How does response bias emerge in lengthy sequential preference judgments?

Masahiro Morii¹ · Takayuki Sakagami² ·
Shinya Masuda³ · Shigetaka Okubo⁴ · Yuki Tamari⁵

Received: 14 March 2017 / Accepted: 8 September 2017 / Published online: 18 September 2017
© The Author(s) 2017. This article is an open access publication

Abstract Internet surveys are currently used in many academic and marketing research fields. However, the results for these surveys occasionally show traces of response bias. In our study, we analyzed how response bias appears in lengthy preference judgments. 1042 respondents participated in lengthy sequential preference judgments. Three stimuli series were used: scene pictures, Attneave nonsense shapes, and point-symmetric figures. One hundred stimuli were selected for each series and individually displayed on a computer screen, with presentation order randomized for each respondent. Respondents were then asked to rate their degree of preference for each stimulus. Mean preference scores increased over the first 10–20 trials, then, gradually decreased from the middle to the last trial. Furthermore, participants tended to produce the maximum and minimum score during early trials. These results demonstrated that response bias can be a function of presentation order.

Communicated by Kazuhisa Takemura.

Electronic supplementary material The online version of this article (doi:[10.1007/s41237-017-0036-6](https://doi.org/10.1007/s41237-017-0036-6)) contains supplementary material, which is available to authorized users.

✉ Masahiro Morii
masa.morii@gmail.com

¹ Global Centre for Advanced Research on Logic and Sensibility, Keio University, Mita 2-15-45, Minato-ku, Tokyo 108-8345, Japan

² Department of Psychology, Keio University, Tokyo, Japan

³ Faculty of Nursing and Medical Care, Keio University, Kanagawa, Japan

⁴ Department of Psychology, Waseda University, Tokyo, Japan

⁵ School of Management and Information, University of Shizuoka, Shizuoka, Japan

Keywords Sequential judgments · Preference · Midpoint response · Internet survey · Visual analog scale

1 Introduction

Questionnaires have been used in psychological, social, and marketing research, national population censuses, and many other fields for a long time. More recently, Internet surveys have also become as widely used as questionnaires. Such survey methods are now common tools for examining respondents' "attitudes" or "thinking". However, they are not without their drawbacks, as survey data occasionally show response bias (Baumgartner and Steenkamp 2001; Choi and Pak 2005; Vaerenbergh and Thomas 2013; Weijters et al. 2009). For example, survey results can be affected by questionnaire design (Tourangeau et al. 2013).

One common response bias is the midpoint response. The midpoint response is defined as choosing the locationally midpoint category (i.e., middle alternative) as an answer. The midpoint response was originally referred to as a "neutral" or "indifferent" attitude. However, this can also refer to answers such as "don't know," "no opinion," "never thought about it," and "undecided" (Raaijmakers et al. 2000; Sturgis et al. 2014). Krosnick (1991) suggested that the midpoint response is generally chosen as a result of "satisficing". "Satisficing" was originally defined as performing satisfactorily and sufficiently, but not optimally (cf. Simon 1957). In survey responses, respondents are asked to expend a great deal of cognitive effort such as interpreting question meanings, recalling memories, integrating information, and reporting their judgments. Because of these intensive processes, respondents tend to choose midpoint responses to minimize cognitive effort. As such, it can be difficult to determine whether the midpoint response is used to represent a neutral attitude, or a form of nonresponse. If, in a survey, respondents provide too many midpoint responses as a means of representing nonresponses the survey results may be concentrated near the midpoint of the response scale to an excessive degree. Consequently, this could cause misunderstanding of "actual" respondents' answers.

In this study, we attempted to analyze the "sequential effect" in lengthy surveys. The sequential effect was originally applied in psychophysics to conduct tests such as judgments of luminance or the length of a line segment (Holland and Lockhead 1968). A similar phenomenon is found in behavioral psychology, which is known as "habituation" where response rate decreases because of the repeated presentation of stimuli (Harris 1943).

