keep the response options of questions where the end points are spelled similarly in the order where the positive option is displayed on top.

2 RELATED WORK

In HCI research studies, we often have the luxury to use elaborate tools to understand the user sentiment about prototypes or designs (SUS [3], UMUX [9], AttrakDiff [15], UEQ [20], UMUX-LITE [21], VisAWI [23], SMEQ [25]). In industry settings, where intercept surveys (e.g. Müller and Sedley [24]) are used, there is a need for keeping surveys short (e.g. Finstad [9]). According to de Haan et al. [7], the most commonly-used items in industry research are Satisfaction, Net Promoter Score (NPS), and Customer Effort Score (CES). Satisfaction surveys are commonly taught in User Research methods textbooks (e.g. Baxter, Courage & Caine [2]; Goodman, Kuniavsky, & Moed, [12]; or Jarrett [16]).

In an extensive literature review on the effect of the order of response options, Smyth et al. [26] remind us that manipulating the order can have a statistically significant effect on the response distributions. They argue that when an effect is found for vertical orientation, it can be explained by two theories: *satisficing* and the heuristic of *Up Means Good*.

Satisficing [17–19] is the notion that not all respondents will put the same amount of effort in answering questions by following the four steps cognitive processes necessary to answer a survey item: comprehension, retrieval, judgment & estimation, and reporting the answer [27]. Specifically *weak* satisficing implies carrying out these steps sloppily, while strong satisficing implies skipping retrieval and judgment altogether [18]. Satisficing is likely to occur if (1) the motivation is low, (2) ability are low, or (3) task difficulty [1]. In in-product surveys, respondents often have low motivation and their cognitive resources may be consumed by other tasks. Thus, we expect satisficing to affect responses of in-product surveys.

Up Means Good was introduced by Tourangeau, Couper and Conrad [28], inspired by the insight that happiness and sadness are often associated with up or down (e.g. being uplifting or being down), respectively (Carbonell [5]). In one experiment, Tourangeau et al. [28] found that when an unfamiliar car brand is placed higher in a list, people are more likely to assume that it is expensive. In another set of experiments, Tourangeau et al. [29] manipulated the position (high on the screen or low on the screen) of items on different topics (Congress and HMOs, a variety of foods, and six physician specialties). They found that the ratings are significantly more positive when the item appears in a top position on the screen than when it appears at the bottom.

When looking specifically at response scale options, Garbarski, Dykema and Schaeffer [11] found that when online respondents were shown the very commonly-used [10] health scale ("*In general would you say your health is...*") they rated their health more positive when the response option *Excellent* was at the top than when it was at the bottom (Mean 3.47 vs. 3.30 p < .001). Recently, Yan [30] published an eye tracking study where the direction of a 5 point fully labeled vertically oriented satisfaction scale was reversed. When the scale presented with "Very dissatisfied" first, respondents had significantly more fixations (average of 5.6 vs. 3.8, p=.03) and longer fixations (2.7 ms. vs 1.7, p=.04) than when "Very dissatisfied" was presented last (Table 1, Yan, 2023) thus indicating higher cognitive demand.

The closest experiments to our study were conducted by Smyth et al. [26] where they manipulated the order (positive to negative) of a five point vertically oriented satisfaction scale. In seven of eight separate conditions, the mean rating was significantly higher when *very satisfied* was shown at the top (Table S2 in Smyth et al online supplement).

If *Up Means Good* explains the above findings, what happens when the heuristic is not followed? One explanation supported by the literature is that respondents might experience higher cognitive load: Christian, Parsons and Dillman [6] found that respondents took longer to answer a negatively oriented vertical response scale in comparison to when it was positively oriented (trimmed mean of 13.1 vs. 10.1 seconds p. < 0.001). In the previously discussed experiments, Smyth et al. [26] found that there were more answer changes when the label *very dissatisfied* was shown on top of the scale. Callegaro [4] argues that these findings are evidence for confusion.

However, this reasoning relies on imperfect proxies (slower response speed and the number of answer changes) for *confusion*. The gap in the literature that this work addresses is to (1) explicitly probe for instances of confusion and (2) study the nature of the confusion.

3 STUDY 1: REVERSING RESPONSE OPTIONS LOWERS SATISFACTION

RQ1: what is the impact of reversing the response option of bipolar, vertically-oriented satisfaction scales?

