Horizontal vs. Vertical Scales vs. Use of a Grid in Online Data Collection: Which Is Better? May 2018

William Mockovak
Bureau of Labor Statistics, 2 Massachusetts Ave, N.E., Washington, DC 20212

Abstract:
When designing online questionnaires, survey designers often have the option of choosing between the use of horizontal or vertical rating scales, or possibly a grid, if several survey questions use a common response scale. Previous research has investigated use of these alternative scales, but with inconsistent findings. As a result, some researchers advocate use of horizontal scales, others recommend the use of vertical scales,[1] whereas the use of grids has been generally discouraged. Although the use of alternative scales has been heavily researched, the familiarity and expertise of online respondents with different question formats have continued to increase, so studies that were done years ago may no longer be relevant. Therefore, this study revisited the topic of question format and its impact on data quality.

Using online instruments that presented the exact same question order, content, and scale direction (from positive to negative), unipolar horizontal and vertical scales were compared across two questions, and with the use of a grid in four additional questions. All individual questions and the grid appeared on separate web pages. Participants were recruited through Amazon's Mechanical Turk and saw only one version of the questions. Across the six questions that compared horizontal or vertical scales (N=193 and 229, respectively), no significant differences were found on five of the questions. On one question, the use of a horizontal scale led to significantly higher ratings. In the comparison of the 4-question grid (N=279) to the same questions presented using horizontal or vertical scales, three of the questions showed no significant differences among any of the contrasts (vertical vs. horizontal vs. grid). The only significant difference occurred on one question where the mean of the grid question was significantly higher than the mean for the horizontal scale, but not significantly higher than the vertical scale. Data quality comparisons among the grid, horizontal, and vertical scales were also explored, but no significant differences were found in scale reliability (measured with Chronbach's alpha), in the amount of straight-lining that occurred, in item non-response, or in the resulting factor structure. In summary, the question formats studied did not consistently affect the results or associated measures of data quality for the questions in this study.

Key Words: Horizontal vs. vertical rating scales, grids

1. Introduction

When designing online questionnaires, some commercial software packages offer the designer the option of choosing between the use of horizontal or vertical rating scales, or if the questions use a common scale, a grid. What research exists to help guide these choices?

Previous research has shown that when using rating scales for measuring attitudes and opinions, the direction and orientation of the scale can be important – for example, whether responses range from positive to negative or from negative to positive, and whether the response options are presented horizontally or vertically.

[1] https://www.une.edu/sites/default/files/Microsoft-Word-Guiding-Principles-for-Mail-and-Internet-Surveys_8-3.pdf

With horizontal scales, research has shown there can be a bias toward selecting responses on the left, and when positive options are also presented on the left, the response bias can be higher than when negative options are presented on the left (Chan, 1991). Bias can also occur with the use of vertical scales, with items at the top being selected more often. In addition, the bias can be worse with vertical scales compared to horizontal scales (Toepoel et al., 2009).

However, the research results about the impact of scale orientation are not conclusive (Toepoel et al., 2009; Keusch, 2012; Yan and Keusch, 2015). For example, Friedman and Friedman (1994) asked participants to rate six occupations on status. Their results showed that equivalent horizontal and vertical rating scales did not always elicit the same responses and, moreover, the direction of the difference was not consistent. Three ratings showed no statistically significant differences, two showed the expected pattern of obtaining a higher rating using a horizontal scale, but one showed the opposite pattern. Friedman and Friedman concluded that other factors must be at work. What could those factors be?

Christian and Dillman (2004) and Toepoel et al. (2009) identify a variety of factors that might affect scale ratings including verbal language, graphical language (size, shape, location, etc.), and numerical language (numbers associated with response options). Other possible factors include satisficing (Krosnick, 1991), cultural differences (Weng, et al., 2008), interpretive heuristics (Tourangeau et al., 2004), as well as the content of the question itself, since some questions are likely to elicit stronger responses that may be less susceptible to secondary factors.

In addition, although some researchers discourage the use of grids (Dillman, 2009), and problems with their use have been documented (Couper et al., 2013), especially on mobile devices, many online surveys continue to use grids to present rating questions.

