Web survey experiments on fully balanced, minimally balanced and unbalanced rating scales

Survey Methods: Insights from the Field

Mingnan Liu, SurveyMonkey, Palo Alto, California, U.S.A.
Sarah Cho, SurveyMonkey, Palo Alto, California, U.S.A.

How to cite this article : Liu M. & Cho, S. (2016). Web survey experiments on fully balanced, minimally balanced and unbalanced rating scales, Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=7650

DOI : 10.13094/SMIF-2016-00002

Abstract : When asking attitudinal questions with dichotomous and mutually exclusive response options, the questions can be presented in one of three ways: a full balanced question, a minimally balanced question, and an unbalanced question. Although previous research has compared the fully vs. minimally balanced rating scales, as far as we know, these three types of rating scales have not been tested in a strict experimental setting. In this study, we report two web survey experiments testing these three types of rating scales among 16 different questions. Different from most previous studies, this study used visual display only without any auditory component. Overall, the univariate distributions across these three scale balancing types are very similar to one another. Similar patterns are found when breaking down the analysis by respondent's education level.

Introduction

Writing a survey question may sound easy but there are many factors one needs to consider in order to create a good question. The wording of a question, the number of response options, the order of answer choices, and context of questions, just to name a few, can all have impact on the data quality and the survey estimates (Krosnick & Presser, 2010). Specifically, when constructing questions for survey respondents about their opinions and attitudes regarding competing dichotomous options, survey researchers have a few choices when it comes to wording their question. One approach is to ask a fully balanced question, which presents both sides of the competing viewpoint. Another approach is to ask a minimally balanced question, which only presents one side of the viewpoint and ends the sentence with "or not". The minimal balancing presents both sides of the competing viewpoint, although one side is more explicitly presented than the other. A third approach is to use an unbalanced question that only presents one side of the viewpoint. Researchers have long been aware of the differential measurement errors associated with different survey questioning approaches. As Brace (2008) pointed in his book, an uncareful use of unbalanced wording could result in substantially different answers which in turn lead to biased estimate (see also Bishop, 2004; Saris & Gallhofer, 2007). However, there are very limited empirical studies that have examined this issue in survey experiments. Hedges (1979) was among the first to experimentally test the impact of different question-balancing formats on survey responses in a

face-to-face survey. Hedges' research found that five out of seven comparisons between balanced and unbalanced questions show significant differences. The reason, as suggested by the author, is due to the vague meaning of the implicit option: that is, when not explicitly presented, respondents may interpret the opposite of the presented viewpoint differently. When both viewpoints are fully presented, however, the meanings of both viewpoints allow less room for variability in interpretation. Hence, the univariate distributions differ between the scale formats. Additionally, Hedges found that the impact of question format also interacts with a respondent's demographic background, especially if the topic of the question is related to the demographics. For example, older respondents, when compared to younger respondents, are more resistant to the impact of question format on health-related questions. This is likely to be due to the fact that health issues are more pertinent to older than to younger respondents. Older respondents may have established an opinion toward health-related issues and hence are less likely to vary their answers based on the question-balancing formats. Later, a series of studies on question-balancing formats was reported by Schuman and Presser (1981). Their telephone survey experiments show that the difference between unbalanced and fully balanced scales, as well as the difference between partially balanced and fully balanced scales, is not significant. However, there is some evidence that the responses change when the balancing is achieved through substantive counterarguments, and different counterarguments may lead to different responses. In addition, Schuman and Presser's findings show evidence of the interaction effect between education and scale-balancing format for some questions but not for others.

In a meta-analysis of Schuman and Presser's (1981) studies, Narayan and Krosnick (1996) show that there is a significant difference between unbalanced and fully balanced questions. Further, the effect is similar across respondents with different education levels. However, when comparing fully balanced scales with scales balanced through substantive counterarguments, it was found that less educated respondents are slightly more susceptible to the scale format effect than more educated respondents. The authors argue that the education effect is due to satisficing (Krosnick, 1991). That is, it is more cognitively challenging for lower-educated respondents to form a viewpoint that is not clearly presented in the question than it is for higher-educated respondents. When two viewpoints are substantively described with more words, lower-educated respondents are more inclined to select the viewpoint otherwise not presented in a less balanced approach. This finding is in line with several later studies that also report an interaction effect between question design format and respondents' cognitive level (Knauper, 1999; Knäuper, Belli, Hill, & Herzog, 1997; Knauper, Schwarz, Park, & Fritsch, 2007).