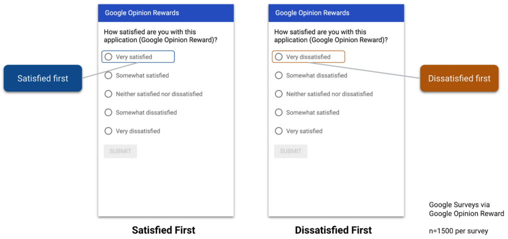
In this first study, we explored to what extent the order of the response options in vertically-oriented, 5-point, bipolar satisfaction questions has an effect on the results. We fielded a survey with a single question that asked for the satisfaction with the Google Opinion Reward app: "*How satisfied are you with this application (Google Opinion Reward)?*" with the response options *very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, very dissatisfied.* We asked about satisfaction with the Opinion Reward app since we used Google Surveys (now discontinued) to run the study, and since its questionnaires are delivered

	***very satisfied* first**	***very disatisfied* first**			
Response option	**Prevalence** $[CI_{95\%}]$	**Prevalence** $[CI_{95\%}]$	**Delta**	**Sig**	**Odd-Ratio**
Very dissatisfied	0.5% [0.2%- 1.0%]	5.0% [4.0%- 6.2%]	10.0x	p < .001	11.2
Somewhat dissatisfied	2.2% [1.6%- 3.1%]	5.0% [4.0%- 6.2%]	2.3x	p < .001	2.3
Neither sat .. nor .. dissat ..	7.4% [6.2%- 8.8%]	8.0% [6.7%- 9.5%]	-	p = .584	N.S.
Somewhat satisfied	29.0% [26.8%-31.3%]	30.5% [28.2%-32.8%]	-	p = .402	N.S.
Very satisfied	60.9% [58.4%-63.4%]	51.5% [49.0%-54.1%]	0.8x	p < .001	1.5

Table 1: Study 1: Descriptive and inferential statistics on the impact of the scale order on the prevalence of the response options.

via the Opinion Reward app. This ensured implicitly that each of the participants was a user of the application.

We used the tool's setting to recruit a sample that represents US Android phone users in terms of age, gender, and geographic location. People who are selected as participants receive a notification to participate which is valid for 24 hours. The system samples participants from the appropriate demographics until the desired number of responses is reached.



How satisfied are you with this application (Google Opinion Reward)?

How satisfied are you with this application (Google Opinion Reward)?

Figure 2: Study 1: The questionnaires used in the control (left) and experimental condition (right).

The study was set up as a between-group experiment with two conditions (see Figure 2): the response options being in the default order (*very satisfied* first, control condition), and in the reversed order (*very dissatisfied* first, treatment condition). For each survey, we collected 1,500 responses. The age group distribution was 18-24 = 14.0%, 25-34 = 21.1%, 35-44 = 19.4%, 55-54 = 15.7%, 55-64 = 14.4%, and 65+ = 15.4%. The gender distribution was 47.7% female, 55.7% male, and 0.1% other. Participants came from all 50 US states.

Figure 3 shows that the distribution of the results was notably affected by the experimental treatment. When the questions were reversed, we observed a shift from positive to negative responses. We found statistically significant impacts of the treatment on how often *very satisfied*, *somewhat dissatisfied*, and *very dissatisfied* were selected (Fisher's Exact tests, all p < .001). In particular, the *very dissatisfied* response option was selected by 5% instead of 0.5% of the respondents. The Odds-Ratio of 11.2 indicates that the very dissatisfied option is 11.2 times more likely to be selected when the response options are reversed. Table 1 shows the descriptive (mean and confidence intervals) and inferential statistics (Fisher's exact test and Odds ratio) for each of the response options in detail.

The control-condition survey resulted in 90% top-2 box satisfaction. That is, at least 90% of the respondents are at least *somewhat satisfied.* For the experimental condition, we would report only 82% top-2 box satisfaction. The difference is statistically significant (Fisher's Exact Test: p < .001). Importantly, this difference is also substantively meaningful in practical settings when reporting to stakeholders.