Another form of behavioral change similar to the sequential effect is known as the "museum-fatigue effect." Gilman (1961) tested this effect by analyzing museum visitors' behavior in relation to their responses to art works. He found that visitors would scrutinize and appreciate art works early in the exhibitions, but as time passed, they would only glance at the works in passing. Gilman believed that the cause of this phenomenon was the visitors' fatigue. This museum-fatigue effect has been found in many experimental observations and laboratory experiments (Falk

et al. 1985; Robinson 1928; Serrell 1997) and, previously, the main factor driving the museum-fatigue effect was believed to be visitors' fatigue. However, Bitgood (2009) demonstrated that other cognitive factors such as satisfaction, information overload, and limitations in attentional capacity can also cause the museum-fatigue phenomenon.

Previous studies have shown that similar phenomena can be found regarding survey responses. In survey research, the sequential effect is known as the "order effect" or the "context effect." For example, Kraut et al. (1975) examined the effect of presenting opinion-survey items in different positions (in either earlier positions, or in later positions). They found that respondents gave less extreme answers when items were placed later in a questionnaire, relative to earlier. Furthermore, Knowles (1988) demonstrated that item answers become more polarized, consistent, and reliable as the number of previous items answered increases. Thus, responses to a particular question are likely to be affected by the sequential position in which the question is presented and the content of previous questions. Such response bias can become significant, especially in lengthy questionnaires, because each question may have an interrelation with the others, and lengthy questionnaires involve a high "cognitive cost."

These response biases have also been observed in preference judgments (Dijksterhuis and Smith 2002; Leventhal et al. 2007). In a study by Leventhal et al. (2007), participants were repeatedly exposed to pleasurable stimuli (pictures of people playing water sports), and were then asked to rate how pleasurable they found each stimulus using paper-based visual analog scales (VAS). The results demonstrated that, overall, a negative slope regarding the degree of pleasure was observed as the sequence progressed. The researchers called this "affective habituation."

As mentioned above, many kinds of behavioral and cognitive factors affect respondents' behaviors in surveys, and these factors can distort survey findings. Previous research papers have proposed methods for detecting these biases, particularly through altering questionnaire designs. One common method for eliminating the sequential effect is to randomize the order of the questions. This procedure has been proven to reduce the sequential effect, but does not provide a completely satisfying solution.

In this study, we conducted an Internet survey to examine the cause of response biases in lengthy, sequential preference judgments. Three stimuli series, landscape pictures, Attneave nonsense shapes, and point-symmetric figures were used, and 100 stimuli were selected for each series. The stimuli were presented one-by-one on a computer screen and respondents were asked to rate each stimulus in regard to their "preference" using a VAS that they manipulated using the computer's mouse. The order of presentation for the stimuli was randomized for each participant. Thus, if some sort of sequential bias was present, the results would show sequential trends that reflected the order in which the stimuli were presented regardless of stimulus contents. We then analyzed whether the sequential effect was present in lengthy preference judgments. We have discussed response bias in lengthy, sequential judgments that focus on preference for stimuli, and how and why these biases appear differently depending on stimuli series (Fig. 1).

Fig. 1 Examples of stimuli (top: Series A, middle: Series B, bottom: Series C)

2 Materials and methods

2.1 Participants

For this experiment, 1042 participants aged from their 20–60s participated in our survey. Each participant was paid JPY 400 (approx. USD 3.50) for their participation. These participants were then randomly assigned one of six groups so that each group possessed an approximately equal number of members in terms of age and sex. After completion of the survey, the data for 255 participants were removed from the analysis, as these were considered to be “untrustworthy” data (see 3.1 and 3.2 for details).

2.2 Stimuli

We used three stimulus series (Series A: landscape pictures, Series B: Attneave nonsense shapes; Series C: point-symmetric figures). Examples of the stimuli are shown in Fig. 1. A total of 135 stimuli for each series were used in the preliminary survey, while 100 stimuli were used in each series during the main survey. Stimuli were selected or generated in accordance with the criteria described in the following section.

- Series A: Landscape pictures

For Series A, we searched for suitable landscape pictures using Google Images (<https://images.google.com/>). During the search, we applied Google Images' "transparent" filter to obtain images that would foster less bias in terms of colors. We disregarded images that featured people and those that had a high vertical length. The search was conducted using the following input and filter:

- Search word: “風景” (“landscape” in Japanese).
- Transparent: red, yellow, green, teal, blue, purple, white, gray, or black.
- Image size: more than 800 pixels × 600 pixels.
- Usage rights: “labeled for reuse with modification”.
- Series B: Attneave nonsense shapes

For Series B, 135 Attneave nonsense shapes were selected from Vanderplas and Garvin (1959). Shapes that had high association values with real, existing objects were eliminated, along with those that were axisymmetric or point-symmetric. Each image chosen for Series B had either 4, 5, 8, 12, 16, or 24 sides.