To evaluate the impact of using different types of scales, an experiment was embedded into a study whose primary purpose was to obtain feedback about the effectiveness of a letter that is sent to respondents to encourage their continued participation in a survey. Unipolar rating questions were presented using horizontal, vertical, or grid scales. Two null hypotheses were of interest:

- Hypothesis 1. The same unipolar questions presented using a horizontal rating scale or a vertical rating scale with the same scale direction (positive to negative) will yield similar mean values and response distributions.

- Hypothesis 2. Unipolar questions presented in a matrix and individually on a separate Web page with either a horizontal or vertical scale with the same scale direction (positive to negative) will yield similar mean values, response distributions, and quality.

2. Method

Participants were asked to read a letter that thanked them for agreeing to participate in a survey and which explained "next steps" in the survey process. The letter was written to encourage continued participation in the survey. Participants were then asked to evaluate the effectiveness of the letter by completing an evaluation consisting of 18 questions. However, only six of the first seven questions were involved in this study.

Four unipolar questions about the uses of the Consumer Price Index (CPI) were presented using three formats: a grid, a horizontal scale, or a vertical scale. Since there were two additional questions on the form that could be presented using either vertical or horizontal unipolar scales, they were also included in the

study. Other questions on the form were not included because they used other types of scales (for example, Yes/no, mark all that apply, check a range, etc.).

Three online instruments were developed that used the exact same questions and question order, but with response scale formats that varied for the six questions being compared.[2] See Attachment 1 for a description of how question formats were distributed on the three online instruments. Participants saw only one version of the questions.

The grid of four CPI-use questions appeared on a single Web page, and no scrolling was required.

Each non-grid question, presented using either a vertical or horizontal scale, appeared alone on a single Web page.

Scale direction always went from positive to negative (left to right) for questions with a horizontal scale (or grid), and from positive to negative (top to bottom) for questions with vertical scales.

Participants were recruited through Amazon's Mechanical Turk by asking "We want your opinions about a letter that encourages people to participate in a survey." Participants were asked to read the survey letter and then answer questions about it.

Each of the three online versions was posted as a separate task (HIT[3]) on the Amazon Turk website. In an attempt to obtain comparable groups, recruiting for each instrument version occurred at the same times and days of three consecutive weeks. Data for the version that used a grid for the four questions of interest was collected in the first week, the version using horizontal scales was collected in the second week, and the version using vertical scales was collected in the third week.

Participants were paid $2.10 to complete the task which took between 5-8 minutes for most participants. The following criteria were used when recruiting participants: HIT approval rate greater than 97, location = U.S., and "Number of HITS Approved" greater than 500.

3. Results

The number of responses to each questionnaire version is shown in Table 1.

Table 1A. Number of Completed Cases Obtained for the Three Questionnaire Versions

Version	Count
1	229
2	193
3	279

Although an attempt was made to recruit the same number of participants for each questionnaire version, some participants answered more than one version of the questionnaire despite implementing steps to prevent this from happening. Participants who completed multiple forms were identified using their IP addresses. Participants' responses for the first form they completed were kept, but deleted for any subsequent forms they completed (10 duplicate reporters were deleted from Version 2, and 59 were deleted

[2] SurveyMonkey was used to develop the instruments.

[3] A Human Intelligence Task, or HIT, is a question that needs an answer. A HIT represents a single, self-contained task that a worker can work on, submit an answer, and collect a reward for completing.
https://www.mturk.com/mturk/help?helpPage=overview

from Version 3). A last-minute attempt to better balance the numbers in each group actually led to more respondents completing Version 3, but rather than delete these cases, they were kept since no bias was suspected.

Since the three questionnaire versions contained identical questions about gender, age, and education, some information about group equivalence was available. Chi-square tests were run on gender, age, and education, but none of the group differences were found to be significant. Table 1B shows the breakdown of the entire sample by gender, Table 1C by age, and Table 1D by education. The sample was roughly evenly split between males and females, skewed younger with 78 percent between ages 18 and 40; and was well educated, with 49 percent reporting graduating from college or higher, and 39.1 percent reporting some college or an Associate's degree.