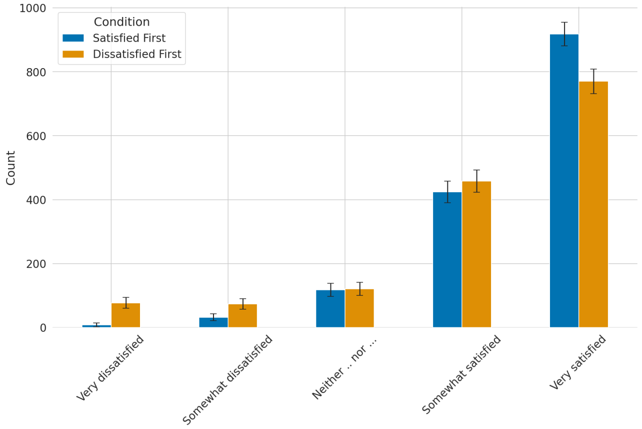


Figure 3: Study 1: Number of responses per response category. Note that for both surveys we consider exactly 1500 responses.

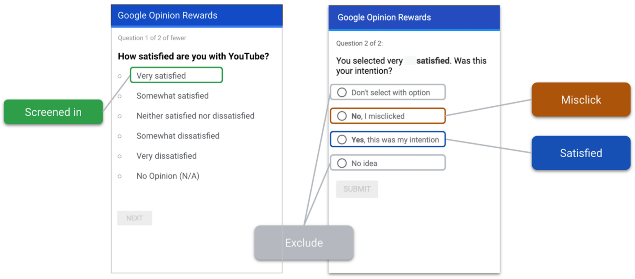
The median response time for each of the conditions give a hint at the cause: in the baseline condition, the median response time was 7.5 seconds $[CI_{95\%}$: 7.2- 7.8]. When the response options were reversed, the median response time increased to 8.1 seconds $[CI_{95\%}$: 7.8- 8.3]. The difference is statistically significant (t(2965.1)=2.40, p=0.017*). Slower response time can indicate an increase in cognitive load when the response options are reversed.

4 STUDY 2: VERIFYING THE RESPONSES

RQ2: is there evidence that the shift results from unintended responses?

To quantify to what extent the shift in satisfaction responses from Study 1 are caused by respondents picking the wrong response option, we conducted a second survey-based study. The first question asked about the satisfaction with a popular entertainment product. If a respondent selected the top response option, they were taken to a second question that asked "*You selected [previous response option inserted here]. Was that your intention?*". Respondents could select "*yes, this was my intention*" and "*no, this was an accident*". To identify and remove spurious responses from the data, we added one distractor item to the top and at the bottom each. This second question only appeared if the respondent had selected the top-most response in the previous question. The survey tool did not allow to go back to the previous question.

The study was set up as a between-group experiment. In the control condition, the satisfaction question was shown in positive order, that is, the very satisfied option was on top. In the experimental condition, the order of the response options was reversed. Figures 4 and 5 show the questionnaires used in this study.


Google Opinion Rewards

Question 1 of 2 of fewer

How satisfied are you with YouTube?

- ○ Very satisfied
- ○ Somewhat satisfied
- ○ Neither satisfied nor dissatisfied
- ○ Somewhat dissatisfied
- ○ Very dissatisfied
- ○ No Opinion (N/A)

NEXT

Google Opinion Rewards

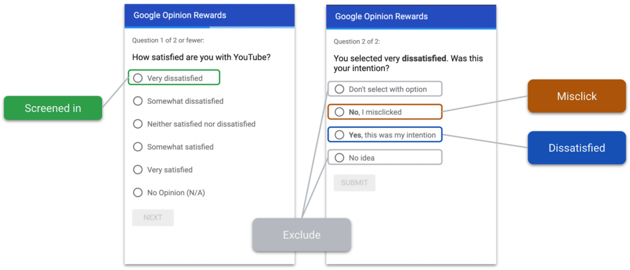
Question 2 of 2:

You selected very **satisfied**. Was this your intention?

- ○ Don't select with option
- ○ **No**, I misclicked
- ○ **Yes**, this was my intention
- ○ No idea

SUBMIT

Figure 4: Study 2: Survey questions used for the control group: *very satisfied* on top.


Google Opinion Rewards

Question 1 of 2 or fewer:

How satisfied are you with YouTube?

- ○ Very dissatisfied
- ○ Somewhat dissatisfied
- ○ Neither satisfied nor dissatisfied
- ○ Somewhat satisfied
- ○ Very satisfied
- ○ No Opinion (N/A)

NEXT

Google Opinion Rewards

Question 2 of 2:

You selected very **dissatisfied**. Was this your intention?

- ○ Don't select with option
- ○ **No**, I misclicked
- ○ **Yes**, this was my intention
- ○ No idea

SUBMIT

Figure 5: Study 2: Survey questions used for the experimental group: *very dissatisfied* on top.