- Series C: Point-symmetric figures

For Series C, 135 point-symmetric figures were developed by Fourier descriptors referring to the study by Zahn and Roskies (1972). The shapes were created using the programming language “R” (see the supplementary material for details regarding the R functions). All shapes had rotational symmetries of either 72°, 90°, 120°, or 180°.

2.3 Preliminary survey

Forty undergraduate and graduate-school students (22 males, 17 females, and one unidentified) from Keio University participated in a preliminary survey. In this survey, participants rated all three stimuli series in terms of their degree of preference. There were 135 stimuli for each series, meaning each respondent rated a total of 405 stimuli. Each stimulus was presented on the screen for 7 s and the three series were presented in the following order: Series A, Series B, and Series C. On the response paper, a VAS with 101 scale marks were presented on a scale and “very preferable” and “very unpreferable” were written in Japanese on the right and left sides of the scale, respectively. The respondents gave their degree of preference by drawing a diagonal line through the mark that corresponded with their degree of preference.

The answers were tallied in the form of scores from 0 to 100, and we then calculated the score variance. As we wished to select stimuli that had low variation regarding participants' rating, we eliminated stimuli that had a large variance. A total of 100 stimuli were then selected for each series and used in the main survey.

2.4 Procedure

Our survey was conducted in December 2013 through an Internet survey firm. Participants were randomly assigned to a combination of two stimuli series; there

were six possible combinations of series in this regard: AB, BA, AC, CA, BC, and CB.

In the survey, the stimuli were sequentially presented one-by-one and the stimuli presentation order was randomized between participants. Participants were asked to give their degree of preference for each stimulus using a VAS with 101 scale marks. On the scale, “very preferable,” on the right side, and “very unpreferable,” on the left side, were presented in Japanese. The slider for responding was not initially presented. Once the participant clicked anywhere on the scale bar, a slider appeared just under the mouse cursor and a “Next” button appeared under the scale bar. Participants could then adjust their answers by dragging the slider on the scale bar. When the “Next” button was clicked, their answer was recorded and the next stimulus appeared. Their answers were given a value from 0 to 600 based on the position of the slider, and these values were later re-scaled into a range from -100 to 100 . An example of the stimulus display is shown in Fig. 2.

3 Results

3.1 Exclusion of “untrustworthy” data by the Internet survey firm

The Internet survey firm analyzed all 1042 participants’ data and removed 75. These 75 were excluded because they satisfied one of the following criteria:

- (1) total response time for all questions was too long (more than 5 h),
- (2) total response time for all questions was too short (less than 270 s),
- (3) more than or equal to 95% of responses fell within the range of -5 – 5 (in the -100 – 100 range) in either or both series.

Consequently, 967 participants remained.

Fig. 2 An example of the stimulus display (left: before clicking on the scale; right: after clicking on the scale)

3.2 Analysis of the interquartile range among participants

Although the “untrustworthy” data had been removed, some data from the remaining 967 participants were still concentrated within a narrow response range. Consequently, we calculated the interquartile range (IQR) for individual data to detect data centralization and removed data with an IQR less than 10 (from -100 to 100). Applying this criterion showed that half of the 100 responses were represented by five scale marks on the response scale. After applying this filter, the number of remaining participants’ data was 619 for Series A, 534 for Series B, and 563 for Series C (as shown in Table 1); these data were used for our following analysis. For all stimulus series, the number of excluded data was larger in the second phase than in the first phase.

3.3 Mean scores for the three series

We calculated the mean scores and the maximums (and minimums) of the mean scores for the three series, which are shown in Table 2. Series A had the highest mean score. In Series B, the mean score and standard deviation were lower than in other series, and the maximum score was less than 0, which represented the scale midpoint. In Series C, the maximum score was higher than that for Series B but the minimum score was lower than that for Series B.

