

RECSM Working Paper Number 30

2013

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

Melanie Revilla

Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra

Edifici ESCI-Born- Office 19.523

Passeig Pujades, 1

08003 Barcelona

melanie.revilla@upf.edu

Abstract

Surveys mainly use questions in which it is allowed to answer only through a closed series of alternatives. The choice of the 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 panelists of an online panel oriented toward 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 give results closer to the reality, the correlation with an external variable is used: the higher this correlation is, the better the scale is. In practice, this information can and should be used to select the better scale for a survey.

Introduction

There are two main kinds of questions: 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 in a list of alternatives. The advantages and disadvantages of both open and closed response format have been discussed already in 1944 by Lazarsfeld and later by many others (e.g. Schuman and Presser 1981; Converse 1984; Krosnick and Schuman 1988; Schwarz and 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 analyzed. This is time consuming and has a cost. Therefore, most survey research uses closed questions.

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 behaviors. Rather, response scales may also serve as a source of information for the respondent. Specifically, respondents may consider the range of behaviors described in the response alternatives to reflect the researcher’s knowledge of, or expectations about, the distribution of these behaviors in the ‘real world’.” (Schwarz and Hippler 1987, pp. 164).

Respondents do not already have in memory an opinion about everything they are asked about in a survey (Converse 1964; Zaller 1992; Tourangeau, Rips, and Rasinski 2000). For most of the questions, they form their opinion in the moment using all the available information at this 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 behavioral frequency reports, respondents assume that the middle of the scale correspond to the “normal” or most common behavior. 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 behavioral frequencies and report higher frequencies in scales that present high rather than low frequency response alternatives” (Schwarz and Hippler 1991, pp.48).

Focusing on sensitive questions, Tourangeau and Smith (1996) also find support for this phenomenon. Their results show that the reported number of sex partners is more than twice higher when the answer categories 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 that case, we can assume that most respondents know how many sex partners they have had. However, they use the information from the answer categories 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 extracted from the scale).

Following this line of research, we hypothesize 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 behaviors will be reported. Our goal is to test if this hypothesis holds in different frames than the ones 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 Web. Because this is a self-completed mode, we expect the social desirability bias that leads respondents to under report “undesirable” behaviors and over report “desirable” behaviors to be lower (Kreuter, Presser and Tourangeau 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 sex partners and the scale goes

only till “5 or more”, you will feel that 20 is “too much”. But if the social desirability bias is lower, you will tend less to lie and under-report it.

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 and Gallhofer 2007). However, very little is known about countries from central and Latin America.

The next section presents the split-ballot experiments performed and analyzed 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 behaviors do not mean better scale. It is also possible that people report more than what they are doing, mainly if it is socially desirable to report so. Therefore, in order to know which scale is performing better, finally, we perform an external validity test.

Experiments

The data comes from a survey completed between the 29th of January and the 24th of April 2013 by 6.000 Netquest¹ panelists 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 4th 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 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 summarizes for the two experiments studied, the topic, the names of the questions, the main differences across the scales and the text of the the questions².

Table 1 here

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”,

¹ Information at: www.netquest.com

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

“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 $M_2 < \text{the one in } M_1 < \text{the one in } M_3$

Indeed, by adding the category “yes, more than once” in M_3 , the scale may suggest to the respondents that the more “normal” behavior is the central category, “yes, once”. Moreover, by proposing two positive categories versus one negative, the scale may suggest that the researcher expect most respondents will 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 3 out of 10 items, they can feel that they did their job and do not make the effort of processing carefully the items left to be sure to check really all the ones that apply. This is one form of weak satisficing (Krosnick 1991).

By opposition, 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 and 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, Dillman, Christian and Stern 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 th 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 programs. Different scales are used: in group 1 the scale contains 8 categories and goes till 3 hours or more. In group 2, it contains 14 categories and goes till 6 hours or more. In group 3, it contains 9 categories and goes till 7 hours or more.

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

H2: proportion of respondents reporting more than 3 hours in $M_1 <$ the one in $M_2 <$ the one in M_3 .

Indeed, M_1 is the scale with the lowest duration per day as labels, 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 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 here

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 appear to be confirmed for online panelists’ 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: hypothesis *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 percentage of respondents that reported spending more than 3 hours on the different media and the means for the recoded variables.

