Effect of using different labels for the scales in a web survey

Melanie Revilla
Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra

Surveys mainly use questions in which it is allowed to answer only through a closed series of alternatives. The choice of labels for these closed alternatives is an important decision. Depending on this choice, different results can be found. This paper focuses on the impact of using low versus high frequencies or durations scales. The novelty is that it studies panellists of an online panel oriented towards marketing surveys. Also, it uses data from countries little studied before: Spain, Mexico and Colombia. Using a split-ballot experimental design, it shows that significant differences in answers are obtained depending on the scale used. In order to determine which scale gives results closer to the reality, the correlation with an external variable is used; the higher this correlation, the better the scale. In practice, this information can and should be used to select the best scale for a survey.

Introduction

There are two main kinds of question: closed questions, i.e. questions in which it is allowed to answer only through a closed series of alternatives, and open questions, i.e. questions for which the respondents can answer whatever they want and do not have to choose their answer from a list of alternatives. The advantages and disadvantages of both open and closed response formats have been discussed already: in 1944 by Lazarsfeld and, later, by many others (e.g. Schuman & Presser 1981; Converse 1984; Krosnick & Schuman 1988; Schwarz & Hippler 1991). In practice, the open question format is more complicated to use for quantitative research since questions need first to be coded in such a way that the answers can be analysed. This is time consuming and has a cost. Therefore, most survey research uses closed questions.

Received (in revised form): 26 September 2013

DOI: 10.2501/IJMR-2014-028

Using closed questions means that the researcher has to take decisions not only about the formulation of the question itself but also about the scale. Decisions about the scale include decisions about the number of response categories, about the presence of a middle point, of a 'don't know' option, about the use and choice of labels, etc.

These decisions are crucial because the 'response scales are not only passive "measurement devices" that respondents use to report their behaviours. Rather, response scales may also serve as a source of information for the respondent. Specifically, respondents may consider the range of behaviours described in the response alternatives to reflect the researcher's knowledge of, or expectations about, the distribution of these behaviours in the "real world"' (Schwarz & Hippler 1987, p. 164).

Respondents do not already have in memory an opinion about everything they are asked about in a survey (Converse 1964; Zaller 1992; Tourangeau *et al.* 2000). For most of the questions, they form their opinion in the moment using all the available information at that instant. The scales are part of this information. They give a hint to the respondents about which ranges of answers are acceptable. Schwarz and Hippler's results (1987) suggest that, for behavioural frequency reports, respondents assume that the middle of the scale corresponds to the 'normal' or most common behaviour. Then, respondents position themselves according to that. If they think they do more than the average, they select higher answer categories. If they think they do less, they select lower answer categories. In brief, 'respondents use the range of the response alternatives as a frame of reference in estimating their own behavioural frequencies and report higher frequencies in scales that present high rather than low frequency response alternatives' (Schwarz & Hippler 1991, p. 48).

Focusing on sensitive questions, Tourangeau and Smith (1996) also find support for this phenomenon. Their results show that the reported number of sexual partners is more than twice as high when the answer category labels are shifted to indicate higher numbers (from 0, 1, 2, 3, 4, 5 or more to 0, 1–4, 5–9, 10–49, 50–99, 100 or more). In this case, we can assume that most respondents know how many sex partners they have had. However, they use the information from the answer category labels to determine what seem to be the desirable answers. If they are out of this range, they can decide to lie in order to present themselves in what they think is a more positive way (based on the information they have extracted from the scale).

Following this line of research, we hypothesise that respondents are using the labels as references in order to choose the option category for their answer.

When the labels at the end point are higher, more of the corresponding behaviours will be reported. Our goal is to test if this hypothesis holds in different frames than those studied so far in the literature.

First, we want to consider a different mode of data collection: the internet. Tourangeau and Smith (1996) find a significant interaction between the mode and the response format. However, they do not consider the web. Because this is a self-completed mode, we expect the social desirability bias that leads respondents to under-report 'undesirable' behaviours and over-report 'desirable' behaviours to be lower (Kreuter *et al.* 2008). Thus, we assume that the effect of switching the labels of the scale will be lower in internet mode. Indeed, if you had 20 sexual partners and the scale goes up to only '5 or more', you will feel that 20 is 'too much'. But if the social desirability bias is lower, you will tend to lie and under-report it less.