More recently, Shaeffer and her colleague have compared minimally balanced and fully balanced scales in a telephone survey and found no significant difference for univariate distribution (Shaeffer, Krosnick, Langer, & Merkle, 2005). However, they found that the concurrent validity tends to be higher for the minimally balanced than for the fully balanced scale. Lundmark, Gilljam and Dahlberg (2015) tested scale balancing technique (fully versus minimally) and scale length (2-, 7-, and 11-point) through probability-based and non-probability-based web surveys in Sweden. The results show that the measurement is better for minimally balanced scales (7- or 11-point scales) than for fully balanced scales (2-point scales). As can be seen from the previous research, although there are studies that compare two types of scale-balancing formats, the literature on scale-balancing effects is still lacking in some aspects. The current study aims to expand this line of research in two directions. First, this study tests the impact of scale-balancing formats through Web survey experiments. The previous studies mostly used telephone survey experiments (except Lundmark et al., 2015). In a telephone survey, in which information is exchanged aurally, an unbalanced or minimally balanced scale is more likely to channel the focus of the respondents to one side of the viewpoint presented, while overlooking the opinion on the opposite side. In a Web survey, however, questions and response options are presented visually. Even for an unbalanced or minimally balanced scale, the two competing arguments are simultaneously presented as response

options. The meanings of both competing viewpoints are immediately clear to the respondents, regardless of the way the questions are worded. Therefore, we hypothesize that the responses to unbalanced and minimally balanced scales will not differ from responses to the balanced scale. Second, previous studies have only made pairwise comparisons between unbalanced, minimally balanced, or fully balanced scales. In this study, we compare four scale formats using a single factorial design. In addition, we tested two types of minimally balanced scales.

In the following sections, we report findings from two Web survey experiments using nonprobability samples. In both experiments, we test four types of rating scales, including fully balanced, unbalanced, and two minimally balanced scales.

Study 1

Data and measures

The experiment was embedded in a survey on smartphone and tablet usage using a nonprobability sample. The survey was conducted between 22 June and 30 June 2015. The survey invitation was displayed on the SurveyMonkey "Thank you" page – a page displayed to respondents after finishing a user-created survey on SurveyMonkey. Some 87,641 respondents viewed the invitation page, 7545 clicked on it and 4659 completed the survey. There were four experimental conditions. Condition 1 asked fully balanced questions by presenting both sides of the statement. For example, "In general, do you think things in this country are heading in the right direction or heading in the wrong direction?" Condition 2 asked the first variation of minimally balanced questions. For example, "In general, do you think things in this country are heading in the right direction, or not?" Condition 3 asked the second variation of minimally balanced questions. For example, "In general, do you or do you not think things in this country are heading in the right direction?" Condition 4 asked unbalanced questions. For example, "In general, do you think things in this country are heading in the right direction?" There were four questions in each condition (see Appendix A for question wordings). Respondents who clicked on the "Thank you" page invite were randomly assigned to one of the four conditions. There were slightly over 1000 respondents for each condition.

Results

Table 1 presents the distributions for each question under each condition. All questions used binary response options. For all but the "follow the news" question, the p-values are well above the traditional 0.05 threshold, indicating that there are no significant differences between responses across these four conditions. For the "follow the news" question, when both sides of the statement were presented in the fully balanced condition, more respondents admitted that they do not follow the news regularly. For the other three conditions, where either minimally balanced questions or unbalanced questions were asked, the breakdown between yes and no is similar. For the other three questions, including "the country's direction" (p=0.33), "economic situation" (p=0.20), and "good/bad time to buy a house" (p=0.27), no significant differences are observed across the four conditions.

Table 1. Response Distributions For Four Types Of Rating Scales, Experiment 1.

	Fully balanced	Minimally balanced 1	Minimally balanced 2	No balancing		p-value
Right/wrong direction						
Right direction	32%	31%	31%	29%	3.36	0.33
Wrong direction	68%	69%	69%	71%		
Total	1014	1039	1013	1008		
Economy situation						
Getting better	43%	38%	39%	40%	4.63	0.2
Getting worse	57%	62%	61%	60%		
Total	1016	1039	1016	1006		
Buy a house						
Good time	65%	63%	65%	62%	3.89	0.27
Bad time	35%	37%	35%	38%		
Total	1004	1036	1009	1003		
Follow the news						
Yes	83%	87%	88%	87%	9.86	0.02
No	17%	13%	12%	13%		
Total	1018	1039	1018	1012		

Next, we conducted an MANOVA, with the four questions as the dependent variable and the experimental condition indicator as the independent variable. The model (Pillai=0.01, F=1.68, df=3, residual df=4006, p=0.06) is not significant at the 0.05 level, which suggests that the four ways of asking the same questions result in indistinguishable responses. As the literature suggested, the impact of question format can differ by the respondent's education level. To test whether this was the case in this experiment, we broke down the analysis by each respondent's education. Specifically, respondents were classified into one of two groups: educated to a level of more than high school, or educated to a level of high school or less.