Table 1B. Percentage of Males and Females in Sample

Gender	Percent	N
Male	46.7%	327
Female	52.3%	367

Table 1C. Distribution of Age

Distribution of Age	Percent
18-21	5.2%
22-30	37.6%
31-40	35.5%
41-50	12.0%
51-60	7.0%
61+	2.7%

Table 1D. Highest Level of Education

Distribution of Education	Percent
Less than high school	0.6%
High school graduate/GED	11.5%
Some college	27.8%
Associate's degree	11.3%
College degree (B.A., B.S., etc.)	36.4%
Some post-college education or training	3.9%
Master's degree or Ph.D.	6.7%
Professional degree (e.g., law, medicine, etc.)	1.9%

3.1 Effect of Scale Type

Six questions were studied. Two of the six questions were compared using only vertical or horizontal scales. Four additional questions that dealt with uses of the Consumer Price Index were presented as individual questions on separate web pages using horizontal or vertical scales, as well as in a 4-question grid (also on a separate page).

3.2 Horizontal vs. Vertical Scales

One of the two questions compared using only a vertical or horizontal scale is shown below with a vertical scale orientation and the assigned weights (the weights were not visible to respondents). This was also the first question in each instrument.

1. What is your general reaction to this letter? How convincing or persuasive would you say it is?
   - 5 o Very persuasive
   - 4 o Persuasive
   - 3 o Somewhat persuasive
   - 2 o A little persuasive
   - 1 o Not at all persuasive

In the analysis, which is summarized in Table 2, data for this question from the two instrument versions that used the vertical scale were combined. An ANOVA showed there were no differences between the groups using the horizontal or vertical scales ($p = .499$).

Table 2. What is your general reaction to this letter? How convincing or persuasive would you say it is?

Orientation of Scale	Mean	Std. Dev	N
Horizontal	3.36	.920	280
Vertical	3.31	.903	423
Overall	3.33	.910	703

The next question that was compared asked, "*How carefully did you read the confidentiality pledge that appears on the bottom of the letter?*" The response options were ordered as in the previous question.

The results are shown in Table 3. The mean score for the group using the horizontal scale was 3.48. The mean score for the group using the vertical scale was 3.10. In this case, the horizontal scale resulted in a significantly higher mean rating ($F=15.355$, $p<.000$).

Table 3. How carefully did you read the confidentiality pledge that appears on the bottom of the letter?

Orientation of Scale	Mean	Std. Dev	N
Horizontal	3.48	1.240	279
Vertical	3.10	1.293	423
Overall	3.25	1.285	702

Four uses of the Consumer Price Index were mentioned in the letter sent to respondents. Respondents were asked to rate the importance of each using the following scale:

- 5 o Very important
- 4 o Important
- 3 o Somewhat important
- 2 o A little important
- 1 o Not at all important

Average ratings obtained using the different scales are shown in Table 4. Higher mean scores indicate more positive ratings. A multivariate analysis of variance was run to determine if there were any significant

differences among the groups and to control for multiple tests. A significant Wilks' lambda (.032, F=5181.46, p<.000) was obtained.

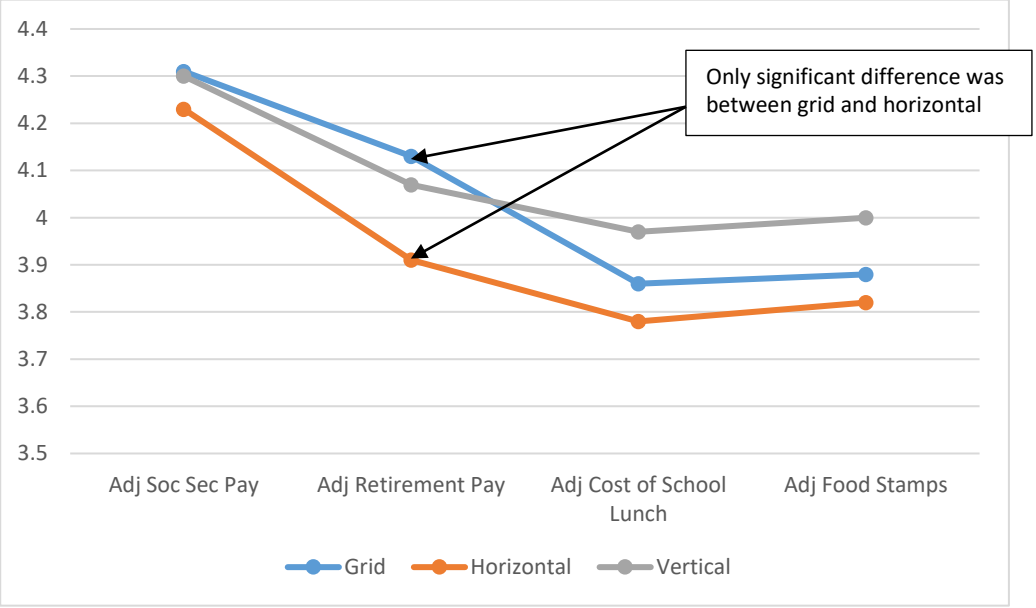
Post tests showed that the only significant difference occurred for the question that asked about the reported use of "Adjusting retirement payments." In this case, only the mean scores for the Grid and Horizontal scale ratings were significantly different.