Second, we want to consider countries that have not been studied with respect to this before: Spain, Mexico and Colombia. Cultural differences have been shown to exist across countries with respect to the impact of scale characteristics on different measures of quality (see e.g. Saris & Gallhofer 2007). However, very little is known about countries from central and Latin America.

The next section presents the split-ballot experiments performed and analysed in order to test the impact of the choice of the labels. Then, the results about the impact of the different scales on the distributions and means are presented. Nevertheless, more reported behaviours do not mean better scales. It is also possible that people report more than what they are doing, mainly if it is socially desirable to do so. Therefore, in order to know which scale is performing better, finally we perform an external validity test.

Experiments

The data come from a survey completed between 29 January and 24 April 2013 by 6,000 Netquest[1] panellists in Spain, Mexico and Colombia (2,000 in each of the three countries; quotas for age and gender used). Netquest is one of the online survey companies with the biggest panels in these countries. The survey of interest is a repetition of the core modules of the fourth round of the European Social Survey (adapted to be completed online).

This survey includes several split-ballot experiments. For each experiment, the respondents within each country were randomly assigned to three different groups. Each group counts around 660 respondents. Each group

[1] Information at: www.netquest.com.

got the same questions but with a different scale (also called 'method'). Group 1 got scale 1 (M_1), group 2 got scale 2 (M_2) and group 3 got scale 3 (M_3). Since the assignment is random, we expect significant differences across groups to be due to the effect of answering with one or another scale, and not to substantive differences. Table 1 summarises for the two experiments studied, the topic, the names of the questions, the main differences across the scales and the text of the questions.[2]

As Table 1 indicates, the first experiment is about political participation, and includes seven items about different kinds of possible political participation. In split-ballot groups 1 and 3, the seven items are presented in a battery and, for each item, the respondents are asked to select either 'yes' or 'no' (group 1) or 'no', 'yes, one time', 'yes, more than one time' (group 3). In split-ballot group 2, a check-all-that-apply format is used using the same seven items ('in the past 12 months, did you do any of the following activities? Please, check all the ones you did in the list below').

Our general hypothesis applied to the political participation experiment means that we expect:

> **H1:** proportion of respondents reporting they did the activity in M_2 < the one in M_1 < the one in M_3.

Indeed, by adding the category 'yes, more than once' in M_3, the scale may suggest to the respondents that the more 'normal' behaviour is the central category, 'yes, once'. Moreover, by proposing two positive categories versus one negative, the scale may suggest that the researcher expects most respondents to answer 'yes' (once or more than once, but 'yes').

The difference between M_1 and M_2 is of a different nature. We should refer to another line of research to justify our hypothesis in that case. The difference here is between a check-all-that-apply format (M_2) and a forced-choice format (M_1 and M_3).

Often, these two formats are used in surveys as if they would be equivalent. However, they are quite different. In the check-all-that-apply format, respondents may tend to process only part of the items. When they already have selected, for instance, three out of ten items, they can feel that they did their job and do not make the effort to process carefully the items left to be sure to check all those that really apply. This is one form of weak satisficing (Krosnick 1991).

[2] The complete questionnaire is available at http://test.nicequest.com/surveys/global_glacier/eb5e4c34-e56e-4f1c-be7d-7354febeb01f.

Table 1 The two split ballot experiments

Topic (question names)	Methods for the different groups	First question – others
Political participation (B13–B19)	M_1 = yes/no M_2 = check-all-that-apply M_3 = no/yes one time/yes more than once	*There are different ways of trying to improve things in [country] or help prevent things from going wrong. During the last 12 months, have you* B13 – … *contacted a politician?* B14 – …*worked in a political party or action group?* B15 – …*worked in another organisation or association?* B16 – …*worn or displayed a campaign badge/sticker?* B17 – …*signed a petition?* B18 – …*taken part in a lawful public demonstration?* B19 – …*boycotted certain products?*
Media use (A1–A6)	M_1 = from 0 to >3h – 8 categories M_2 = from 0 to >6h – 14 categories M_3 = from 0 to >7h – 9 categories	A1 – *How much time, in total, do you spend watching television?* A2 – *And on an average weekday, how much of your time watching television is spent watching news or programmes about politics and current affairs?* Idem with radio (A3 and A4) and newspapers (A5 and A6)

