

Google



CPUX



AAPOR ✓

Detecting Response Scale Inconsistency in Real Time

Using realtime paradata, dependent interviewing, and natural language processing (NLP) in web surveys

Mario Callegaro - Google

Carol Haney - Qualtrics



callegaro@google.com



carolh@qualtrics.com

Which response scale orientation most accurately reflects the respondents' true attitude?

Seemingly conflicting respondent answers

Within the same survey some respondents give conflicting answers on the same topic

Open ended answer in opposite sentiment

Some respondents will give glowingly positive open-ended evaluations of a subject immediately after having provided a low rating for the same subject

Is the culprit interpreting the scale incorrectly?

How does the response scale orientation affect the ratings?

Recent research on scale orientation for Self-rated health (SRH) - Desktop respondents

Original SHR Question wording:

Would you say your health in general is excellent, very good, good, fair, or poor?

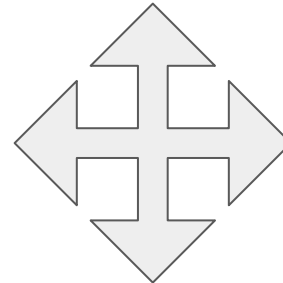
	Garbarski, Schaeffer & Dykema (2015)	Garbarski, Schaeffer & Dykema (2016)	Garbarski, Schaeffer & Dykema (2018)
Scale orientation	Vertical	Vertical	Vertical and Horizontal
Scale manipulation	Positive to Negative and Negative to Positive		
Online sample	U.S. KnowledgePanel	U.S. KnowledgePanel	U.S. Amazon Mturk
Results	Higher mean (healthier respondents) when scale ordered from Excellent to Poor		

2 by 4 design. Mobile vs. Desktop by scale orientation

U.S. Mobile (≈ 650 per condition) and Desktop (≈ 450 per condition) respondents from Dynata. Total number: 4,521 respondents

Random assignment to unipolar & bipolar block scales (counterbalanced) in one of four conditions (stay within the same condition for the whole study):

- Horizontal orientation, Negative on Left
- Horizontal orientation, Positive on Left
- Vertical orientation, Negative on Top
- Vertical orientation, Positive on Top



Topics: physical and mental health, financial situation & work satisfaction

Screenshots examples of scale questions

In general, would you say your physical health is ...

Excellent

Very good

Good

Fair

Poor



In general, how would you rate the quality of your social relationships?

Poor Fair Good Very good Excellent



9:05
Messages
ca1.qualtrics.com

In general, would you say your quality of life is:


Excellent


Very good

Good

Fair

Poor







8:38
Messages
ca1.qualtrics.com

In general, how would you rate your mental health, including your mood and your ability to think?