Table 3 here

According to Table 3, there are significantly higher percentages of respondents reporting they spend more than three hours watching the television (“A1”) and watching programs about politics on the television (“A2”) when the answer categories go up to 6 or 7 hours (M_2 and M_3). There are also higher percentages of respondents reporting they spent more than three hours listening to the radio (“A3”) and listening to programs about politics (“A4”) in M_2 and M_3 . However, the differences are statistically significant for Mexico, but not 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. In front of a sensitive question, they feel more the need to use all the information from the scale to decide about their answers. 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 everyday and report more easily longer durations of television watching.

For the time spent reading newspapers (“A4”) and readings about politics in the newspapers (“A5”), very few respondents report that they spend more than 3 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 the order. Here, the order suggests that when using a high duration scale, people tend to report more time spent in average on these activities. The differences are almost always significant for Mexico. In Colombia, they are not significant for the radio. In Spain, they are not significant for the

radio and for the 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 between on the one hand the first scale (going till more than 3hours) and on the other hand the two others (going till more than 6 or more than 7 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 less reported activities is not necessarily a scale whose quality is worse.

What we would like to determine is: what is the scale that allows getting 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 is, 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 the following:

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 as “pp” in Table 4). We expect that the higher the political interest of the respondent is, the higher the number of political actions the

respondent has done is. Therefore, the scale that leads to the highest correlation (in absolute value³) is the one we will consider the 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 programs 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 correlations⁴ (absolute values) of the variables of the two experiments with political interest in the three countries and for the different scales.

Table 4 here

For the political participation experiment, the lowest correlations are found for the check-all-that-apply scale. This is in line with previous 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 M₁ is a bit higher than the one for M₃. It 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 M₃, except in Spain for A4. It suggests that in general this is the scale that works better and should be preferred. However, the improvement is not always significant compared to M₂. But the correlations are multiplied up to a factor two compared to M₁. For instance, in Mexico for A6, it goes from .13 to .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 Hipple (1987, 1991).

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 relatively low number of response categories in some cases. However, we also computed the Spearman 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.

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 until “more than 6 hours” does not seem of significantly better quality than the one going until “more than 3 hours”, but the one going until “more than 7 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 that 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, Dillman, Christian and Stern 2006) and holds for panelists respondents of online surveys. Also, it holds in different countries with different cultures and languages from what previous research had studied. However, differences across countries exist in the size and the 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 check-all-that-apply format as often as possible. Instead, they should better 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 behavior is not normal and do not tend to under-report social undesirable behaviors. 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 of what is the norm and evaluate their own behavior 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 about the generalization of the results: we only studied 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 (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 proposed as a multitrait-multimethod design already 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 and 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 panelists.

References

Berg, I.A. and G.M. Rapaport (1954). "Response bias in an unstructured questionnaire". *The Journal of Psychology*. Taylor & Francis

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

Campbell, D. T. and D.W. Fiske (1959). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 6:81-105

Converse P. (1964). *The nature of belief systems in mass publics*. In D. A. Apter (ed.), *Ideology and Discontent*, New York: Free Press, 206-261.

Converse, J.M. (1984). "Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s". *Public Opinion Quarterly*. 48(1B): 267–282. DOI: 10.1093/poq/48.1B.267

Kreuter, F., Presser, S., and R. Tourangeau (2008). "Social Desirability Bias in CATI, IVR and Web Surveys: The Effects of Mode and Question Sensitivity". *Public Opinion Quarterly*, 72(5): 847–865. DOI: 10.1093/poq/nfn063

Krosnick, J.A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". *Applied Cognitive Psychology*, 5, 213-236.

Krosnik J. A., and H. Schuman (1988). "Attitude intensity, importance and certainty and susceptibility to response effects". *Journal of Personality and Social Psychology*, 54, 940-952.

Labovitz, S. (1967). "Some observations on measurement and statistics." *Social Forces* 46: 151-160.

Labovitz, S. (1970). "The Assignment of Numbers to Rank Order Categories". *Sanford American Sociological Review* , 35(3): 515-524. Published by: American Sociological Association Article Stable URL: <http://www.jstor.org/stable/2092993>

Lazarsfeld, P.F. (1944). "The controversy over detailed interviews – An offer for negotiation". *Public Opinion Quarterly*, 8(1): 38-60. DOI:10.1086/265666

Saris, W.E. and I. Gallhofer (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley

Schuman H., and S. Presser (1981). *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York: Academic Press.