In contrast, in the forced-choice format, respondents have to give an answer to each of the items. They are forced to answer the items one by one. Nevertheless, we cannot force the respondents to process the items carefully and give the proper answer. Besides, it has been shown that there is a general tendency of human beings to say 'yes' or 'I agree' (Berg & Rapaport 1954). This tendency is referred to as 'yes-saying' or 'acquiescence bias' (Krosnick 1991). This is another form of satisficing. Therefore, the forced-choice format using a yes/no scale may lead to a higher proportion of 'yes' than the reality.

There is some empirical evidence (Smyth *et al*. 2006) showing that the proportion of 'yes' is indeed much higher when using a forced-choice format compared with a check-all-that-apply format. Smyth *et al*. (2006) also show that the results obtained are closer to the reality when using the forced-choice format. Therefore, it seems that the main problem comes from the fact that respondents do not process all the items in the check-all-that-apply format.

Concerning the media experiment, it includes six items. As indicated in

Table 1, they are about the time spent watching television, listening to the radio and reading the newspapers. The respondents are asked about the total time they spent on these three media, and about which part of that time is dedicated to politics and current affairs programmes. Different scales are used: in group 1 the scale contains eight categories and goes up to three hours or more. In group 2, it contains 14 categories and goes up to six hours or more. In group 3, it contains nine categories and goes up to seven hours or more.

Our general hypothesis applied to the media use experiment means that we expect:

> **H2:** proportion of respondents reporting more than three hours in M_1 < the one in M_2 < the one in M_3.

Indeed, M_1 is the scale with the lowest value as label of the highest response category, then comes M_2 and finally M_3. The number of response categories varies also between the different scales, but we expect that the most important difference will come from the choice of the labels, suggesting that different durations are normal and acceptable.

Reported activities

First, we look at the distributions and means of the different variables of interest, in order to see if changing the format of the scale has the expected impact on the answers.

Political participation experiment

Table 2 presents the percentages of respondents reporting that they did each activity when the different scales are used and for the three countries. For M_2, this corresponds to the percentage of respondents that select the corresponding item in the check-all-that-apply scale. For M_3, it is the sum of the percentages of respondents answering 'yes, once' and the ones answering 'yes, more than once'. Table 2 also reports the average number of items out of the seven in the battery that the respondents reported they did (last row: 'avg').

Table 2 shows that the percentages of respondents reporting they did the different activities are always lower when M_2 is used, for all activities and in the three countries. Most of the time, these differences in proportions are statistically significant. Using a check-all-that-apply format clearly elicits less reporting of the political activities. Results from previous studies

Table 2 Percentages of respondents reporting they did the different activities and average number of reported activities

	Mexico			Colombia			Spain		
% yes	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
B13	28.66	22.91*	36.23*	26.88	25.90	37.97*	17.54	7.96*	18.15
B14	20.38	15.71*	29.05*	20.78	17.51	30.84*	13.34	8.11*	17.39*
B15	33.83	27.31*	42.35*	33.59	27.25*	44.67*	28.79	15.62*	35.23*
B16	19.50	17.91	25.23*	20.00	13.17*	27.23*	14.84	11.26*	19.94*
B17	36.78	30.54*	40.21	49.84	43.11*	58.94*	60.72	50.45*	59.97
B18	16.40	12.33*	22.94*	24.22	18.26*	30.40*	40.18	37.39	46.03*
B19	32.05	25.99*	33.79	30.63	25.15*	34.15	34.78	29.58*	41.23*
Average	1.88	1.53*	2.30*	2.06	1.70*	2.64*	2.10	1.60*	2.38*

Notes: M_1 = yes/no; M_2 = check-all-that-apply; M_3 = no, yes-once, yes-more than once; the stars in the column of M_2 (resp. M_3) indicate when the difference in proportions or means between M_1 and M_2 (resp. M_1 and M_3) is significant: * means at the 5% level, ** means at the 10% level

appear to be confirmed for online panellist respondents in Spain, Mexico and Colombia.