Poor Fair Good Very good Excellent



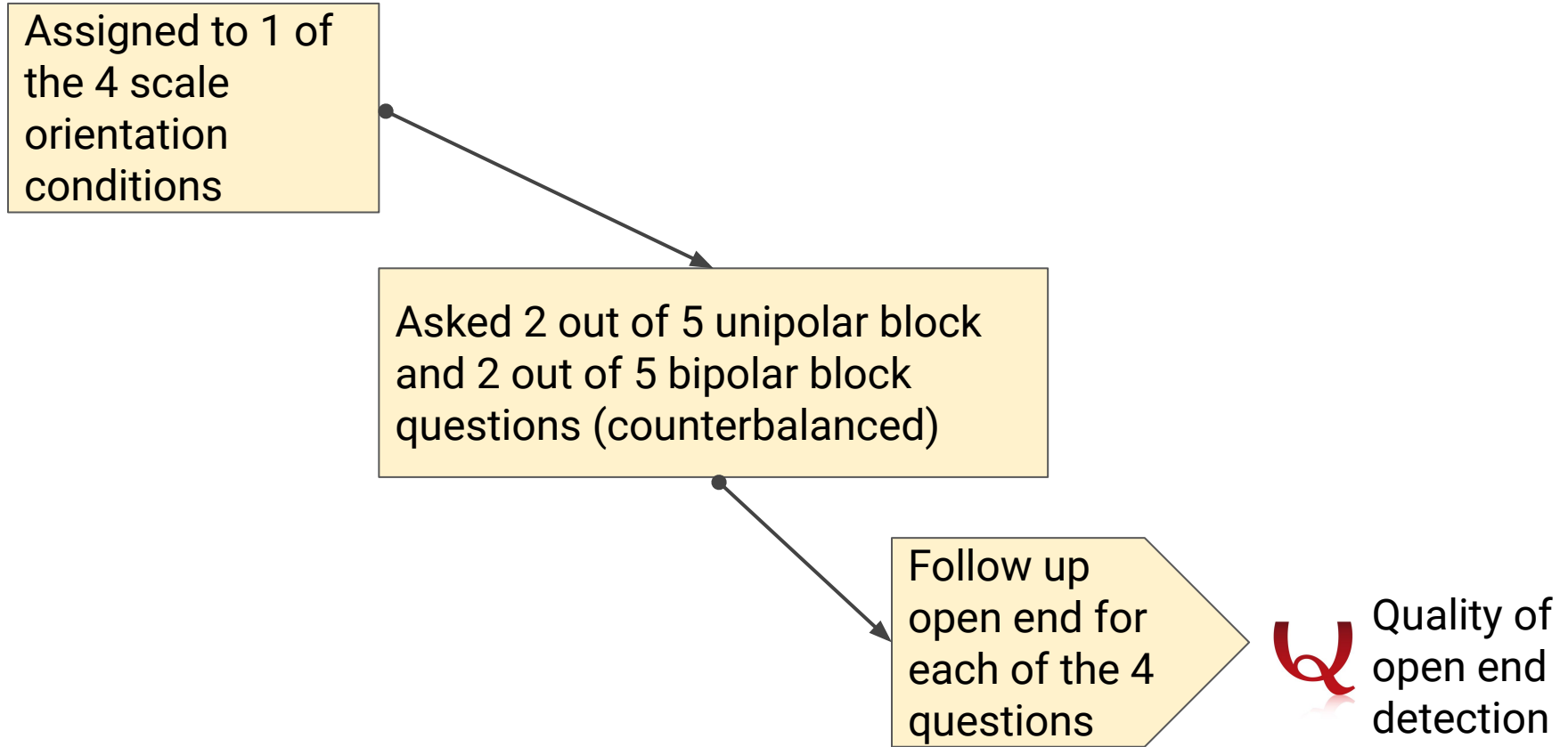


Real time data quality check and automatic sentiment scoring

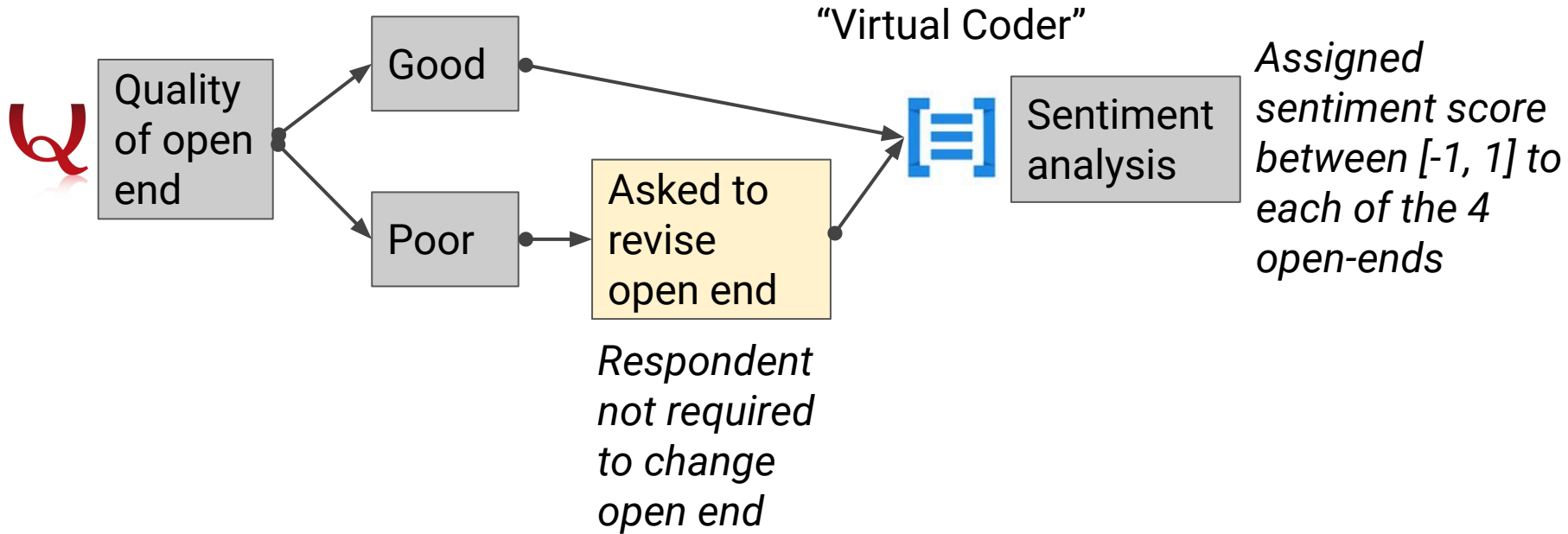
- **Real time paradata**
 - Check on quality of open ended answers using Qualtrics' Real-Time Gibberish Detection
- **Reactive dependent interviewing “Virtual coder”**
 - Google `AnalyzeSentiment` score of positive, negative or neutral used to detect inconsistency between the response scale and the open end answer
- **Proactive dependent interviewing “Respondent controlled”**
 - All respondents given a chance to change their original rating and explain why



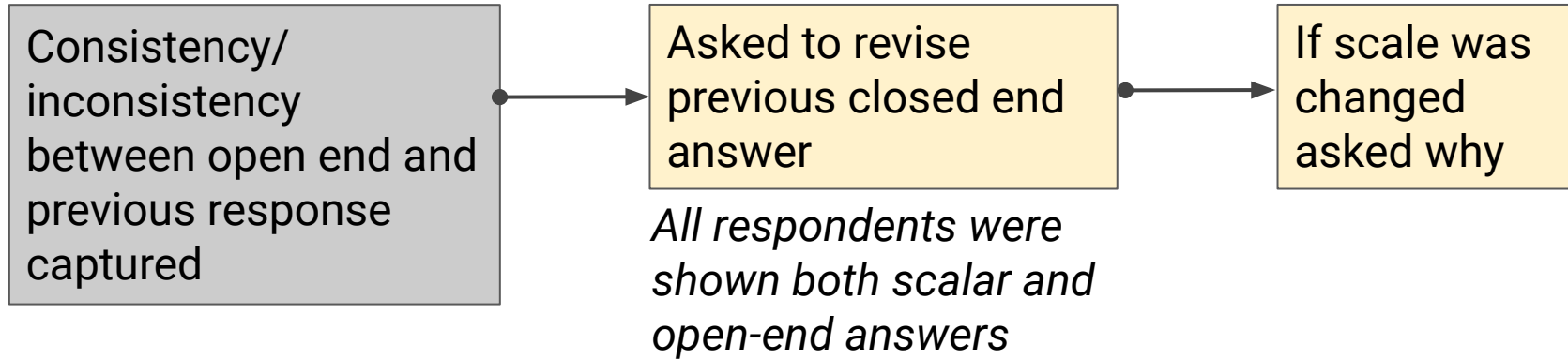
Questionnaire Flowchart I



Questionnaire Flowchart II



Questionnaire Flowchart III



Two dependent interviewing approaches:

1. **Respondent-Controlled**

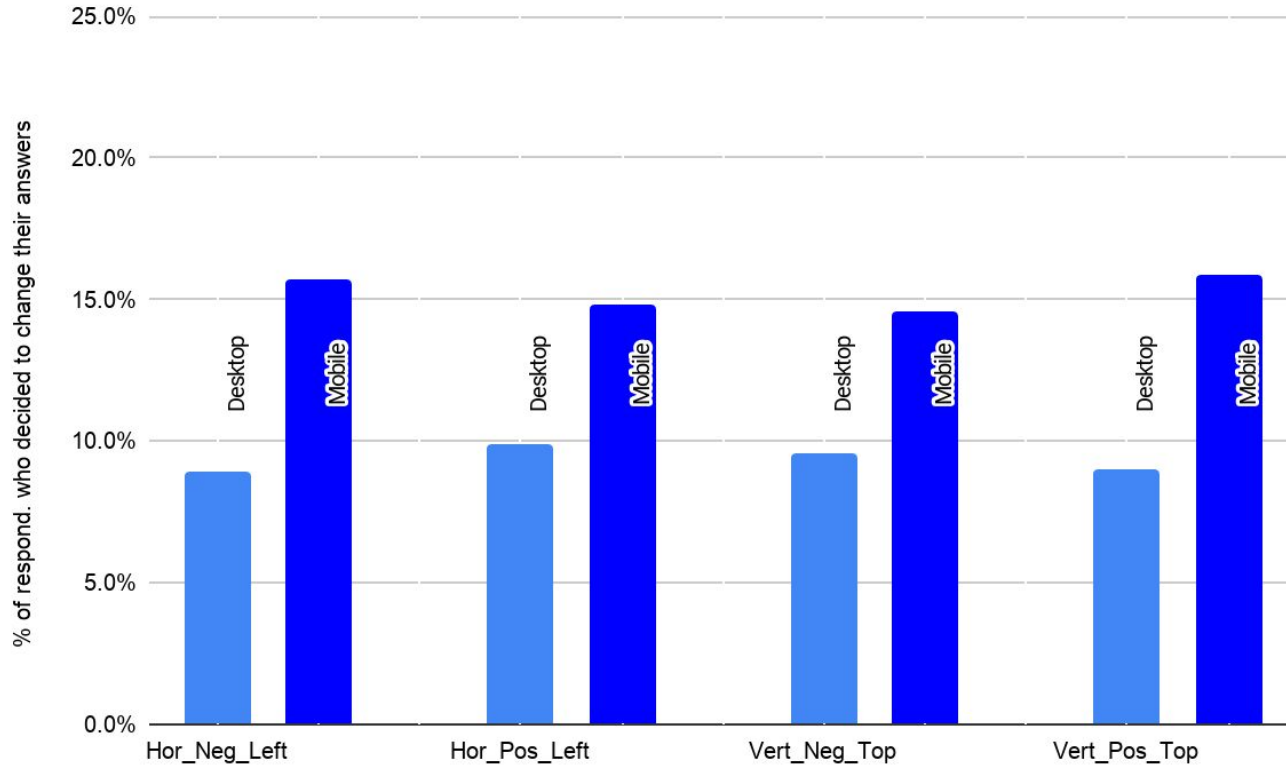
Ask respondents if they want to change the scale question, and ask about scale confusion

2. **Virtual Coder-Controlled**

Qualtrics integrated with its Real-Time Gibberish Detection and with Google NLP API “to assess quality of open ended and to flag inconsistencies

between scale question and open end

Respondents wanting to change their answers by device, at least once. N=4,521

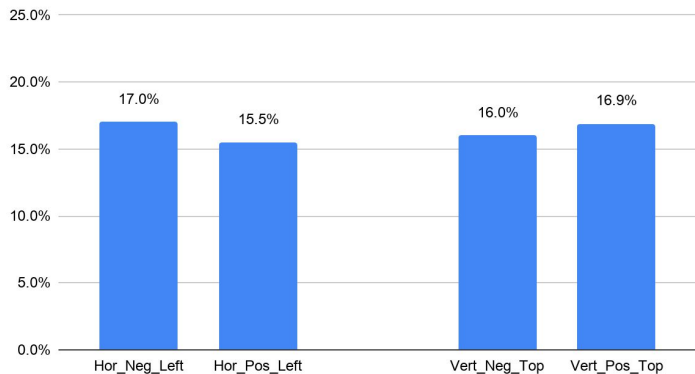


The mobile darker bar color means that the difference with the lighter desktop bar color next to it is statistically significant at $p < .05$

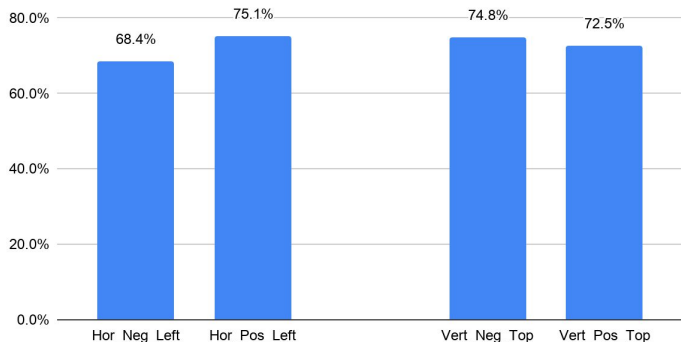
Results for unipolar scales across devices

Respondent-Controlled Base: all unipolar responses

I would like to change my answer to the rating question N= 1,447

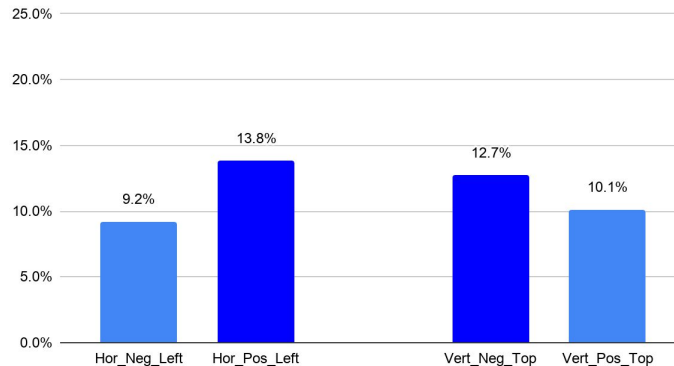


The first time I answered the rating question, I accidentally chose the wrong answer. N= 1,072 out of 1,447