Table 2. Response Distributions For Four Types Of Rating Scales By Respondent's Education Level, Experiment 1.

	Low education						High education					
	Fully balanced	Minimally balanced	Minimally balanced	No balancing		p-value	Fully balanced	Minimally balanced	Minimally balanced	No balancing		p-value
Right/wrong direction												
Right direction	36%	26%	33%	30%	4.23	0.24	31%	32%	31%	28%	2.58	0.46
Wrong direction	64%	74%	67%	70%			69%	68%	69%	72%		
Total	175	189	169	192								
Economy situation												
Getting better	41%	33%	43%	39%	4.26	0.23	43%	40%	39%	40%	4.12	0.25
Getting worse	59%	67%	57%	61%			57%	60%	61%	60%		
Total	175	189	168	192								
Buy a house												
Good time	60%	47%	55%	51%	7.59	0.06	66%	67%	68%	65%	1.84	0.61
Bad time	40%	53%	45%	49%			34%	33%	32%	35%		

Total	174	187	168	192								
Follow the news												
Yes	72%	80%	75%	79%	4.33	0.23	86%	89%	90%	89%	7.99	0.05
No	28%	20%	25%	21%			14%	11%	10%	11%		
Total	175	191	168	192								

As Table 2 shows, the patterns of the responses between the two groups are very similar to the combined results in Table 1 and none of the comparisons is statistically significant. This suggests that the respondents' education does not interact with the question format and the answers are similar across the four types of questions for both high- and low-educated respondents.

Study 2

Data and measures

Study 2 was a replication of Study 1 where different questions were used to test the same four different variations of scale-balancing formats. Altogether, there were 12 questions tested within each condition in Study 2. The survey was conducted using SurveyMonkey Audience, an online nonprobability Web panel. The experiment was conducted between 18 June and 22 June 2015, with a total sample size of 2000. Respondents were randomly assigned to one of the four conditions. Due to a programming error, the randomization failed for the first 448 cases and these cases were subsequently removed from the analysis. To compensate for those cases, the same survey experiment was deployed again on SurveyMonkey Audience for 500 further participants. The analysis combines both samples. All but the "sexual orientation" question gave two response options. In the fully balanced condition, the sexual orientation question contained five options – heterosexual, gay, lesbian, bisexual, and transgender. The question was dichotomized into heterosexual versus LGBT in the analysis. In the other three conditions, the sexual orientation question used a yes/no dichotomous response option. *Results* Altogether, Study 2 completed 2061 surveys, with about 500 in each condition. A chi-square test was performed for each question to test whether the responses under the four conditions are similar or different. As Table 3 shows, the p-values for all 12 chi-square tests are all above the 0.05 level, suggesting that responses under the four conditions are similar. A MANOVA test was also conducted, with the 12 questions as the dependent variable and the experimental condition as the independent variable. The test (Pillai=0.03, F=1.36, df=3, residual df=1778, p=0.07) is not significant at the 0.05 level, indicating that no significant difference exists across these four approaches for asking questions.

Table 3. Response Distributions For Four Types Of Rating Scales, Experiment 2.

	Fully balanced	Minimally balanced	Minimally balanced	No balancing		p-value
Terrorism						
Doing enough	41%	44%	45%	45%	2.53	0.47
Not doing enough	59%	56%	55%	55%		
Total	509	486	511	502		
Terrorist attach						
Doing all it can	56%	49%	50%	52%	5.79	0.12
Should do more	44%	51%	50%	48%		