With respect to the difference between the two forced-choice formats, except for the question B17 in Spain, the percentage of respondents answering 'yes' is higher for M_3 than for M_1 and, in most cases, significantly higher at the 5% level. Therefore, when providing a category with a higher frequency label, respondents tend to report that they did more activities. This is in line with the mechanisms proposed by Schwarz and Hippler (1987, 1991).

The trends are confirmed when looking at the average number of reported activities out of the seven in the battery. This average number is always lower for the check-all-that-apply format, then for the 'yes/no' format, and finally for the 'no/yes once/yes more than once'. Differences are all statistically significant.

Overall, the results are quite clear: H1 holds in all three countries.

Media use experiment

For the second experiment, in order to be able to compare the answers in the three scales, we recode all the variables in the following categories: nothing (0), less than 1h (1), from 1h to 2h (2), from 2h to 3h (3), more than 3h (4). Table 3 gives the percentages of respondents that reported spending more than three hours on the different media, and the means for the recoded variables.

Table 3 Percentages of respondents reporting they spend more than 3 hours on the different media and means for the recoded variables going from 0 to 4 (excluding the few missing)

	Mexico			Colombia			Spain		
	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
% report more 3h									
A1	9.01	16.15*	15.44	13.13	24.70*	20.89**	17.09	27.03*	22.19*
A2	1.27	3.12*	2.72	2.16	3.15	4.27	1.07	3.42*	2.31
A3	14.33	20.56*	18.20	18.44	20.36	21.90	12.89	12.91	14.24
A4	2.99	5.90*	4.93	7.03	6.96	9.20	3.02	3.80	4.55
A5	0.44	0.59	1.22	0.47	0.60	1.30	0.15	0	0.75*
A6	0.67	0.81	1.34	0.42	0.61	1.78**	0.44	0	0.23
Means for the recoded variables (going from 0 to 4)									
A1	1.91	2.16*	2.29*	2.12	2.37*	2.40	2.37	2.55*	2.47
A2	1.12	1.24*	1.36*	1.28	1.37*	1.52*	1.30	1.40*	1.41
A3	1.67	1.82*	1.85	1.68	1.79	1.89	1.49	1.52	1.61
A4	1.18	1.31*	1.36	1.33	1.40	1.51	1.06	1.07	1.11
A5	0.74	0.82*	0.90**	0.81	0.85	0.94*	0.74	0.75	0.77
A6	1.01	1.04	1.17*	1.02	1.05	1.19*	1.02	1.01	1.05

Notes: M_1 = 8 categories till >3h by half hours; M_2 = 14 categories till >6h by half hours; M_3 = 9 categories till >7h by hours; the stars in the column of M_2 (resp. M_3) indicate when the difference in proportion or means between M_1 and M_2 (resp. M_2 and M_3) is significant: * means at the 5% level, ** means at the 10% level

According to Table 3, there are significantly higher percentages of respondents reporting they spend more than three hours watching television ('A1') and watching programmes about politics on television ('A2') when the answer categories go up to six or seven hours (M_2 and M_3). There are also higher percentages of respondents reporting they spend more than three hours listening to the radio ('A3') and listening to programmes about politics ('A4') in M_2 and M_3. However, the differences are statistically significant for Mexico, but not for Colombia and Spain.

This may be because watching a lot of television is seen more negatively than listening a lot to the radio. If the time spent watching the television is a more sensitive question, then, respondents may react more to the changes of scales. Presented with a sensitive question, they feel more the need to use all the information from the scale to decide about their answer. When the time proposed in the label of the end point of the scale is higher, they may deduce that it is socially acceptable to spend more hours watching television every day and report more easily longer durations of television watching.

For the time spent reading newspapers ('A4'), and reading about politics in the newspapers ('A5'), very few respondents report that they spend more than three hours in all three scales, which seems quite normal. However, there are somehow more respondents reporting they spend on average more than three hours when M_3 is used.

Looking at the means for the recoded variables (going from 0 to 4), we see that the mean when M_1 is used is always lower than the mean using M_2, which is always lower than the mean using M_3, except in the case of A1 in Spain. The values of the means do not mean much here since we use the code of the variables and not directly the durations. But our interest is not in the values themselves; it is in the order. Here, the order suggests that, when using a high duration scale, people tend to report more time spent on average on these activities. The differences are almost always significant for Mexico. In Colombia, they are not significant for radio. In Spain, they are not significant for radio and for newspapers. It seems that the impact of using high versus low duration scales is more important in Mexico than in Colombia, and more in Colombia than in Spain.