“Virtual Coder”-Controlled Base: all unipolar responses

Inconsistency between scale and open end detected. N= 1,032

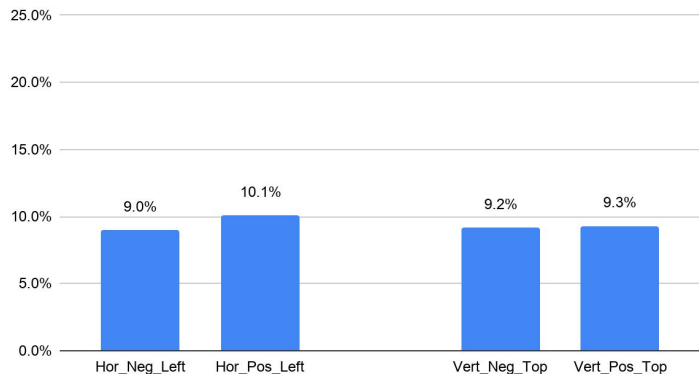


A darker bar color means that the difference with the lighter bar color is statistically significant at $p < .05$

Results for bipolar scales across devices

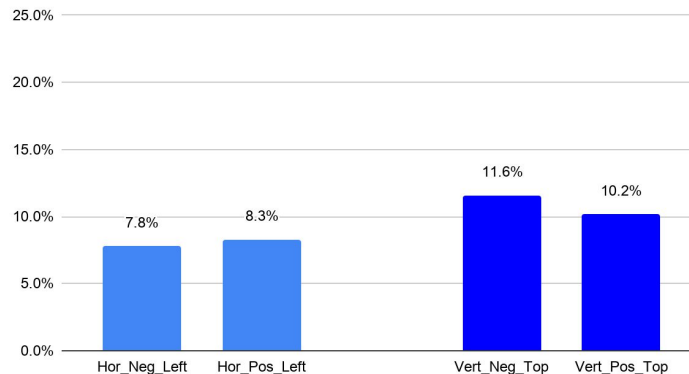
Respondent-Controlled Base: all bipolar responses

I would like to change my answer to the rating question. N= 848

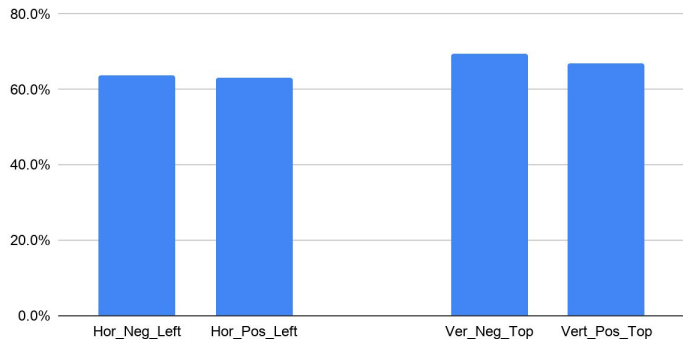


“Virtual Coder”-Controlled Base: all bipolar responses

Inconsistency between scale and open end detected N= 857



The first time I answered the rating question, I accidentally chose the wrong answer. N= 559 out of 848

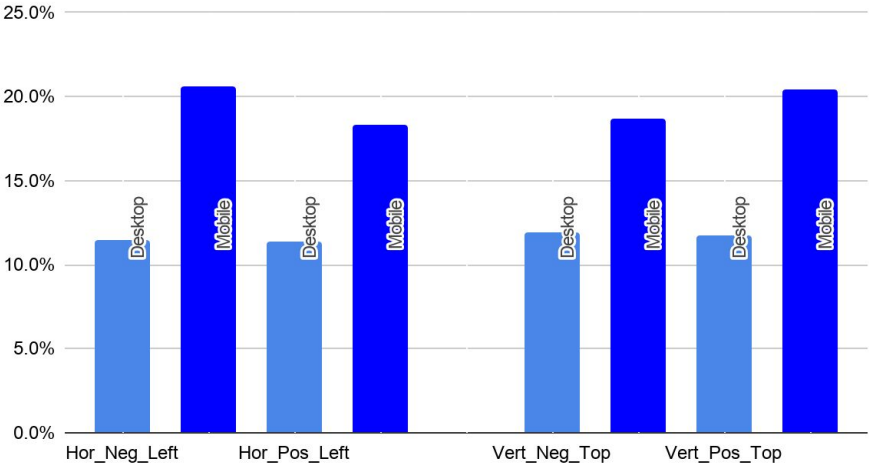


A darker bar color means that the difference with the lighter bar color is statistically significant at $p < .05$