Total	510	483	513	498		
Insurance						
It should	72%	66%	70%	69%	4.19	0.24
It shouldn’t	28%	34%	30%	31%		
Total	487	467	500	473		
Health care						
It should	29%	34%	35%	32%	3.76	0.29
It shouldn’t	71%	66%	65%	68%		
Total	501	478	508	482		
Global warming						
Probably happening	83%	81%	81%	84%	1.57	0.67
Probably not	17%	19%	19%	16%		
Total	521	492	520	498		
Homosexual: speak						
Allowed	96%	94%	95%	96%	2.9	0.41
Not allowed	4%	6%	5%	4%		
Total	519	497	522	503		
Homosexual: teach						
Allowed	96%	94%	93%	96%	7	0.07
Not allowed	4%	6%	7%	4%		
Total	514	495	519	503		
Homosexual: remove book						
Allowed	15%	16%	16%	13%	1.53	0.68
Not allowed	85%	84%	84%	87%		
Total	516	492	519	499		
Smartphone						
I have	75%	75%	76%	77%	0.44	0.93
Don’t have	25%	25%	24%	23%		
Total	518	496	521	510		
Gender						
Male	37%	33%	39%	33%	5.53	0.14
Female	63%	67%	61%	67%		
Total	520	498	524	509		

Sexuality						
LGBT	7%	7%	6%	5%	3.83	0.28
Heterosexual	93%	93%	94%	95%		
Total	503	494	525	507		
Know gay/lesbian						
Know someone	92%	88%	88%	89%	6.65	0.08
Don't know	8%	12%	12%	11%		
Total	522	497	523	508		

Similar to Study 1, in Study 2 we also tested whether the impact of scale format differs by the respondent's education level. As Table 4 shows, for respondents with an education of high school or less, two of the 12 items examined show significant difference across the four conditions. For the question on whether homosexuals should be allowed to teach (=9.06, p=0.03), the percentage of "Allowed" is the highest in the unbalanced condition and lowest in the first minimal balancing condition where the statement ended with "or not". For the question on whether the respondent knows any one who identifies as gay or lesbian (=13.12, p<.01), the percentage of respondents saying they know such people is the lowest in the minimal balancing condition where the statement ends with "or not". The percentage of knowing gay or lesbian people is high in the other minimal balancing and unbalanced conditions. For respondents with more than a high school education, the only significant question was the one asking whether respondents know gay or lesbian people (=10.88, p=.01). In comparison to the lower-educated respondents, a higher percentage of respondents with a higher education stated that they know gay or lesbian people in the fully balanced condition than in the other conditions. Given the small number of items with a significant difference, and the relatively small difference even for the significant questions, we conclude that education level is not a moderating variable for scale balancing.

Table 4. Response Distributions For Four Types Of Rating Scales By Respondent's Education Level, Experiment 2.

	Low education						High education					
	Fully balanced	Minimally balanced	Minimally balanced	No balancing		p-value	Fully balanced	Minimally balanced	Minimally balanced	No balancing		p-value
Terrorism												
Doing enough	40%	37%	36%	38%	0.22	0.97	41%	45%	46%	46%	3.29	0.35
Not doing enough	60%	63%	64%	63%			59%	55%	54%	54%		
Total	47	38	58	48			460	448	452	453		
Terrorist attach												
Doing all it can	52%	37%	40%	34%	3.70	0.30	56%	50%	51%	53%	4.35	0.23
Should do more	48%	63%	60%	66%			44%	50%	49%	47%		
Total	48	38	58	47			460	445	454	450		
Insurance												
It should	69%	59%	60%	67%	1.46	0.69	72%	67%	71%	69%	3.89	0.27
It shouldn't	31%	41%	40%	33%			28%	33%	29%	31%		
Total	42	34	55	43			443	433	445	429		
Health care												
It should	27%	44%	29%	33%	2.94	0.40	30%	33%	35%	32%	3.49	0.32
It shouldn't	73%	56%	71%	67%			70%	67%	65%	68%		
Total	44	34	55	48			455	444	452	433		