We collected 100 responses per condition. From those, we removed the respondents who had selected one of the distractor items in the second question. This left us with 97 valid responses for the control condition, and 93 valid responses for the experimental condition.

45 of the respondents in the control condition were female, 51 were male, one did not disclose their gender. The number of responses per age brackets were: 18-24 = 22, 25-34 = 21, 35-44 = 21, 45-54 = 17, 55-64 = 7, 65+ = 9. 51 of the respondents in the experimental condition were female, 42 were male. The number of responses per age brackets were: 18-24 = 15, 25-34 = 16, 35-44 = 11, 45-54 = 18, 55-64 = 13, 65+ = 18.

In the control condition, 96 of 97 (99.0% – $CI_{95\%}$: 93.8% – 99.9%) respondents with valid answers confirmed that selecting the top-most response option (*very satisfied*) had been their intention. In the experimental condition, only 46 of 93 (49.5% – $CI_{95\%}$: 39.5% - 59.4%]) respondents with valid answers confirmed that selecting the top-most response option (*very dissatisfied*) had been their intention. A Fisher's Exact test showed that the difference is statistically significant (p<0.001) with a large effect size (Odds Ratio=98.09).

This confirms that when the *very dissatisfied* response option is on top, we expect in about 50% of the cases when it is selected, it is selected unintentionally.

5 STUDY 3: UNDERSTANDING THE REASONS BEHIND THE DIFFERENCE

RQ3: why do these unintended responses occur when the options are reversed?

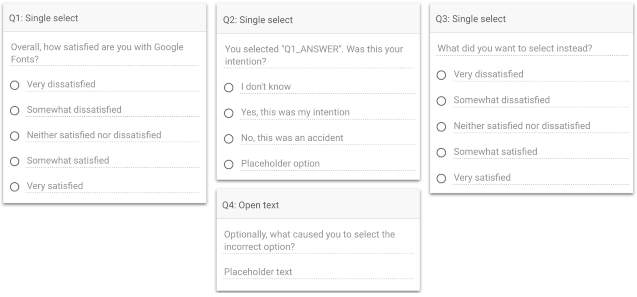
To better understand the cause of respondents selecting unintended responses, we conducted a third study on a public website http://fonts.google.com with the friendly permission of the team behind that website. The website is largely visited by designers, both professionals and enthusiasts, from around the world. To field the survey, we used *HaTS* [24], a tool that fields intercept surveys to website visitors. It randomly selects participants from website visitors who have spent at least 20 seconds on the website. The conversion rate was about 2.5% which is a typical rate for such surveys. The tool is run in production products. To comply with privacy regulations, the tool does not collect respondent demographics. Thus, we cannot report age or gender distributions for this study. The survey consisted of 4 questions [see Fig 6]:

Q1: The first question asked the standard satisfaction question. The response options were kept in the reversed order: *very dissatisfied* was always the top response option.

Q2: The second question repeated the selection made in the first question and asked whether that had been the intended response (e.g. "*You selected very dissatisfied. Was this your intention?*"). To be able to filter out respondents who participated in the survey without paying attention, the top and the bottom response options were distractors which did not provide reasonable responses. The two middle options allowed respondents to indicate whether their first response had been intended or not. If participants recorded their response as intended, the survey closed at this point.

Q3: If respondents indicated that they had selected the wrong response option, they were taken to a third question that allowed them to indicate which response option they had wanted to select instead.

Q4: Optionally, participants could explain what caused them to select the incorrect response option.


Q1: Single select

Overall, how satisfied are you with Google Fonts?

- ○ Very dissatisfied
- ○ Somewhat dissatisfied
- ○ Neither satisfied nor dissatisfied
- ○ Somewhat satisfied
- ○ Very satisfied

Q2: Single select

You selected "Q1_ANSWER". Was this your intention?

- ○ I don't know
- ○ Yes, this was my intention
- ○ No, this was an accident
- ○ Placeholder option

Q4: Open text

Optionally, what caused you to select the incorrect option?

Placeholder text

Q3: Single select

What did you want to select instead?

- ○ Very dissatisfied
- ○ Somewhat dissatisfied
- ○ Neither satisfied nor dissatisfied
- ○ Somewhat satisfied
- ○ Very satisfied

Figure 6: Study 3: Questions used in the survey.