In summarizing results from comparisons that asked for ratings of the four uses of the CPI, the horizontal and vertical scales produced statistically equivalent ratings for all four uses of the CPI, with the grid ratings significantly higher on only one variable, and that was in comparison with the horizontal scale rating (Tukey HSD, p=.039; LSD p=.015; Bonferroni p=.044). These comparisons are shown graphically in Figure 1. Although significantly different in only one comparison, horizontal ratings were consistently lower than ratings obtained using either vertical scales or a grid.

Table 4. Average Importance Ratings for Uses of the CPI by Type of Rating Scale

Reported Use of the CPI	**Grid**	**Std. Dev.**	**Horizontal**	**Std. Dev.**	**Vertical**	**Std. Dev.**
Adjusting social security payments	4.31	0.848	4.23	.888	4.30	0.857
Adjusting retirement payments	4.13	0.922	3.91	1.027	4.07	0.952
Adjusting the cost of school lunch programs	3.86	1.074	3.78	1.117	3.97	1.022
Adjusting food stamp benefits	3.88	1.071	3.82	1.084	4.00	0.998
Sample size	278		194		226	

Figure 1. Ratings of the Importance of Four Uses of the CPI Using Different Scales



Returning to Hypothesis 1, six questions were compared that used either horizontal or vertical scales. Of the six questions, use of a horizontal scale led to significantly higher ratings in only one instance: a question that asked how carefully the confidentiality pledge was read. In all other comparisons, the mean differences were not statistically significant. These inconsistent results led to acceptance of the null hypothesis.

For Hypothesis 2, four questions allowed comparison of grid questions with questions having horizontal or vertical scales. In this case, only one grid question produced a significantly different mean, with the grid question version producing a higher mean than the same question using a horizontal scale. Given the lack of consistent differences in mean values using different scaling approaches, other measures of question quality were explored.

Previous researchers have also found that grids can result in increased item non-response (Couper et al., 2013). However, that was not an issue with the grid used in this study. As shown in Table 5, although some non-response occurred, it was non-significant and very minor.

Table 5. Item Non-Response among Different Scale Approaches

Reported Use of the CPI	**Grid**	**Horizontal**	**Vertical**
Adjusting social security payments	0.7%	0%	1.7%
Adjusting retirement payments	0.4%	0%	1.3%
Adjusting the cost of school lunch programs	0.4%	0%	1.3%
Adjusting food stamp benefits	0.4%	0%	1.3%

Internal consistency or reliability is a measure of the quality of questions because it shows how closely the questions are related. Cronbach's alpha is a frequently used measure of scale reliability, so this measure was computed for the four questions that asked about the importance of uses of the CPI.

Table 6 shows how Cronbach's alpha (non-standardized) varied for questions presented using a grid or separately using horizontal or vertical scales. In all cases the values of Chronbach's alpha indicate acceptable reliability, since a value of 0.7 is viewed as a general cutoff.[4]

Table 6. Cronbach's Alpha for Different Scaling Approaches

Type of Scale	Cronbach's alpha	Scale Statistics	Std. Dev.
Grid	0.778	16.19	3.047
Horizontal	0.744	15.74	3.107
Vertical	0.755	16.33	2.913

All the scale formats have good internal reliability, and the differences among Chronbach's alpha in Table 6 are not statistically significant (Chi square = 0.8039, df=2, p = 0.669). See Diedenhofen and Musch (2016). These results suggest that the different scaling approaches result in essentially equivalent scales in terms of reliability.