Table 3 shows the mean score for each pattern (which concerns each two-series combination and the order of series presentation). For Series A, when respondents were presented with this series in their second phase (after providing answers for either Series B or C), the mean scores were higher than when Series A appeared in the first phase. For Series B, the mean scores were higher in the first phase than in the second phase. For Series C, the mean score was lower after the respondents had answered for Series A, but the mean score was higher after they had first given answers for Series B. It was expected that the scores for Series B in Pattern “B to A” and “B to C” would be identical because the answering conditions for Series B were identical. However, there was a significant difference between these scores

Table 1 The number of remaining data after the elimination by the IQR criteria; the number of eliminated participants is noted in the brackets

Pattern	Series A	Series B	Series C
A to B ($n = 160$)	157 (3)	123 (37)	–
B to A ($n = 161$)	155 (6)	140 (21)	–
A to C ($n = 156$)	148 (8)	–	120 (36)
C to A ($n = 171$)	159 (12)	–	155 (16)
B to C ($n = 158$)	–	140 (18)	138 (20)
C to B ($n = 161$)	–	129 (32)	150 (11)
Total ($n = 967$)	619 (29)	534 (106)	563 (83)

Bold values indicate that the particular series was presented in the first phase

Table 2 Mean, maximum, and minimum scores for each series

Series	Mean (SD)	Maximum	Minimum
A ($n = 619$)	20.92 (14.16)	51.46	−23.00
B ($n = 534$)	−20.68 (6.94)	−3.32	−32.66
C ($n = 563$)	−9.77 (12.90)	38.79	−38.28

Table 3 Mean scores for each series in regard to all six two-series combinations

Pattern	Series A	Series B	Series C
A to B	19.50	−25.11	−
B to A	22.29	−20.64	−
A to C	18.85	−	−17.44
C to A	22.91	−	−9.57
B to C	−	−13.66	2.02
C to B	−	−24.28	−14.69

Bold values indicate that the particular series was presented in the first phase

[$t(278) = 2.62$, $p < 0.01$]. Similarly, there was a marginal difference between the scores for Series C in Patterns “C to A” and “C to B” [$t(303) = 1.93$, $p = 0.055$]. We asked the Internet survey firm whether there were any sampling biases or any other bias among these patterns, but were informed that there was no bias regarding pattern assignment. We could not identify why these differences were observed. We consequently decided to use the individual normalized scores in some of the following analyses to eliminate these unaccountable biases.

3.4 Analysis of sequential trends during evaluations

- Sequential changes in evaluations

As stimuli presentation order was randomized for each participant in our survey, we arranged all response data corresponding to presentation order (from 1 to 100) and each participant rated a different stimulus at each point. If no sequential effect was present in the series, the results would not show any sequential trends. The mean re-scaled scores (from −100 to 100) are shown in Fig. 3; these data show that small differences existed between the first and second phases for each series.

We calculated the individual normalized scores to offset the series order effect, and the results of this are shown in Fig. 4. The re-scaled scores showed a rise during the first 10–20 trials for every series. Furthermore, this trend is even more remarkable when the normalized scores are considered. The normalized scores rose at the beginning of the sequence before gradually decreasing towards the last point. For Series A, the start point was higher than 0. Meanwhile, the start points for both Series B and C were less than 0, and these were the lowest scores in the entire series.

- Response ratio for the middle area and change in standard deviation

Fig. 3 Sequential trends regarding presentation order (mean re-scaled scores from -100 to 100)

Fig. 4 Sequential trends regarding presentation order (individual normalized scores)

The response ratio for the middle area (from -20 to 20 on the re-scaled score from -100 to 100) is shown in Fig. 5. The starting point marked the lowest score for the ratio. However, once the ratio rose to approximately 0.4 , it remained at that point. Series A shows the most noticeable trend; at the starting point the ratio was lower than that of the other two series at the same point, but the ratio at the end was the highest of all series (~ 0.45).

- Sequential changes of various indexes

The sequential variation of standard deviation (SD) is shown in Fig. 6. For Series A, the SD was highest at the starting point and then slightly decreased. For Series B, the SD was lower at the starting point. However, the SD gradually increased during

Fig. 5 Sequential trends regarding response ratio for the middle area (from -20 to 20)

Fig. 6 Sequential variation of standard deviation

the approximately first 20 trials before maintaining this level to the end. For Series C, the SD remained relatively higher than the other two series from the start to the end and did not show any sequential trends.