The survey collected 8,951 responses. 4,595 (51.3%) respondents indicated whether or not their responses were intended. Of those

4,595 responses, 470 (10.2%) were flagged as unintended. The chart in Figure 7 shows what fraction of responses were given by mistake for each of the response options. It shows that most of the mistakes happened for the *very dissatisfied* response option. Selecting *very dissatisfied* accounts for 370 of those 470 unintended responses.

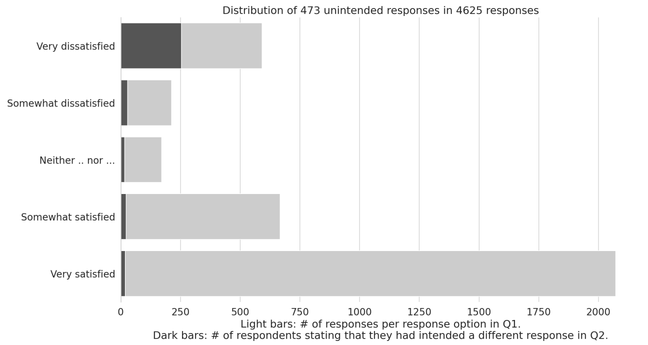


Figure 7: Study 3: Distribution of the responses where respondents confirmed their response as intended.

Figure 8 shows the confusion matrix for all cases where participants corrected their response to the first question. The x-axis contains the intended response, the y-axis the recorded response. The by far most common case was that respondents had intended to report themselves as being *very satisfied* instead of *very dissatisfied.*

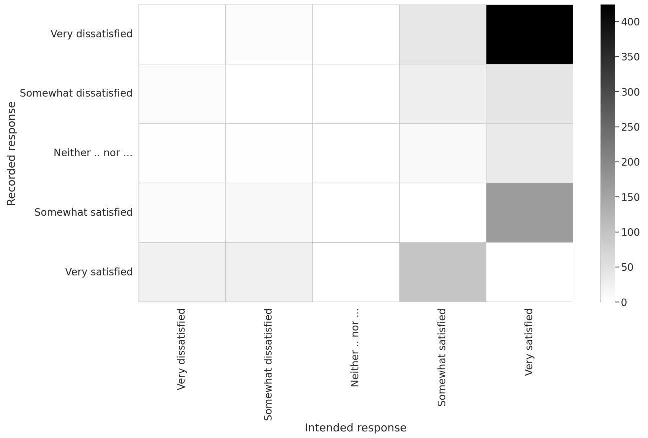


Figure 8: Study 3: Confusion matrix mapping the initially recorded response (y-axis) against the actually-intended response (x-axis).

186 (39.3%) of the 473 respondents who corrected their response left an answer to the question "*what caused you to select the incorrect option?*" 76 (40.9%) of the 186 comments mention that the respondent expected the positive option to be on top. A few examples are:

- "*surprised to to see very satisfied right at bottom*",
- "*it was the first option that appeared on top. To me, logically top (best) bottom (worst)*",
- "*The order of the selections was the opposite of what I'm used to. Positive options are usually at the top and negative options are usually at the bottom*",
- "*read the options to quickly, and normally I see the positive on top and not the negative*"
- "*Up is good, down is bad*",
- "*Positive should be at the top and negative at the bottom... or maybe you're messing with me and you're only pretending that the negative was at the top and this is for some behavioral study*".

55 (29.6%) of the 186 comments mention that the respondents did not pay enough attention to the question, e.g.:

- "*Doing two things at the same time and not reading good enough. In my head good is at the top and worst at the bottom*",
- "*I clicked too fast without reading, I imagined the most positive answer was at the top of the list.*"
- "*Didn't read the options so I choose by instinct*",
- "*I clicked without reading the entire sentence*",
- "*read the last part only*".

A few comments mentioned the similarity between the labels of the endpoints:

- "*Without a visual cue, "dissatisfied" looks very similar to "satisfied""*,
- "*[...], the first word for both options was "Very", and both ended in "Satisfied", the only difference being "Dis." A more varied vocabulary would have made it less ambiguous*".