Schwarz, N., and H.J. Hippler (1987). "What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives." In *Social Information Processing and Survey Methodology*, ed. H. Hippler, N. Schwarz, and S. Sudman. New York: Springer-Verlag.

Schwarz, N., and H.J. Hippler (1991). "Response alternatives: The impact of their choice and order". In P. Biemer, R.M. Groves, N.A. Mathiowetz & S. Sudman (Eds), *Measurement errors in surveys* (pp41-56). Chichester, England: Wiley.

Smyth, J.D., Dillman, D.A., Christian, L.M. and M.J. Stern (2006). "Comparing Check-All and Forced-Choice Question Formats in Web Surveys." *Public Opinion Quarterly*, 70(1): 66-77. DOI:10.1093/poq/nfj007

Tourangeau, R., and T.W. Smith (1996). "Asking Sensitive Questions: the Impact of Data Collection Mode, Question Format, and Question Context". *Public Opinion Quarterly*, 60(2): 275-304. DOI: 10.1086/297751

Tourangeau R., L. J. Rips, and K. Rasinski (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. Zanna M. P., and J. K. Rempel

Tables

Table 1: The two split ballot experiments

| Topic (quest. 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 | <i>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</i> 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- <i>How much time, in total, do you spend watching television?</i> A2- <i>And on an average weekday, how much of your time watching television is spent watching news or programs about politics and current affairs?</i> Idem with radio (A3 and A4) and newspapers (A5 and A6). |

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

| | Mexico | | | Colombia | | | Spain | | |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | 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* |
| Avg | 1.88 | 1.53* | 2.30* | 2.06 | 1.70* | 2.64* | 2.10 | 1.60* | 2.38* |

Note: M₁=yes/no; M₂= check-all-that-apply; M₃= no, yes-once, yes-more than once. The stars in the column of M₂ (resp. M₃) indicate when the difference in proportions or means between M₁ and M₂ (resp. M₁ and M₃) is significant: * means at the 5% level, ** means at the 10% level.

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 | | |
|---|--------|--------|-------|----------|--------|---------|-------|--------|--------|
| | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 |
| % 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 | .74 | .82* | .90** | .81 | .85 | .94* | .74 | .75 | .77 |
| A6 | 1.01 | 1.04 | 1.17* | 1.02 | 1.05 | 1.19* | 1.02 | 1.01 | 1.05 |

Note: M1=8categories till >3h by half hours; M2=14categories till >6h by half hours; M3=9categories till>7h by hours. The stars in the column of M2 (resp. M3) indicate when the difference in proportion or means between **M1 and M2** (resp. **M2 and M3**) is significant: * means at the 5% level, ** means at the 10% level.

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

| Expt | Corr | Mexico | | | Colombia | | | Spain | | |
|----------|-------|--------|-----|-------|----------|-----|-------|-------|-----|-------|
| | | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 |
| Pol.part | B1-pp | .42 | .35 | .38 | .37 | .37 | .38 | .37 | .31 | .37 |
| Media | B1-A2 | .24 | .27 | .36** | .24 | .19 | .25 | .26 | .25 | .37* |
| | B1-A4 | .22 | .21 | .29 | .25 | .21 | .30** | .21 | .26 | .21 |
| | B1-A6 | .13 | .23 | .25 | .11 | .18 | .19 | .22 | .21 | .33** |

Note: For the political participation experiment, the stars in the column M2 (resp. M3) indicate when the difference in correlations between M1 and M2 (resp. **M1 and M3**) is significant. For the media use experiment, the stars in the column M2 (resp. M3) indicate when the difference in correlations between M1 and M2 (resp. **M2 and M3**) is significant. * means at the 95% level, ** means at the 90% level.

Appendix

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

| Expt | Corr | Mexico | | | Colombia | | | Spain | | |
|----------|-------|--------|-----|-----|----------|-----|-----|-------|-----|-----|
| | | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 |
| Pol.part | B1-pp | .42 | .36 | .39 | .39 | .38 | .39 | .37 | .29 | .35 |
| Media | B1-A2 | .30 | .29 | .37 | .28 | .21 | .24 | .28 | .24 | .38 |
| | B1-A4 | .24 | .23 | .30 | .27 | .23 | .33 | .25 | .29 | .24 |
| | B1-A6 | .23 | .23 | .29 | .15 | .19 | .21 | .24 | .22 | .34 |