Since the amount of straight-lining[5] is viewed by some researchers as a measure of satisficing and, hence, quality (Kaminska et al., 2010), the amount of straight-lining that occurred was calculated for each of the response scale formats. Table 7 presents the average amount of straight-lining that occurred with each scaling approach. An ANOVA showed no significant differences among the groups for any of the differing response values.

[4] https://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/

[5] Straight-lining was defined as occurring when a respondent selected the same response for each of the four questions that asked about the importance of uses of the CPI. Since there are five possible response options, there are five possible measures of straight-lining.

Table 7. Average Amount of Straight-lining That Occurred for Each Type of Response Format

	Value of Code				
Format	**5**	**4**	**3**	**2**	**1**
Grid	.21	.11	.01	.01	.00
Horizontal	.16	.08	.01	.00	.00
Vertical	.18	.09	.01	.00	.00
Overall	.19	.10	.01	.01	.00

The amount of straight-lining is often associated with the time taken to answer survey questions. However, although the time required to complete the entire questionnaire was collected, the time required to complete individual questions or subsections of the questionnaire was not collected and, therefore, will not be analyzed.

A final step for comparing the scaling approaches was a Principal Component Analysis. A one component solution was optimal for each of the scaling approaches. Table 8 shows the variance explained by the single component, and Table 9 shows the component score coefficients obtained for each scaling approach. These results show comparable results among the three scaling approaches. These results led to acceptance of the second null hypothesis.

Table 8. Percent of Variance Explained by Principal Component Analysis for Each Scaling Approach

	% of Variance Explained by One Component
Grid	60.9%
Horizontal	57.3%
Vertical	58.3%

Table 9. Component Score Coefficients for Each Scaling Approach

	Component Coefficient		
Reported Use of the CPI	**Grid**	**Horizontal**	**Vertical**
Adjusting social security payments	.337	.353	.339
Adjusting retirement payments	.327	.315	.350
Adjusting the cost of school lunch programs	.296	.309	.273
Adjusting food stamp benefits	.320	.343	.341

4. Discussion

Couper et al. (2013) present a nice summary on the impact of using grids, pointing out that the evidence is fairly strong that they reduce completion time, but the results are less consistent when looking at item-missing rates or inter-item correlations. Although alternative scale formats have been widely researched, the familiarity of online respondents with different question designs has continued to increase, so studies that were done years ago may no longer be relevant. Therefore, this study revisited the topic of question format and its impact on data quality.

The main conclusion from this study is that the question formats studied did not consistently affect the means or associated measures of data quality.

When means obtained using a four-item grid were compared to the same questions using horizontal or vertical scales, there were no significant differences among three of the questions (vertical vs. horizontal

vs. grid). Individual questions and the grid appeared on separate web pages. The only significant difference occurred on one question where the mean of the grid question was significantly higher than the mean for the horizontal scale, but not significantly higher than the mean for the vertical scale (horizontal and vertical did not differ).

Six item-by-item questions that used horizontal or vertical scales were also compared, with only one question leading to a significant mean difference (the use of a horizontal scale led to a significantly higher rating).

Data quality comparisons among the grid, horizontal, and vertical scales were also explored, but no significant differences were found in scale reliability (measured with Chronbach's alpha), in the amount of straight-lining that occurred, in item non-response, or in the resulting factor structure. As a result, a key conclusion is that the question formats studied did not consistently affect the results or associated measures of data quality. The current study did not look at item completion times, because this measure was not available with the software used.[6]

The lack of consistent differences in this study is perhaps not surprising given the test materials, previous inconsistent research, as well as the participants.

Recent research has shown that the number of questions in a grid is important, as is the number of scale points. The grid used in this study consisted of only four questions, so it's less visually demanding and complex than larger grids. Also, a five-point scale was used, and in a recent study that compared matrix and item-by-item questions with 2, 3, 4, 5, 7, 9, and 11 response options, Liu and Cernat (2016) discovered that measurement models revealed measurement equivalence between the two question types when there were fewer than seven response options.