Figure 9 shows the average *misclick*-rate by response time. At the 3 second mark – clustering all responses that were made within 2.5 to 3.5 seconds, the fraction of unintentional selections of *somewhat-* and *very dissatisfied* is notably higher than for longer response times. For responses times below 3.5 seconds, 58 of 98 (59.2%, $CI_{95\%}$=[49.2%-68.5%]) of the responses were marked as unintentional. In comparison, for responses times of 3.5 seconds and above, for responses times below 3.5 seconds, 341 of 977 (34.9%, $CI_{95\%}$=[32.0%-37.9%]) of the responses were marked as unintentional. This mean difference of 24.1% points is statistically significant (Fisher's Exact test: p=0.000***) with a medium effect size (Odds Ratio=2.70). This quantitative evidence corroborates the explanation that not paying attention caused respondents to select the wrong response option.

In summary, the findings of this study show: people select the wrong response options because they are not paying close attention to the wording, expect the positive response option to be on top, and do not notice that the scale is reversed due to the similar spelling of satisfied and dissatisfied.

6 DISCUSSION

Reversing Response Options Significantly Lower Satisfaction

Study 1 showed that for a vertically-presented bipolar satisfaction question (*very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, very dissatisfied*) the *very dissatisfied* response option is selected 10 times more frequently when the scale is reverted, that is, *very dissatisfied* is the top response option, than when *very satisfied* is the top response option. Study 1 successfully

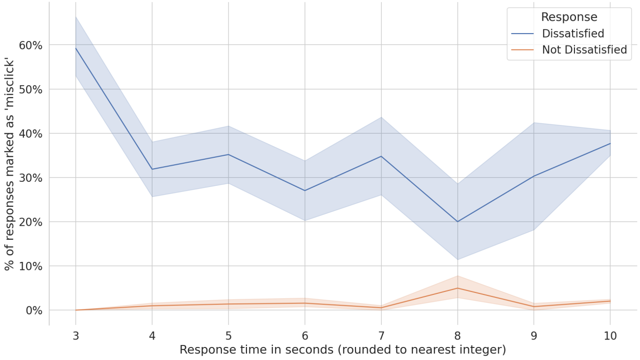


Figure 9: Study 3: Fraction of responses with selection error by response time.

replicated previous findings [11, 26, 29] for short surveys delivered via mobile phones.

The top-2-box satisfaction drop of 8% points can mean the difference between celebrating/promoting a product or not. When a researcher enables or disables the random flip option, we could expect a shift of about 4% points in either direction (the exact shift will depend on the sample and product). For products with large sample sizes, this can trigger different investments and business decisions.

Reason: Respondents Provide Unintended Responses

Study 2 showed that about half of the times that *very dissatisfied* is selected from a reversed scale, respondents indicated that this selection was unintended. This effect does not occur when the scale is in positive order. Study 3 confirmed this finding at scale: 42.9% of the time that *very dissatisfied* was selected, respondents later corrected their selection. Study 3 further showed that reversing the response options invalidates about 10% of the responses. These two studies demonstrate that the effect of order of the responses can be explained by participants selecting options that they did not intend to select.

One alternative interpretation might be that we are seeing a social desirability effect (e.g. Grimm [13]) at play: participants who selected *very dissatisfied* felt compelled to change their rating due to social pressure. The qualitative feedback, however, does not confirm this. We would also expect a much larger fraction of respondents who selected *somewhat dissatisfied* to change their response as well.

Explained by Lack of Attention and Expectation of “Good Means Up”

The qualitative feedback collected in Study 3 provides novel insight into the reason for the unintended responses. The two main cases were participants not reading the response options properly and expecting the positive option to be on top. Because the two constructs *satisfied* and *dissatisfied* spell very similarly, it was easy for the respondents to overlook that the response options were reversed. This is finding is quantitatively confirmed by the fact that the most frequent occurrence is correcting an unintentionally-selected *very dissatisfied* to *very satisfied.*

The quantitative evidence confirms that most unintended responses happen when people respond very fast - in our case about 3.5 seconds and faster. This is evidence for the occurrence of unintentional responses due to low motivation and having limited cognitive resources available due to the main task or distractions in the surroundings of the respondent. This is in line with research where speed in answering is correlated with satisficing behaviors (Zhang & Conrad [32]; Hanby & Taylor [14]).

These results corroborate the *Up means good* (Tourangeau, Couper and Conrad [28]) heuristic. In surveys with vertically-ordered response options, a significant share of participants will expect the positive response option to be on top. We also found that it is important to meet this expectation, as respondents may not bother to verify whether the scale of the survey they are about to answer meets their expectations.