External validity test

Differences in distributions are quite clear. For the political participation, the check-all-that-apply format elicits much less reporting of the different activities than the forced-choice format 'yes/no' – besides, the 'yes/no' format elicits less reporting than the 'no/yes once/yes more than once' format. For the media experiment, the main difference appears to be between, on the one hand, the first scale (going up to more than three hours) and, on the other hand, the two others (going up to more than six or more than seven hours). It is in Mexico that this difference is the clearest.

But which of the previous results is the closest to the reality? In the political participation experiment, it may be that respondents do not report enough activities in the check-all-that-apply format, because they do not go through all the items carefully. But it can also be that they report too much in the closed format, because of acquiescence bias. A scale that leads to fewer reported activities is not necessarily a scale whose quality is worse.

What we would like to determine is: what is the scale that allows us to get results closer to the true values? What is the scale that performs better to reproduce the reality?

To answer these questions, we could use external measures that would give us objective information about the true values. Nevertheless, we did

not have access to such data. Therefore, we use another approach to try to determine which scale is performing better, meaning giving the closest results to the reality. We perform an external validity test.

The external validity of a scale may be quantified by looking at the correlation between the answers obtained when using this scale and other variables that are known to correlate with the variable of interest. The higher this correlation, the better the external validity of the scale. The variable used as correlate should be measured using the same scale in the different split-ballot groups such that what is changing the correlation is the difference in scale for the variables of our experiments.

The variable selected for the external validity test is as follows:

> **B1:** How interested would you say you are in politics? Are you … very interested (1), quite interested (2), hardly interested (3), or not at all interested (4)?

This question is a measure of political interest. It should be correlated with political participation, measured by the number of reported actions in the battery of the political participation experiment (referred to as 'pp' in Table 4). We expect that, the higher the political interest of the respondent, the higher the number of political actions the respondent has done. Therefore, the scale that leads to the highest correlation (in absolute value)[3] is the one we will consider better.

For the overall time spent on the different media (television, radio, newspapers), we did not find a suitable variable to test for external validity. However, we also expect political interest to be correlated positively with the time spent on the different media dedicated to news or programmes about politics and current affairs (questions A2, A4, A6). For these three variables, we could therefore test external validity, too.

Table 4 gives the Pearson correlations[4] (absolute values) of the variables of the two experiments with political interest in the three countries and for the different scales.

For the political participation experiment, the lowest correlations are found for the check-all-that-apply scale. This is in line with previous

[3] Otherwise, the sign of the correlation is expected to be negative because the scales are going from high to low for B1 and from low to high for the total number of political activities reported.

[4] In line with results, for instance, from Labovitz (1967, 1970) or Borgatta and Bohrnstedt (1980), we believe that Pearson correlations make sense for these analyses even if the scales have a relatively low number of response categories in some cases. However, we also computed the Spearman's rho to see if results would change by considering the variables as ordinal ones. As can be seen in Appendix 1, the results are very similar and all the main conclusions remain true.

Table 4 Pearson correlations total number of reported activities and media use for political affairs with political interest (absolute values)

		Mexico			Colombia			Spain		
Experiment	**Correlation**	**M_1**	**M_2**	**M_3**	**M_1**	**M_2**	**M_3**	**M_1**	**M_2**	**M_3**
Political participation	B1–pp	0.42	0.35	0.38	0.37	0.37	0.38	0.37	0.31	0.37
Media	B1–A2	0.24	0.27	0.36**	0.24	0.19	0.25	0.26	0.25	0.37*
	B1–A4	0.22	0.21	0.29	0.25	0.21	0.30**	0.21	0.26	0.21
	B1–A6	0.13	0.23	0.25	0.11	0.18	0.19	0.22	0.21	0.33**