We analyzed the absolute difference between each score and that of the preceding question as well as each score's correlation with the next score, and these results are shown in Figs. 7 and 8, respectively. In all three series, it was found that the absolute difference from the last value decreased as a function of the presentation order. Figure 8 shows that the correlation coefficients with succeeding questions were high (more than 0.5) for all series even though they were relatively lower for Series A in early trials (less than 0.5).

Fig. 7 Change in absolute difference from the previous score

Fig. 8 Correlation with the preceding question

Figure 9 shows the correlations of scores for two questions as a function of distance between questions. The scores showed a high correlation when the distance between questions was small, but this decreased as the distance increased.

3.5 Location of maximum and minimum values

We conducted an analysis to determine where individual maximum (Max) and minimum (Min) values appeared. Figure 10 shows the distribution of individual Max and Min values in each set of 10 trials. If the same Max or Min value appeared more than once, the first point to appear was counted. The results showed that the Min values were most frequently observed during the first 10 trials. The Max value

Fig. 9 Correlation coefficients as a function of distance between questions

Fig. 10 The distributions of Max and Min values for each set of 10 trials

was also frequently observed early during Series A, but this tendency was not significantly present in Series B or C.

We also analyzed evaluations after the maximum and minimum points. Figure 11 shows the data for re-scaled scores after Max and Min values. As can be seen, the scores remained high during the period after the Max values and low during the period after the Min values. The values immediately after the Max and Min values strongly depended on the previous values.

Fig. 11 Scores after the Max and Min values

4 Discussion

The purpose of this study was to examine response bias in lengthy sequential judgments and to examine how this form of bias differs between different stimuli series. In our survey, the stimuli were presented one-by-one in random order and participants were asked to rate their preference for stimuli using a VAS. We arranged all response data corresponding to presentation order and analyzed the sequential trend of rating. Results showed that there were differences among three stimuli series. Furthermore, for all stimuli series, the mean scores were different between they were evaluated in the first or second phase. As we intended to focus on analyzing sequential trends in respondents' behaviors, we calculated individual normalized scores and analyzed them. With individual normalized scores, we could compare sequential patterns regardless of stimuli series and phases. The results showed common response biases for all series. As the trials proceeded, midpoint responses increased (Fig. 5) and the absolute difference from the previous score decreased (Fig. 7). Furthermore, the correlation coefficients decreased as a function of the distance of two questions as shown in Fig. 9. These findings are consistent with previous studies (Knowles 1988; Kraut et al. 1975).

On the other hand, the mean scores for each series differed as shown in Tables 2 and 3. The mean score was relatively higher for Series A, and it was relatively lower for Series B. Furthermore, comparing six stimulus-patterns (involving each two-series combination and stimuli series order) showed different results. When respondents evaluated Series A (which was found to be the most preferable stimulus series) during the first phase, they tended to prefer Series B or C much less during the second phase. On the other hand, when respondents evaluated Series B (the most unpreferable stimulus series) during the first phase, they tended to evaluate Series A or C more preferably during the second phase. This finding could be regarded as a result of the contrasting effects between the first and the second phase. During the

second phase, respondents evaluated the stimuli by comparing them to stimuli from the first phase.

We found different sequential changes among the three series (Fig. 4). In Series A, the normalized score was high at the starting point and increased for a small number of trials before consistently decreasing. Meanwhile, for Series B and C, the normalized scores were low at the starting point. After the scores for these series increased over approximately 20 trials, they then gradually decreased from the mid to the last trial. In addition, although the standard deviation gradually decreased for Series A as the trial proceeded, they did not decrease for Series B and C as shown in Fig. 6. It was considered that this difference might be a result of the difference in the mean scores between the three series. The stimuli from Series A were preferred more than those from Series B and C (Tables 2, 3). Namely, for each series the global preference scores affected the starting point for this sequential pattern. As the trial proceeded, the series' values rose from their respective starting points during the approximately first 10–20 trials, and then converged at the middle point.

It is noteworthy that the Max and Min values frequently emerged early in each series (Fig. 10). It was expected that these values would be evenly distributed across the full series sequences because presentation order was randomized for each respondent. However, our results showed that the Max and Min values did not appear evenly throughout the series. Furthermore, the distribution (location) of Max and Min values differed between the three series. For all three series, the Min values were clearly concentrated early in the sequence. On the other hand, the Max values were less concentrated in the early sequence stages for Series B and C. We believe that this frequency difference for Max values was caused by the differences between the series regarding the mean values. For Series A, which had high overall rating scores, the Max values appeared early during the sequence. On the other hand, in Series B and C, the overall values and Max values were lower than those for Series A. It followed that the Max values for Series B and C were dispersed in comparison to those of Series A.