Global warming												
Probably happening	83%	64%	66%	78%	6.22	0.10	83%	83%	83%	84%	0.37	0.95
Probably not	17%	36%	34%	22%			17%	17%	17%	16%		
Total	48	39	59	46			471	453	460	451		
Homosexual: speak												
Allowed	89%	76%	86%	92%	5.68	0.13	96%	95%	96%	96%	0.88	0.83
Not allowed	11%	24%	14%	8%			4%	5%	4%	4%		
Total	47	41	58	50			470	456	463	452		
Homosexual: teach												
Allowed	89%	75%	84%	96%	9.06	0.03	96%	95%	94%	96%	3.60	0.31
Not allowed	11%	25%	16%	4%			4%	5%	6%	4%		
Total	47	40	56	50			465	455	462	452		
Homosexual: remove book												
Allowed	24%	27%	20%	12%	3.61	0.31	14%	15%	15%	13%	0.76	0.86
Not allowed	76%	73%	80%	88%			86%	85%	85%	87%		
Total	46	41	56	50			468	451	462	448		
Smartphone												
I have	57%	66%	65%	58%	1.19	0.75	77%	76%	77%	79%	0.96	0.81
Don’t have	43%	34%	35%	42%			23%	24%	23%	21%		
Total	47	41	57	50			469	455	463	459		
Gender												
Male	37%	33%	39%	33%	7.57	0.06	37%	34%	38%	34%	2.89	0.41
Female	41%	22%	42%	24%			63%	66%	62%	66%		
Total	46	41	60	50			472	457	463	458		
Sexuality												
LGBT	9%	3%	3%	6%	2.17	0.54	7%	7%	6%	4%	3.83	0.28
Heterosexual	91%	98%	97%	94%			93%	93%	94%	96%		
Total	47	40	60	50			454	454	464	456		
Know gay/lesbian												
Know someone	77%	60%	88%	86%	13.12	0.00	94%	91%	88%	89%	10.88	0.01
Don’t know	23%	40%	12%	14%			6%	9%	12%	11%		
Total	47	40	58	50			473	457	464	457		

Discussion

Finding the best way of writing survey questions that minimize measurement errors has long been the goal for survey researchers. The balanced scale is one format that many researchers and practitioners follow although no conclusive evidence points to the validity of this question type. Among the limited research done so far, a reduced version of the balanced scale – that is, the minimally balanced scale – produces similar estimates (Shaeffer et al., 2005). In this study, the minimally balanced scales was further simplified to unbalanced scales. Many would argue that presenting only one side of the viewpoint while omitting the other side may mislead the respondents and hence result in biased responses. However, the evidence provided in this study shows the opposite. Two Web survey experiments consistently showed that the response distribution across four types of scales, including fully balanced, minimally balanced, and unbalanced scales, are statistically indistinguishable. Also, respondents’ cognition level, as operationalized by education attainment, does not interact with the balancing format. For respondents with high or low education levels, the different scale formats do not affect the response distributions in any meaningful way. This suggests that an unbalanced scale will not bias the survey result any more than a fully balanced scale can do. The concern over unbalanced scales is more understandable in the telephone survey context where the information is communicated aurally. The unbalanced scale draws a disproportionate amount of a respondent’s attention to the viewpoint presented while largely overlooking the counterpart. In the Web survey, both sides of the viewpoint are visually presented as response options and hence this is less likely to confuse the respondents about the meaning of the question. As

more Web surveys are moving towards mobile Web, the webpage real estate becomes even more valuable: a simplified question wording with equal response distribution is valued more than ever. We hope that future studies will replicate our study and examine whether the findings are replicable in other survey contents and populations. The survey population for this study, either recruited from the Thank you page or the SurveyMonkey Audience, is in principal more motivated and more experienced survey takers than the general population. Replication of this study with participants recruited from other sources will definitely strengthen the generalizability of this study.

Appendix A. Question wordings for experiment 1.

Appendix B. Question wordings for experiment 2.

References

1. Bishop, G. F. (2004). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Rowman & Littlefield Publishers.
2. Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers.
3. Hedges, B. M. (1979). Question Wording Effects: Presenting One or Both Sides of a Case. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *28*(2), 83–99. http://doi.org/10.2307/2987682
4. Knauper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*, 347–370.
5. Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal Of Official Statistics*, *13*(2), 181–199.
6. Knauper, B., Schwarz, N., Park, D., & Fritsch, A. (2007). The perils of interpreting age differences in attitude reports: Question order effects decrease with age. *Journal of Official Statistics*, *23*(4), 515–528.
7. Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.
8. Lundmark, S., Gilljam, M., & Dahlberg, S. (2015). Measuring Generalized Trust An Examination of Question Wording and the Number of Scale Points. *Public Opinion Quarterly*, nfv042. http://doi.org/10.1093/poq/nfv042
9. Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*(1), 58–88.
10. Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research* (Vol. 548). John Wiley & Sons.
11. Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.
12. Shaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the Quality of Data Obtained by Minimally Balanced and Fully Balanced Attitude Questions. *Public Opinion Quarterly*, *69*(3), 417–428. http://doi.org/10.1093/poq/nfi028