Notes: for the political participation experiment, the stars in the column M_2 (resp. M_3) indicate when the difference in correlations between M_1 and M_2 (resp. M_1 and M_3) is significant; for the media use experiment, the stars in the column M_2 (resp. M_3) indicate when the difference in correlations between M_1 and M_2 (resp. M_2 and M_3) is significant; * means at the 95% level, ** means at the 90% level

literature about check-all-that-apply versus forced-choice scales. The correlations for the two forced-choice scales are similar in Colombia and Spain. Adding a category 'yes, more than once' impacts the distributions of the answers but not the correlations with the external variable B1. In Mexico, the correlation for M_1 is a bit higher than the one for M_3. This suggests that the scale 'yes/no' is the one performing better or, at least, since the difference is not statistically significant, that it performs as well as the scale with an extra category.

For the media use experiment, the highest correlations with political interest are always found for M_3, except in Spain for A4. This suggests that, in general, this is the scale that works better and should be preferred. However, the improvement is not always significant compared to M_2. But the correlations are multiplied up to a factor of two compared to M_1. For instance, in Mexico for A6, it goes from 0.13 to 0.25.

Conclusion

First, when labels indicating higher frequencies or durations are used, this increases the proportions of respondents reporting higher frequencies or times spent on the corresponding activities. This supports our hypotheses, and is in line with Schwarz and Hippler (1987, 1991).

Nevertheless, higher reports do not necessarily mean the results are closer to the reality. When looking at the external validity test, there is no clear evidence in the political participation experiment that adding the category 'yes, more than once' improves the quality. In the media use experiment, the scale going up to 'more than six hours' does not seem

of significantly better quality than the one going up to 'more than three hours', but the one going up to 'more than seven hours' and increasing hour by hour (instead of half an hour by half an hour) leads in general to higher correlations with the external variable measuring political interest.

Besides, the results for the political participation experiment show that using a check-all-that-apply format versus a forced-choice format leads to a significantly lower proportion of selected items. In this case, the test of external validity also indicates that the check-all-that-apply format is of lower quality. It seems that the respondents are not considering all the items carefully enough when answering a check-all-that-apply scale. This is in line with previous research on that topic (Smyth *et al*. 2006) and holds for panellist respondents of online surveys. Also, it holds in different countries with different cultures and languages from those that the previous research had studied. However, differences across countries exist in terms of the size and significance of the effects. We observe in this study that, for several variables, the differences across scales are stronger in Mexico than in Colombia and Spain.

Overall, the study suggests therefore that practitioners should avoid the use of the check-all-that-apply format as often as possible. Instead, it would be better to use a forced-choice format. Within the forced-choice formats, if the topic is central to the mind of the respondent and sensitive, they should provide answer categories with high enough labels such that respondents do not feel that their behaviour is not normal and do not tend to under-report socially undesirable behaviours. On the other hand, if the topic is not very central and not sensitive, they should use labels following the expected population distribution such that respondents can use the middle of the scale as a reference point as to what is the norm, and evaluate their own behaviour as lower (higher) than the average. In any case, practitioners should be aware that the choices they are making about the labels of the response categories may influence their results.

We should however be careful not to generalise about the results: we studied only two experiments so more work would be needed to confirm what we suggested in the previous paragraph. This is one of the main limits to our conclusions. We should mention at least two more. First, the number of answer categories varied from one scale to another. Therefore, not only the labels were different but also the numbers of response options. This means that the effect found can come from differences in labels or number of scale points. However, the variations in numbers of response categories are small, except for M_2 in the media experiment, such that we do not expect a large impact of these variations. Besides, the results are

more in line with an explanation in terms of different labels: indeed, the means for M_2, for example, are almost always in between the ones for M_1 and for M_3 (see Table 3, mean of the recoded variables). This is what we expect based on the labels but not based on the number of categories. Nevertheless, an improved design for future research could be one where only the labels are changed but the numbers of categories are kept equal. Second, the tests of external validity were quite weak, because we did not have many adequate variables to correlate with. Also, the external validity test is not a very strong or precise indicator of quality. In order to get a more precise estimation of the quality of the different scales, repetitions of the same question for the same respondents but using different scales would be needed. This is what Campbell and Fiske had already proposed as a multitrait-multimethod design in 1959. With such an approach, not only external validity, but also measurement validity and reliability, can be estimated, which gives much more information to the researchers about which scale it is better to use. An important line of research is working in this direction (see for instance Saris & Gallhofer 2007). However, it has concentrated on specific geographical areas and data-collection modes. It would be interesting to extend it to new countries like the ones studied here, and to online panellists.