Figure 11 shows that after the individual Max (Min) value, respondents rated following stimuli as more preferable (unpreferable). Furthermore, correlation coefficients were high when the distance between two questions was narrow, and they decrease as a function of distance between two questions. These results indicated that their ratings were attracted toward previous ratings even if these scores should be independent each other, which consistent with the non-zero correlation with the succeeding question as shown in Fig. 8.

One possible explanation for the increasing midpoint responses may relate to the satisficing process explained by Krosnick (1991). Because many questions required a high degree of cognitive resources, the respondents increased their ratio of midpoint responses to conserve those resources. The habituation process can also explain these increasing midpoint responses. Repeated exposure to stimuli made respondents reluctant to give extreme responses. However, although these two processes can explain the increasing midpoint responses, they cannot explain the preference increase during the early trials. This preference increase can be considered as a result of the sensitization process. Sensitization is defined as when repeated exposure to stimuli results in a progressive response amplification, which is

the opposite to habituation (Shettleworth 2010). Our results could be explained as follows: after the mean scores for the stimuli increased during the early trials through the process of sensitization, they then began to decrease and the midpoint response increased as the habituation process became more pronounced in the middle to final trials. This dual-process can also explain why the distribution of Max and Min values tended to appear early during each series sequence as shown in Fig. 10. A dual-process with sensitization and habituation was proposed for explaining response plasticity in classical conditioning experiments conducted by Groves and Thompson (1970). In their study, animals first showed an increase in responsiveness (sensitization) and then later a decrease in responsiveness (habituation) to repeated stimulation. Furthermore, Groves et al. (1969) showed that the habituation and sensitization process varied depending on stimulus frequency and intensity. The sequential trends of normalized scores observed in our survey seemed to resemble behavioral patterns explained by their dual-process theory. However, “evaluation” for preference seems to have different characteristics from “behavioral” data such as animal behavior obtained from laboratory experiments. Further examinations are required to examine whether sequential trends observed in our survey resulted from sensitization and habituation process. In future study, we would like to create a model which can explain these sequential changes.

There is another explanation for sequential trends in preference judgments. This is that “midpoint responses” observed in later trials could be an “actual (true) value” and the whole process is a “de-biasing” process. In this idea, the initial rise and larger deviation could be regarded as noises caused by certain factors such as sensitization. In any explanations, we would not explain the sequential changes with a single process and we need to regard as a dual-process. We could not identify which explanation is appropriate: therefore, further analysis is required to do so.

It should be noted that our study was conducted as an Internet survey. For Internet surveys, researchers cannot directly observe respondents’ behaviors and responses may, unbeknownst to the researcher, be subject to bias. In fact, some participants answered all questions in such a short time that it could be inferred that they did not seriously consider their answers. Other participants took a very long time to answer, which could have been because they took a break while responding to the questions. These answers could be regarded as “inexact answers”. However, it is difficult to establish a boundary between “exact” and “inexact” answers.

One limitation of our study was that the stimuli and procedure used were quite simple. Further study is required to generalize our result to different stimuli and question formats. However, as we found by analyzing the IQR filter, the amount of eliminated data for Series A (the relative attractiveness series) was less than that for Series B (relative unattractiveness series). This indicates that stimulus series attractiveness seems to affect response bias: less attractive stimuli may cause an increase in response bias. Furthermore, for all stimulus series, the amount of eliminated data was larger when they were rated in the second phase than when they were rated in the first phase. This also could imply that response biases appear later in lengthy surveys. Our finding that stimulus attractiveness and questionnaire length cause response bias could also apply to general surveys. These points should be addressed in future research.

Our survey format results have strong applicability for survey research, and the factors raised within this paper should be considered in this field. In future studies, we would like to analyze respondents' behavior individually through laboratory experiments. Tracking eye movements or a mouse cursor in terms of response behavior would provide us with further understanding of the behavioral characteristics of this form of behavior.