This study also used MTurk participants, an audience that is not likely to be representative of the general American population in terms of their familiarity with surveys and online survey question design. As Toepoel et al. (2009) noted, elderly respondents appear to be more sensitive to verbal, graphical, and numerical language. However, the group participating in this study skewed younger with 78 percent between ages 18 and 40; and was also well educated. Therefore, it's likely that they are more experienced with different formatting approaches, and because of their experience, less likely to be impacted by them.

Another feature, which has not been discussed much in previous research, concerns the nature of the task and its general interest to, or emotional impact on, participants. It seems likely that less controversial topics would be less likely to lead to extreme reactions; therefore, possibly muting the effects of scale differences. This study's topic was not controversial and also possibly of low interest to a part of the MTurk population.

The use of mobile devices is another possible confounding factor, because some software packages such as SurveyMonkey, automatically convert grids to item-by-item displays when a grid gets too large. That did not happen in this study because the grid was small enough to appear unaffected on a mobile device. However, it is not known how many MTurk participants used a mobile device to complete the task, and if that proportion varied among the different groups. In future studies, the use of a mobile device is something that should be controlled, or at least monitored, for its possible effects on data quality (Martinsson et al., 2017).

[6] SurveyMonkey was used.

Attachment 1 – Table Showing Question Formats Used on Different Versions of the Online Instruments

		Question Asked					
				Uses of the Consumer Price Index (CPI)			
	Scale Used	**General Reaction to letter**	**Confidentiality question**	**Social Security**	**Retirement**	**School Lunch**	**Food Stamps**
Version 1	Horizontal						
	Vertical	X	X	X	X	X	X
	Grid						
Version 2	Horizontal			X	X	X	X
	Vertical	X	X				
	Grid						
Version 3	Horizontal	X	X				
	Vertical						
	Grid			X	X	X	X

References

Chan, J. (1991). Response-order Effects in Likert-type Scales. Educational and Psychological Measurement, 51, pp. 531-540.

Christian, L. and Dillman, D. (2004). The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions. Public Opinion Quarterly, Vol. 68, No. 1, Pp. 57-80, April.

Couper, M., Tourangeau, R., Conrad, F., and Zhang, C. (2013). The Design of Grids in Web Surveys. Social Science Computer Review, Jun; 31(3): 322-345.

Diedenhofen, B., & Musch, J. (2016). Cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*, 51–60.

Dillman, D., Smyth, J. & Christian, L. (2009). Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, New York: Wiley.

Friedman, L. and Friedman, H. (1994). A Comparison of Vertical and Horizontal Rating Scales. The Mid-Atlantic Journal of Business, Vol. 30, No. 1, March.

Kaminska, O., McCutcheon, A.L. & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. Public Opinion Quarterly, Vol. 74 (5), 956-984.

Keusch, F. (2012). The Direction of Rating Scales and Its Influence on Response Behavior in Web Surveys. AAPOR Annual Conference, http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2012/01_keusch_E8.pdf

Krosnick, J. (1991). Response Strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology. 5 (3): 213-236, doi:10.1002/acp.2350050305.

Liu, M. and Cernat, A. (2016). Item-by-item Versus Matrix Questions: A Web Survey Experiment. Social Science Computer Review, https://doi.org/10.1177/0894439316674459

Martinsson, J.; Dumitrescu, D.; Markstedt, E. (2017). The Effects on Data Quality of Horizontal and Vertical Question Orientation and Scales of Different Length for Respondents Using Smartphones, Tablets and PCs. General Online Research Conference (GOR).

Menold, N. and Bogner, K. (2016). Design of Rating Scales in Questionnaires. GESIS Survey Guidelines, December, Version 2.0. Active URL: https://www.gesis.org/fileadmin/upload/SDMwiki/MenoldBogner_Design_of_Rating_Scales_in_Questionnaires.pdf

Toepoel, V., Das, M., van Soest, A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales. Journal of Official Statistics, Vol. 25, No. 4, pp. 509-528.

Tourangeau, R.; Couper, M.; and Conrad, F. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, Volume 68, Issue 3, September, Pages 368–393, https://doi.org/10.1093/poq/nfh035.

Wang, R., Hempton, B., Dugan, J., and Komives, S. (2008). Cultural Differences: Why Do Asians Avoid Extreme Responses? Vol. 1, No. 3, http://www.surveypractice.org/index.php/SurveyPractice/article/view/224/html