Appendix 1: Spearman rho total number of reported activities and media use for political affairs with political interest (absolute values)

		Mexico			Colombia			Spain		
Experiment	Correlation	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
Political participation	B1–pp	0.42	0.36	0.39	0.39	0.38	0.39	0.37	0.29	0.35
Media	B1–A2	0.30	0.29	0.37	0.28	0.21	0.24	0.28	0.24	0.38
	B1–A4	0.24	0.23	0.30	0.27	0.23	0.33	0.25	0.29	0.24
	B1–A6	0.23	0.23	0.29	0.15	0.19	0.21	0.24	0.22	0.34

References

Berg, I.A. & Rapaport, G.M. (1954) Response bias in an unstructured questionnaire. *Journal of Psychology*, **38**, pp. 475–481.

Borgatta, E. & Bohrnstedt, G. (1980) Level of measurement once over again. *Sociological Methods & Research*, **9**, pp. 147–160.

Campbell, D.T. & Fiske, D.W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **6**, pp. 81–105.

Converse, J.M. (1984) Strong arguments and weak evidence: the open/closed questioning controversy of the 1940s. *Public Opinion Quarterly*, **48**, 1B, pp. 267–282.
Converse, P. (1964) The nature of belief systems in mass publics, in Apter, D.A. (ed.) *Ideology and Discontent*. New York: Free Press, pp. 206–261.
Kreuter, F., Presser, S. & Tourangeau, R. (2008) Social desirability bias in CATI, IVR and web surveys: the effects of mode and question sensitivity. *Public Opinion Quarterly*, **72**, 5, pp. 847–865.
Krosnick, J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, **5**, pp. 213–236.
Krosnik, J.A. & Schuman, H. (1988) Attitude intensity, importance and certainty and susceptibility to response effects. *Journal of Personality and Social Psychology*, **54**, pp. 940–952.
Labovitz, S. (1967) Some observations on measurement and statistics. *Social Forces*, **46**, pp. 151–160.
Labovitz, S. (1970) The assignment of numbers to rank order categories. *American Sociological Review*, **35**, 3, pp. 515–524.
Lazarsfeld, P.F. (1944) The controversy over detailed interviews – an offer for negotiation. *Public Opinion Quarterly*, **8**, 1, pp. 38–60.
Saris, W.E. & Gallhofer, I. (2007) *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley.
Schuman, H. & Presser, S. (1981) *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York: Academic Press.
Schwarz, N. & Hippler, H.J. (1987) What response scales may tell your respondents: informative functions of response alternatives, in Hippler, H., Schwarz, N. & Sudman, S. (eds) *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.
Schwarz, N. & Hippler, H.J. (1991) Response alternatives: the impact of their choice and order, in Biemer, P., Groves, R.M., Mathiowetz, N.A. & Sudman, S. (eds) *Measurement Errors in Surveys*. Chichester: Wiley, pp. 41–56.
Smyth, J.D., Dillman, D.A., Christian, L.M. & Stern, M.J. (2006) Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly,* **70**, 1, pp. 66–77.
Tourangeau, R. & Smith, T.W. (1996) Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, **60**, 2, pp. 275–304.
Tourangeau, R., Rips, L.J. & Rasinski, K. (2000) *The Psychology of Survey Response*. Cambridge, MA: Cambridge University Press.
Zaller, J.R. (1992) *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

About the author

Melanie Revilla is a postdoctoral researcher at the Research and Expertise Centre for Survey Methodology (RECSM) and an associate professor at Universitat Pompeu Fabra (UPF, Barcelona, Spain). Her research interests cover most aspects of survey methodology. She is particularly interested in everything related to questionnaire design, prevention and correction for measurement errors, and the effects of the mode of data collection.

Address correspondence to: Dr Melanie Revilla, RECSM, Universitat Pompeu Fabra, Edifici ESCI- Born- Office 19.523, Passeig Pujades 1, 08003 Barcelona, Spain

Email: melanie.revilla@upf.edu