Results for both scales, by device, respondent-controlled

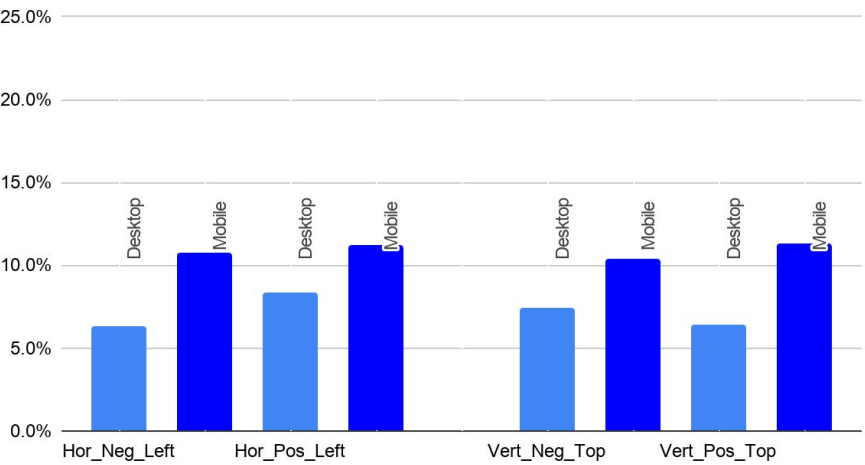
Unipolar

I would like to change my answer to the rating question. N= 1,477



Bipolar

I would like to change my answer to the rating question. N= 848

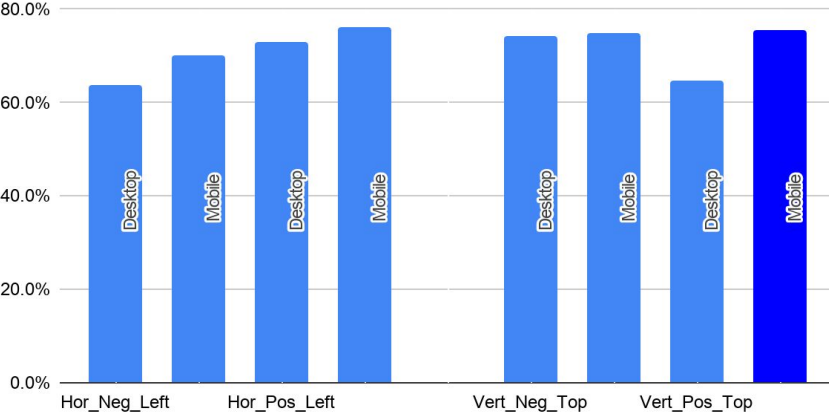


The mobile darker bar color means that the difference with the lighter desktop bar color next to it is statistically significant at $p < .05$

Results for unipolar scales, by device

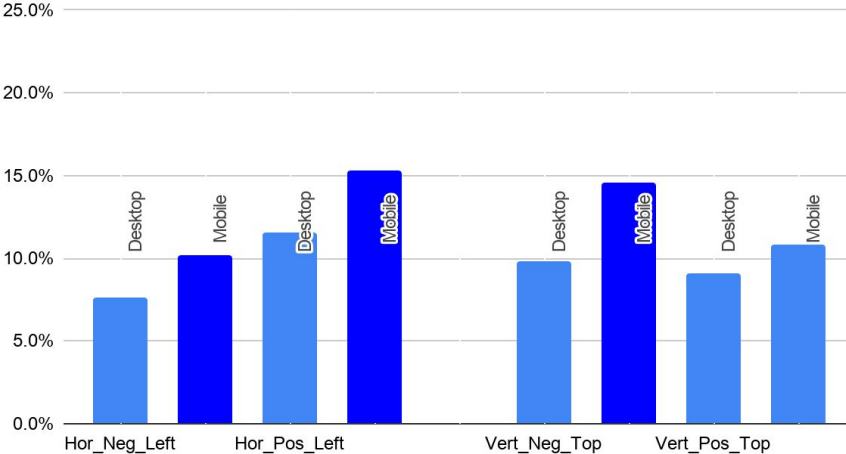
Respondent-Controlled

The first time I answered the rating question, I accidentally chose the wrong answer. N= 1,072



“Virtual Coder”-Controlled

Inconsistency between scale and open end detected. N= 1,032

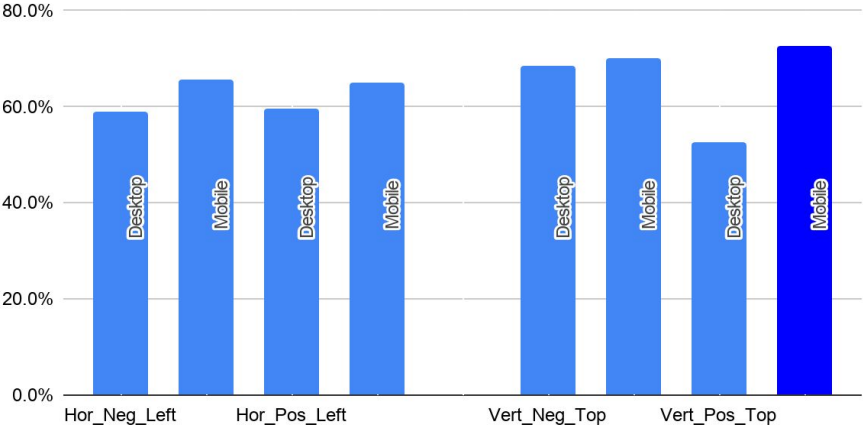


The mobile darker bar color means that the difference with the lighter desktop bar color next to it is statistically significant at $p < .05$

Results for bipolar scales, by device

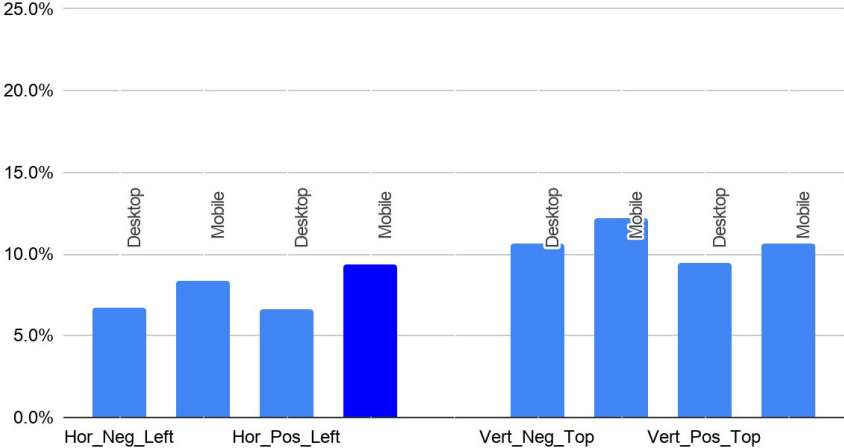
Respondent-Controlled

The first time I answered the rating question, I accidentally chose the wrong answer. N= 559



