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

AAPOR 74th Annual Conference, Toronto, Canada. May, 19 2019
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?*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale manipulation</td>
<td>Vertical</td>
<td>Vertical</td>
<td>Vertical and Horizontal</td>
</tr>
<tr>
<td>Positive to Negative and Negative to Positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online sample</td>
<td>U.S. KnowledgePanel</td>
<td>U.S. KnowledgePanel</td>
<td>U.S. Amazon Mturk</td>
</tr>
<tr>
<td>Results</td>
<td>Higher mean (healthier respondents) when scale ordered from Excellent to Poor</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

In general, would you say your **quality of life** is:
- Excellent
- Very good
- Good
- Fair
- Poor

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
Assigned to 1 of the 4 scale orientation conditions

Asked 2 out of 5 unipolar block and 2 out of 5 bipolar block questions (counterbalanced)

Follow up open end for each of the 4 questions

Quality of open end detection
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

<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>17.0%</td>
<td>15.5%</td>
<td>16.0%</td>
<td>16.9%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>68.4%</td>
<td>75.1%</td>
<td>74.8%</td>
<td>72.5%</td>
</tr>
</tbody>
</table>

**“Virtual Coder”-Controlled**  
Base: all unipolar responses

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

<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>9.2%</td>
<td>13.8%</td>
<td>12.7%</td>
<td>10.1%</td>
</tr>
</tbody>
</table>

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

```
<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0%</td>
<td>10.1%</td>
<td>9.2%</td>
<td>9.3%</td>
<td></td>
</tr>
</tbody>
</table>
```

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

```
<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.8%</td>
<td>8.3%</td>
<td>11.6%</td>
<td>10.2%</td>
<td></td>
</tr>
</tbody>
</table>
```

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

**“Virtual Coder”-Controlled**
Base: all bipolar responses

Inconsistency between scale and open end detected N= 857

```
<table>
<thead>
<tr>
<th></th>
<th>Hor_Neg_Left</th>
<th>Hor_Pos_Left</th>
<th>Vert_Neg_Top</th>
<th>Vert_Pos_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.0%</td>
<td>60.0%</td>
<td>60.0%</td>
<td>60.0%</td>
<td></td>
</tr>
</tbody>
</table>
```

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

The mobile darker bar color means that the difference with the lighter desktop bar color next to it is statistically significant at $p < 0.05$
Results for bipolar scales, by device

**Respondent-Controlled**

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

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 $p < .05$. 

Trimmed mean in seconds
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.

Higher inconsistency showed for mobile respondents.

Bipolar scales showed higher inconsistency for vertically oriented scale.

Unipolar question showed higher inconsistency for Hor_Pos_Left & Vert_Neg_Top.
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>3 bipolar fully labeled items</td>
<td>Agree - Disagree vs. Disagree - Agree</td>
<td>Unipolar scale Least to Most and Most to Least</td>
</tr>
<tr>
<td>Online sample</td>
<td>Russian MOOC students</td>
<td>U.S. KnowledgePanel</td>
<td>Russian MOOC students</td>
</tr>
<tr>
<td>Results</td>
<td>Mixes results: Only one item show stat. Sign differences in horizontal (primacy)</td>
<td>Higher acquiescence response style when scale presented with agree first</td>
<td>Primacy effects</td>
</tr>
</tbody>
</table>
## Previous research on scale orientation

**Desktop vs. Mobile respondents**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>6 items with fully labels response options</td>
<td>List of checkboxes</td>
<td>Set of different scales endpoint labeled</td>
</tr>
<tr>
<td>Online sample</td>
<td>Russian Opt-in Panel</td>
<td>Dutch LISS panel</td>
<td>Spanish Opt-in Panel</td>
</tr>
<tr>
<td>Results</td>
<td>Higher primacy effects in Desktop vs. Mobile</td>
<td>Higher primacy effects for Mobile vs. Desktop</td>
<td>More answer changes on Mobile Vs. Desktop Larger primacy effects on Mobile vs. Desktop</td>
</tr>
</tbody>
</table>
### Age & Gender by device N= 4,521

<table>
<thead>
<tr>
<th>Age</th>
<th>% Desktop</th>
<th>% Mobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-24</td>
<td>2.4</td>
<td>15.6</td>
</tr>
<tr>
<td>25-34</td>
<td>11.6</td>
<td>21.0</td>
</tr>
<tr>
<td>35-44</td>
<td>12.0</td>
<td>19.8</td>
</tr>
<tr>
<td>45-54</td>
<td>16.8</td>
<td>17.2</td>
</tr>
<tr>
<td>55-64</td>
<td>21.5</td>
<td>15.0</td>
</tr>
<tr>
<td>65+</td>
<td>35.7</td>
<td>11.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>% Desktop</th>
<th>% Mobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>55.2</td>
<td>44.0</td>
</tr>
<tr>
<td>Female</td>
<td>44.8</td>
<td>56.0</td>
</tr>
</tbody>
</table>
Education by device N= 4,481

<table>
<thead>
<tr>
<th>Education</th>
<th>% Desktop</th>
<th>% Mobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>17.6</td>
<td>26.0</td>
</tr>
<tr>
<td>Associate degree and some college</td>
<td>30.8</td>
<td>35.0</td>
</tr>
<tr>
<td>Master and Bachelor</td>
<td>46.6</td>
<td>34.9</td>
</tr>
<tr>
<td>Doctoral and professional degree</td>
<td>5.0</td>
<td>4.1</td>
</tr>
</tbody>
</table>
Qualtrics - Google NLP API integration:

URL: https://language.googleapis.com/v1beta2/documents:analyzeSentiment
Method: POST

Query Parameters
- key = AlzaSyDYSqxekJ1ljz-9YDnJQ1aDHIEAtNDv5TY

Body Parameters
- encodingType = String (UTF8)
- document.type = String (PLAIN_TEXT)
- document.content = String (${q://QID27/ChoiceTextEntryValue})

Custom Headers
- fields = Set a Value Now

Fire and Forget

Set Embedded Data
- documentSentiment.score = documentSentiment.score
FAQs
FAQs

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)
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

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

- Poor
- Fair
- Good
- Very good
- Excellent

All things considered, how satisfied or dissatisfied are you with your life as a whole nowadays?

- Extremely dissatisfied
- Somewhat dissatisfied
- Neither satisfied nor dissatisfied
- Somewhat satisfied
- Extremely satisfied

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

- Excellent
- Very good
- Good
- Fair
- Poor

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

- Poor
- Fair
- Good
- Very good
- Excellent
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

Would you describe in detail how you feel about the work you do? Please say more than a single word answer.

Long exhausting and with multiple task masters.

When asked to rate your mental health, in general, you said:

"Excellent."

When asked why your mental health was "Excellent", you said: "Really amazingly resilient".

While these answers seem to be consistent, 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"
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