Limitations

The findings from the literature supporting *Up means good* heuristic, including our own, have been tested in Western cultures where reading is left to right. With very few exceptions we are aware of (e.g. Yang et al. [31]) more research is needed in other cultures. One reason we are seeing such clear effects seems to be the very similar spelling of *satisfied* and *dissatisfied.* We do not think that the evidence from this study can be fully generalized to scales where the endpoints are visually distinct. Finally, this research studied single-question satisfaction questions, as they are used in intercept surveys in products. They findings may not generalize to longer questionnaires where the participants might pay more attention.

7 CONCLUSIONS

We present evidence that some of the current practice of randomly reversing the response options of scales leads to a significant fraction of invalid results for bipolar, vertically-oriented satisfaction questions (*very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, very dissatisfied*). We see a much lower satisfaction because participants who are actually very satisfied pick the *very dissatisfied* response option by accident. In other words, randomly reversing or flipping the scale can confuse part of half the sample and introduce a bias with serious real-world implications.

The impact of this practice is that the stakeholders get reported much lower satisfaction values than what is actually true - up to 8 percent points like in our study. This level of bias can make the difference between celebrating the success of a product and investing a significant amount of resources into addressing potentially trivial satisfaction issues. On the basis of this evidence, we recommend fielding such satisfaction questions following the *Up is Good* heuristic: the top-most response option should be the most positive one.

Future work needs to investigate whether this effect is particular to the construct satisfaction, and/or pairs of similarly looking words, as the labels *very satisfied* and *very dissatisfied* are spelled alike. The same effect might not occur for visually distinct endpoint labels, such as *extremely satisfied* and *not at all satisfied.* Future work furthermore needs to investigate whether the effect also occurs in longer surveys, where respondents have more time to adjust to the response options.

ACKNOWLEDGMENTS

We are grateful to the Google Fonts team for kindly allowing us to run parts of this research on https://fonts.google.com.

REFERENCES

[1] S. Anand. 2008. *Satisficing*. Sage, Los Angeles, 707–799.

[2] K. Baxter, C. Courage, and K. Caine. 2015. *Understanding Your Users: A Practical Guide to User Research Methods*. Elsevier Science. https://books.google.de/books?id=wHjYrQEACAAJ

[3] John Brooke. 1996. *"SUS-A quick and dirty usability scale." Usability evaluation in industry*. CRC Press. https://www.crcpress.com/product/isbn/9780748404605 ISBN: 9780748404605.

[4] Mario Callegaro. 2013. *Paradata in Web Surveys*. John Wiley & Sons, Ltd, Chapter 11, 259–279. https://doi.org/10.1002/9781118596869.ch11 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118596869.ch11

[5] Jaime G. Carbonell. 1993. *Derivational Analogy: A Theory of Reconstructive Problem Solving and Expertise Acquisition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 727–738.

[6] Leah Melani Christian, Nicholas L. Parsons, and Don A. Dillman. 2009. Designing Scalar Questions for Web Surveys. *Sociological Methods & Research* 37, 3 (2009), 393–425. https://doi.org/10.1177/0049124108330004

[7] Evert de Haan, Peter C. Verhoef, and Thorsten Wiesel. 2015. The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing* 32, 2 (June 2015), 195–206. https://doi.org/10.1016/j.ijresmar.2015.02.004

[8] Anna DeCastellarnau. 2018. Effects of Stem and Response Order on Response Patterns in Satisfaction Ratings. *Quality & Quantity* 52, 3 (2018), 1523––1559. https://doi.org/10.1007/s11135-017-0533-4

[9] Kraig Finstad. 2010. The Usability Metric for User Experience. *Interacting with Computers* 22, 5 (2010), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004 Modelling user experience - An agenda for research and practice.

[10] Dana Garbarski. 2016. Research in and Prospects for the Measurement of Health Using Self-Rated Health. *Public Opinion Quarterly* 80, 4 (09 2016), 977–997. https://doi.org/10.1093/poq/nfw033 arXiv:https://academic.oup.com/poq/article-pdf/80/4/977/8639142/nfw033.pdf

[11] Dana Garbarski, Nora Schaeffer, and Jennifer Dykema. 2018. The Effects of Features of Survey Measurement on Self-Rated Health: Response Option Order and Scale Orientation. *Applied Research in Quality of Life* 14 (04 2018). https://doi.org/10.1007/s11482-018-9628-x