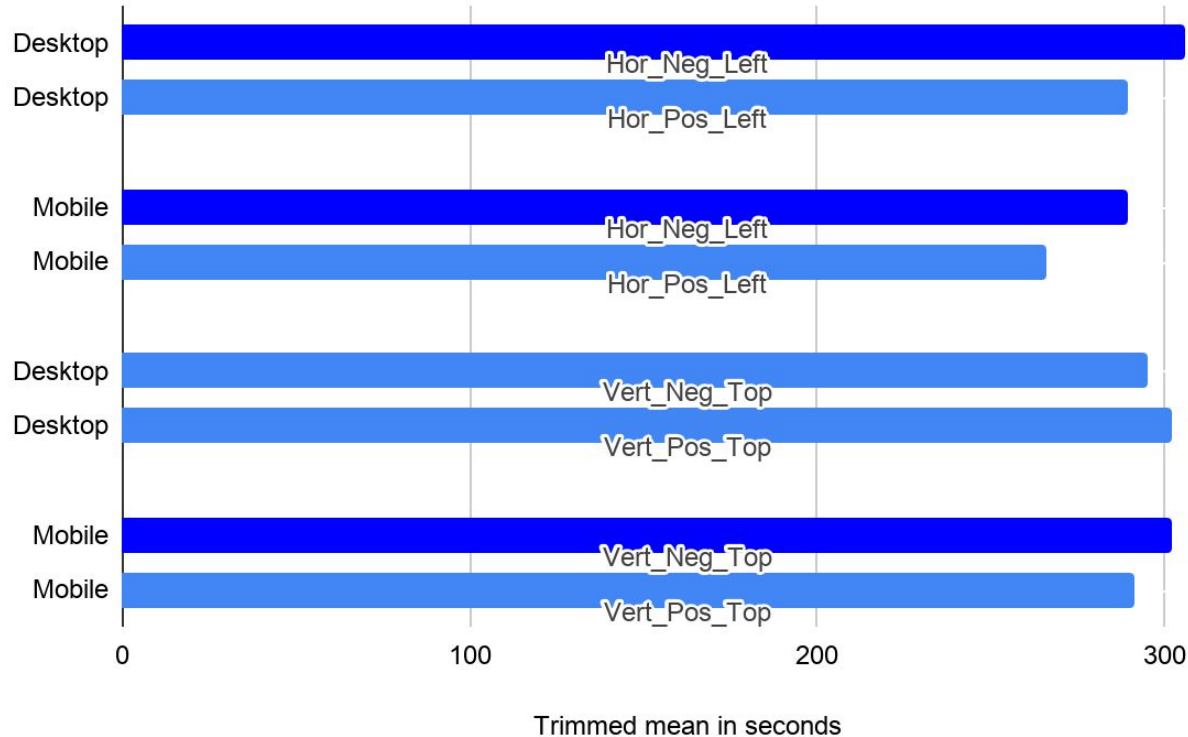
“Virtual Coder”-Controlled

Inconsistency between open end and scale detected. N= 857



The mobile darker bar color means that the difference with the lighter desktop bar color next it is statistically significant at $p < .05$

Answering a scale from the negative end almost always took longer, independent of device



The darker bar color means that the difference with the lighter bar color next to it is statistically significant at $p < .05$

What we learned:

Mobile respondents wanted to change response options more often than desktop respondents

Unipolar scales showed higher inconsistency overall than bipolar scales

Answering a scale from the negative end almost always takes longer

Unipolar question showed higher inconsistency for Hor_Pos_Left & Vert_Neg_Top

Higher inconsistency showed for mobile respondents

Bipolar scales showed higher inconsistency for vertically oriented scale

Questions?

Detecting Scale Inconsistency in Real Time

Using realtime paradata, dependent interviewing, and natural language processing (NLP)

Mario Callegaro - Google

Carol Haney - Qualtrics



callegaro@google.com



carolh@qualtrics.com

Appendix

Extra details, Demos by device, References, FAQs, Screenshots, and Examples of open end answers

Our experiment resulted in two approaches

1 Respondent-Controlled

We ask respondents to help guide consistency of response throughout the questionnaire, where respondents were asked:

1. If they would like to improve their open end question response, if it was detected to be a poor response
2. Prompting based on consistency/inconsistency between the scale response and open-end response, if they want to change their scale question
3. If they did opt to change their scale question, if the reason was based on scale confusion

2 “Virtual Coder”-Controlled

Qualtrics web survey integrated with Google NLP API “reviewed” inconsistencies between scale question and follow-up open end in real-time and assigned a category of inconsistency between scale response and follow-up open end response in the following method:

1. Numeric representation of the open end where *negative* responses were auto-coded if auto-coded sentiment < 0, *positive* responses if > 0, and *neutral* responses if = 0
2. Responses were tagged as inconsistent if:
 - a. Scale was positive (4,5) and auto-coded sentiment was < 0, or
 - b. Scale was negative (1,2) and auto-coded sentiment was > 0

Recent research on scale orientation

Desktop respondents

	Maloshonok & Terentev (2016)	Liu & Keusch (2017)	Terentev & Maloshonok (2019)
Scale orientation	Vertical & Horizontal	Vertical	Vertical
Items	3 bipolar fully labeled items	Agree - Disagree vs. Disagree - Agree	Unipolar scale Least to Most and Most to Least
Online sample	Russian MOOC students	U.S. KnowledgePanel	Russian MOOC students
Results	Mixes results: Only one item show stat. Sign differences in horizontal (primacy)	Higher acquiescence response style when scale presented with agree first	Primacy effects

Previous research on scale orientation

Desktop vs. Mobile respondents

	Mavletova (2013)	Lutig & Toepoel (2016)	Revilla & Couper (2018a, 2018b)
Scale orientation	Vertical	Vertical	Horizontal & Vertical
Items	6 items with fully labels response options	List of checkboxes	Set of different scales endpoint labeled
Online sample	Russian Opt-in Panel	Dutch LISS panel	Spanish Opt-in Panel
Results	Higher primacy effects in Desktop vs. Mobile	Higher primacy effects for Mobile vs. Desktop	More answer changes on Mobile Vs. Desktop Larger primacy effects on Mobile vs. Desktop

Age & Gender by device N= 4,521


Age	% Desktop	% Mobile
18-24	2.4	15.6
25-34	11.6	21.0
35-44	12.0	19.8
45-54	16.8	17.2
55-64	21.5	15.0
65+	35.7	11.3

Gender	% Desktop	% Mobile
Male	55.2	44.0
Female	44.8	56.0

Education by device N= 4,481

Education	% Desktop	% Mobile
High school or less	17.6	26.0
Associate degree and some college	30.8	35.0
Master and Bachelor	46.6	34.9
Doctoral and professional degree	5.0	4.1



Qualtrics - Google NLP API integration:

 **Web Service**



URL: [Test](#)



Method:


Query Parameters

=  



Body Parameters

=  

=  



=  

Custom Headers

=  

Fire and Forget

Set Embedded Data

=  

[Add Below](#) [Move](#) [Duplicate](#) [Delete](#)

FAQs

FAQs I

How similar were mobile to desktop respondents?