Compliance with ethical standards

Funding This work was supported by MEXT's Grants-in-Aid for Scientific Research (A) (Grant number JP16H02050) and the MEXT-Supported Program for the Strategic Research Foundation at Private Universities (2012–2014).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Baumgartner H, Steenkamp J (2001) Response styles in marketing research: a cross-national investigation. *J Mark Res* 38(2):143–156
- Bitgood S (2009) Museum Fatigue: a Critical Review. *Visitor Studies* 12(2):93–111
- Choi BCK, Pak AWP (2005) A catalog of biases in questionnaires. *Prev Chronic Dis* 2(1):1–13
- Dijksterhuis A, Smith PK (2002) Affective habituation: subliminal exposure to extreme stimuli decreases their extremity. *Emotion* 2(3):203–214
- Falk JH, Koran JJ, Dierking LD, Dreblow L (1985) Predicting visitor behavior. *Curator* 28(4):249–257
- Gilman BI (1961) Museum fatigue. *Sci Mon* 2(1):62–74
- Groves PM, Thompson RF (1970) Habituation: a dual-process theory. *Psychol Rev* 77(5):419–450. doi:[10.1037/h0029810](https://doi.org/10.1037/h0029810)
- Groves PM, Lee D, Thompson RF (1969) Effects of stimulus frequency and intensity on habituation and sensitization in acute spinal cat. *Physiol Behav* 4:383–388
- Harris JD (1943) Habituation response decrement in the intact organism. *Psychol Bull* 40(6):385–422
- Holland MK, Lockhead GR (1968) Sequential effects in judgments of loudness. *J Exp Psychol Hum Percept Perform* 3(1):92–104. doi:[10.3758/BF03205747](https://doi.org/10.3758/BF03205747)
- Knowles E (1988) Item context effects on personality scales: measuring changes the measure. *J Pers Soc Psychol* 55(2):312–320. doi:[10.1037//0022-3514.55.2.312](https://doi.org/10.1037//0022-3514.55.2.312)
- Kraut AI, Wolfson AD, Rothenberg A (1975) Some effects of position on opinion survey items. *J Appl Psychol* 60(6):774–776. doi:[10.1037/0021-9010.60.6.774](https://doi.org/10.1037/0021-9010.60.6.774)
- Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measure in surveys. *App Cognit Psychol* 5(3):213–236
- Leventhal AM, Martin RL, Seals RW, Tapia E, Rehm LP (2007) Investigating the dynamics of affect: psychological mechanisms of affective habituation to pleasurable stimuli. *Motiv Emot* 31(2):145–157. doi:[10.1007/s11031-007-9059-8](https://doi.org/10.1007/s11031-007-9059-8)
- Raaijmakers QAW, van Hoof A, Hart H, Verbogt TFMA, Vollebergh WAM (2000) Adolescents' midpoint responses on Likert-type scale items: neutral or missing values? *Int J Public Opin Res* 12(2):208–216
- Robinson ES (1928) *The behavior of the museum visitor*. Am Assoc Mus, Washington, DC
- Serrell B (1997) Paying attention: the duration and allocation of visitors' time in museum exhibitions. *Curator* 40(2):108–125
- Shettleworth SJ (2010) *Cognition, evolution and behavior*, 2nd edn. Oxford University Press, New York
- Simon HA (1957) *Models of man: social and rational*. Wiley, New York

- Sturgis P, Roberts C, Smith P (2014) Middle alternatives revisited: how the neither/nor response acts as a way of saying “I don’t know”? *Sociol Methods Res* 43(1):15–38. doi:[10.1177/0049124112452527](https://doi.org/10.1177/0049124112452527)
- Tourangeau R, Conrad FG, Couper MP (2013) *The science of web surveys*. Oxford University Press, New York
- Vaerenbergh YV, Thomas TD (2013) Response styles in survey research: a literature review of antecedents, consequences, and remedies. *Int J Public Opin* 25(2):195–217
- Vanderplas JM, Garvin EA (1959) The association value of random shapes. *J Exp Psychol* 57(3):147–154
- Weijters B, Geuens M, Schillewaert N (2009) The proximity effect: the role of inter-item distance on reverse-item bias. *Int J Res Mark* 26(1):2–12
- Zahn CT, Roskies RZ (1972) Fourier descriptors for plane closed curves. *Trans Computers* 21(3):269–281