[12] Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed. 2012. Chapter 12 - Surveys. In *Observing the User Experience (Second Edition)* (second edition ed.), Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed (Eds.). Morgan Kaufmann, Boston, 327–383. https://doi.org/10.1016/B978-0-12-384869-7.00012-7

[13] Pamela Grimm. 2010. *Social Desirability Bias*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444316568.wiem02057 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444316568.wiem02057

[14] Tyler Hamby and Wyn Taylor. 2016. Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples. *Educational and Psychological Measurement* 76, 6 (2016), 912–932. https://doi.org/10.1177/0013164415627349

[15] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003: Interaktion in Bewegung*, G. Szwillus and J. Ziegler (Eds.). B. G. Teubner, Stuttgart, 187–196.

[16] Caroline Jarrett. 2021. *Surveys that Work: A Practical Guide for Designing and Running Better Surveys*. Rosenfeld Media.

[17] Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 3 (1991), 213–236. https://doi.org/10.1002/acp.2350050305 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305

[18] Jon A. Krosnick. 1999. Survey Research. *Annual Review of Psychology* 50, 1 (1999), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

[19] Jon A. Krosnick. 2018. *Improving Question Design to Maximize Reliability and Validity*. Springer International Publishing, Cham, 95–101. https://doi.org/10.1007/978-3-319-54395-6_13

[20] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*, Andreas Holzinger (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–76.

[21] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: When There’s No Time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI ’13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. https://doi.org/10.1145/2470654.2481287

[22] Neil A. Morgan and Lopo Leotte Rego. 2006. The Value of Different Customer Satisfaction and Loyalty Metrics in Predicting Business Performance. *Marketing Science* 25, 5 (2006), 426–439. https://doi.org/10.1287/mksc.1050.0180

[23] Morten Moshagen and Meinald T. Thielsch. 2010. Facets of visual aesthetics. *International Journal of Human-Computer Studies* 68, 10 (2010), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

[24] Hendrik Müller and Aaron Sedley. 2014. HaTS: Large-Scale in-Product Measurement of User Attitudes & Experiences with Happiness Tracking Surveys. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (Sydney, New South Wales, Australia) *(OzCHI ’14)*. Association for Computing Machinery, New York, NY, USA, 308–315. https://doi.org/10.1145/2686612.2686656

[25] Jeff Sauro and Joseph S. Dumas. 2009. Comparison of Three One-Question, Post-Task Usability Questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI ’09)*. Association for Computing Machinery, New York, NY, USA, 1599–1608. https://doi.org/10.1145/1518701.1518946

[26] Jolene D. Smyth, Glenn D. Israel, III Milton G. Newberry, and Richard G. Hull. 2019. Effects of Stem and Response Order on Response Patterns in Satisfaction Ratings. *Field Methods* 31, 3 (2019), 260–276. https://doi.org/10.1177/1525822X19860648

[27] Roger Tourangeau. 2018. The survey response process from a cognitive viewpoint. *Quality Assurance in Education* 26, 2 (2018), 169–181. https://doi.org/10.1108/QAE-06-2017-0034

[28] Roger Tourangeau, Mick P. Couper, and Frederick Conrad. 2004. Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly* 68, 3 (09 2004), 368–393. https://doi.org/10.1093/poq/nfh035 arXiv:https://academic.oup.com/poq/article-pdf/68/3/368/5167941/nfh035.pdf

[29] Roger Tourangeau, Mick P. Couper, and Frederick G. Conrad. 2013. “Up Means Good”: The Effect of Screen Position on Evaluative Ratings in Web Surveys. *Public Opinion Quarterly* 77, S1 (01 2013), 69–88. https://doi.org/10.1093/poq/nfs063 arXiv:https://academic.oup.com/poq/article-pdf/77/S1/69/17163883/nfs063.pdf

[30] Ting Yan. 2023. Which Scale Direction is More Difficult for Respondents to Use? An Eye-tracking Study. *Survey Practice* 16, 1 (21 9 2023). https://doi.org/10.29115/SP-2023-0015

[31] Yongwei Yang, Rich Timpone, Mario Callegaro, Marni Hirschorn, Vlad Achimescu, and Maribeth Natchez. 2019. Response Option Order Effects in Cross-Cultural Context. An experimental investigation. In *Conference of the European Association for Survey Research (ESRA)*. Zagreb.

[32] Chan Zhang and Frederick Conrad. 2014. Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods* 8, 2 (Jul. 2014), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453