Mobile respondents tended to be younger, more female, and slightly less educated than desktop respondents

How good was Google NLP to code the open end answer and extract the sentiment?

We manually checked the quality of Google NLP sentiment classification and found it being really good, given the amount of text written in the open text

How long were the open ended answers, on average?

Length of answer: mean = 57; median = 42 characters;

of words per answer: mean = 11; median = 8

FAQs II

Did you use the Auto-Next feature in Qualtrics for the survey?

No, we did not use this feature. Respondents had to press the right arrow to go to the next page.

When did you field the study?

The study was fielded with U.S. respondents from using Dynata (former Research Now - Survey Sampling International) on April 12 - April 23, 2019

What was the proportion of mobile vs. desktop respondents?

Mobile respondents comprised 60% among all respondents

FAQs III

How many questions did each respondent answer in total?

25 required questions, using soft prompt only for gibberish

How long was the survey on average?

The average (trimmed mean) was of about 4.8 minutes

What percent of respondents were “cleaned out” because of inattentive, bots or other issues?

~15%, cleaned during field over time, using speeding and poor open-ends (nonsense and gibberish, using all four open-ends taken into account - one poor open-end was not cleaned)

FAQs IV

How were unipolar and the unipolar answers scales showed?

Same order: unipolar first, bipolar second, or they were randomized?

We showed them in randomized order

Questionnaire Screenshots

Desktop screenshots examples of scale questions

In general, would you say your physical health is ...

Excellent

Very good

Good

Fair

Poor



In general, how would you rate the quality of your social relationships?

Poor Fair Good Very good Excellent



In general, would you say your physical health is ...


Poor

Fair

Good

Very good

Excellent

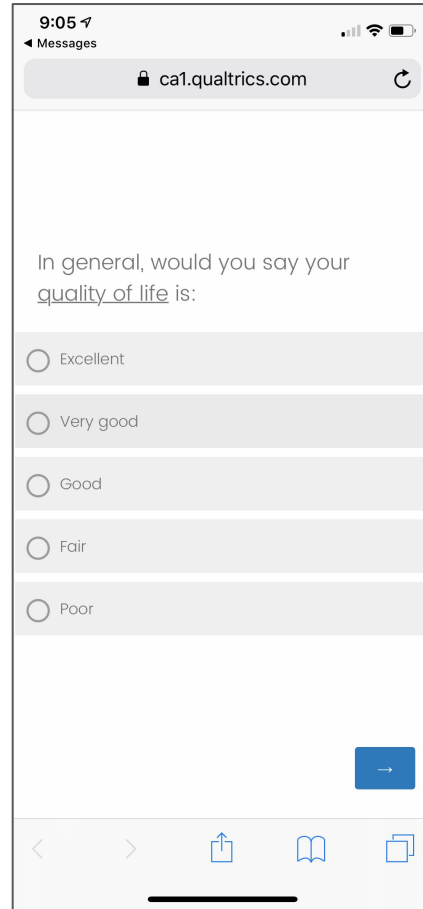
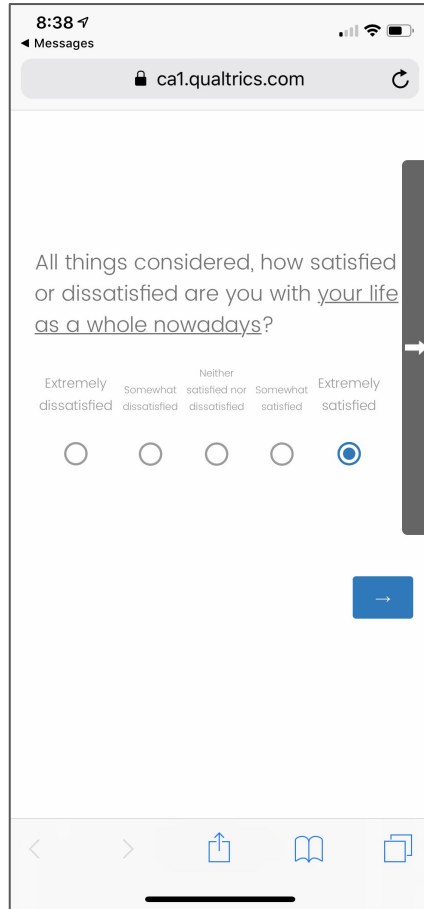
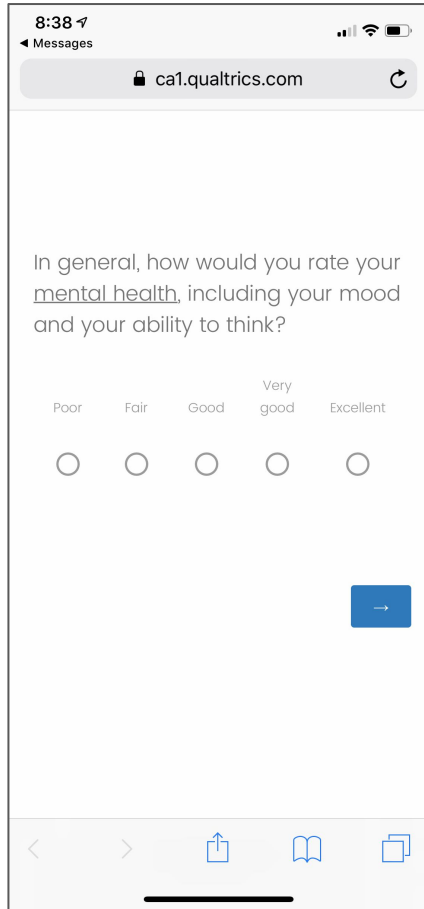


In general, how would you rate the quality of your usual social activities?

Excellent Very good Good Fair Poor



Mobile screenshots example of scale questions



Flowchart II Example

Would you describe in detail your quality of life, overall? Please say more than a single word answer.

Excellent



Your responses to the past two questions are a bit short. Would you like to review and revise those two descriptions before you continue to the end of the survey?

Yes

No



This question shows if previous open end:

- Matches scale point
- Is detected as “Gibberish”
- Both are one-word answers
- Does not look like an answer written in any discernible language

Flowchart III Example

When asked to rate your physical health, in general, you said: **"Excellent"**.

When asked why your physical health was "Excellent", you said: "I am miserable. Sicker than a dog."

These answers seem to be inconsistent, but we could be wrong.

Would you like to change your answer to the rating portion of this question (highlighted in yellow, above)?

I would like to change my answer to the rating question where I selected "Excellent"

I do not want to change my answer



You now have an opportunity to change your answer to the rating question, as requested. We have selected your previous answer; please change your answer to whatever you choose.

In general, would you say your physical health is ...

Excellent

Very good

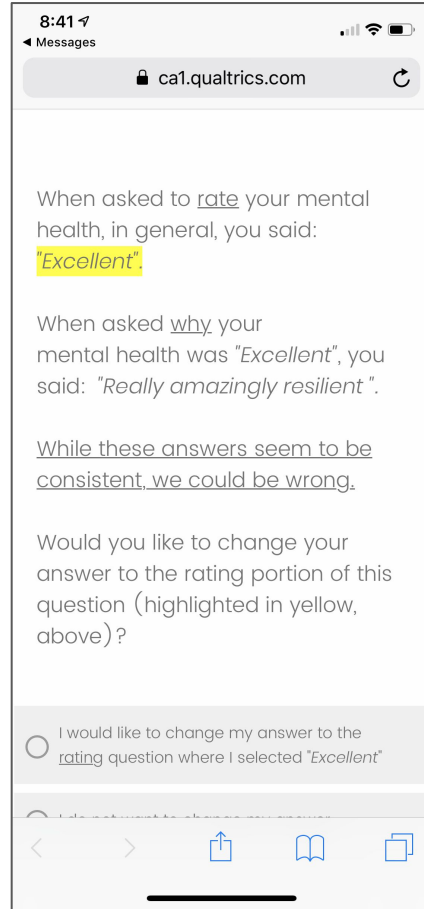
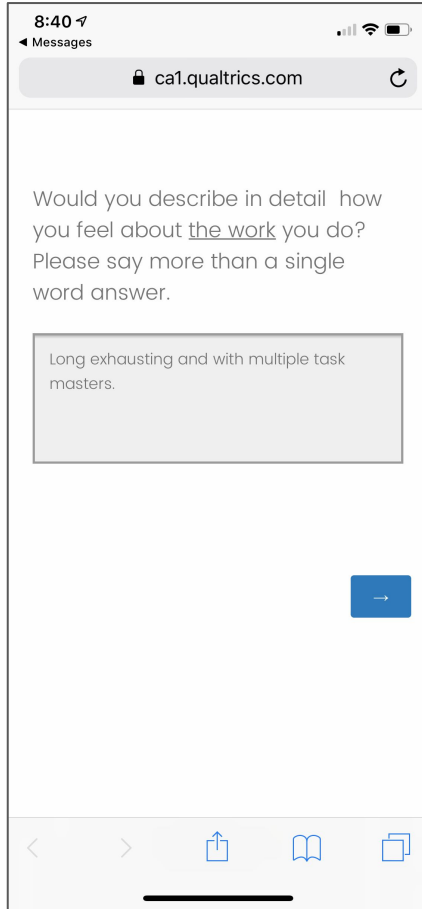
Good

Fair

Poor



Flowchart III Example on mobile device



Open ended examples

Examples of open ended

As a whole, my life is pretty good. I am retired and I am in good health. I get to spend time with my family and friends and take trips whenever I want to.

I am very satisfied with my financial situation, I have enough money to buy the things I want.

Not great, I feel rushed or I procrastinate.

The job could pay more than minimal wage!

Very good they offer great perks.

References

- Garbarski, D., Schaeffer, N. C., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research*, 24(6), 1443–1453. <https://doi.org/10.1007/s11136-014-0861-y>
- Garbarski, D., Schaeffer, N. C., & Dykema, J. (2016). The effect of response option order on self-rated health: A replication study. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 25(8), 2117–2121. <https://doi.org/10.1007/s11136-016-1249-y>
- Garbarski, D., Schaeffer, N. C., & Dykema, J. (2019). The effects of features of survey measurement on self-rated health: Response option order and scale orientation. *Applied Research in Quality of Life*, 14(2), 545–560. <https://doi.org/10.1007/s11482-018-9628-x>
- Liu, M., & Keusch, F. (2017). Effects of scale direction on response style of ordinal rating scales. *Journal of Official Statistics*, 33(1), 137–154. <https://doi.org/10.1515/jos-2017-0008>
- Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, 34(1), 78–94. <https://doi.org/10.1177/0894439315574248>
- Maloshonok, N., & Terentev, E. (2016). The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in Human Behavior*, 62, 506–515. <https://doi.org/10.1016/j.chb.2016.04.025>
- Mavletova, A., Couper, M. P., & Lebedev, D. (2018). Grid and item-by-item formats in pc and mobile web surveys. *Social Science Computer Review*, 36(6), 647–668. <https://doi.org/10.1177/0894439317735307>
- Revilla, M., & Couper, M. P. (2018a). Comparing grids with vertical and horizontal item-by-item formats for pcs and smartphones. *Social Science Computer Review*, 36(3), 349–368. <https://doi.org/10.1177/0894439317715626>
- Revilla, M., & Couper, M. P. (2018b). Testing different rank order question layouts for PC and smartphone respondents. *International Journal of Social Research Methodology*, 21(6), 695–712. <https://doi.org/10.1080/13645579.2018.1471371>
- Terentev, E., & Maloshonok, N. (2019). The impact of response options ordering on respondents' answers to rating questions: Results of two experiments. *International Journal of Social Research Methodology*, 22(2), 179–198. <https://doi.org/10.1080/13645579.2018.1510660>