UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 34

2013

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici ESCI-Born - Despatxos 19.517-19.527
Passeig Pujades, 1 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Quality of different scales in an online survey in Mexico and Colombia

Melanie Revilla

RECSM, Universitat Pompeu Fabra

Carlos Ochoa

Netquest

Acknowledgement:

We are very grateful to Netquest for providing us with the necessary data for this paper and especially to Germán Loewe that made this collaboration possible. We would also like to thank Willem Saris, Salvador Masdeu and Oriol Barras that supported us at different levels during the whole process.

Abstract:

The formulation of theories and hypotheses is done at the level of concepts. In order to test them, these concepts are often operationalized using survey questions. However, survey questions never measure the concepts of interest perfectly, because of measurement errors. In order to correct for measurement errors, one needs information about their size, or the size of their complement, the quality. For the USA and Europe, a lot is already known about the quality of questions depending on the scale characteristics. However, in other parts of the world, this was not studied yet. Therefore, in this paper, we use a multitrait-multimethod approach to estimate the quality of 27 questions in Mexico and Colombia. These first results about quality for central and Latin American countries show quality estimates relatively similar in their relationships with the scale characteristics to what was observed in US and European countries.

Resumen:

La formulación de teorías e hipótesis se hace al nivel de los conceptos. Para poder testearlos, estos conceptos son a menudo operacionalizados usando preguntas de encuestas. Sin embargo, las preguntas de encuestas nunca miden perfectamente los conceptos de interés. Siempre hay errores de medición. Para corregir por estos errores de medición, es necesario tener información sobre su tamaño, o su complemento, la calidad. Para EEUU y Europa, ya mucho se sabe sobre la calidad de las preguntas dependiendo de las características de la escala utilizada. Pero en otras partes del mundo, esto no ha sido estudiado ya. Por eso, en este articulo, utilizamos experimentos multirasgos-multimétodos para estimar la calidad de 27 preguntas en México y Colombia. Estos primeros resultados sobre calidad en América central y latina

demuestran que la relación entre la calidad y las características de las escalas es bastante similar a lo que se había encontrado para EEUU y Europa.

Keywords:

Quality, measurement errors, multitrait-multimethod (MTMM) experiments, web surveys, central and Latin America

Palabras claves:

Calidad, errores de medición, experimentos multirasgos-multimétodos, encuestas online, América central y Latina

Introduction

Research usually starts by the formulation of theories, based on some observations. From these theories, hypotheses are derived. Then, these hypotheses are tested. The test determines if they should be accepted or rejected. In some cases, the hypotheses can be tested by conducting an experiment. In others, observational or non-experimental designs are used.

The formulation of theories and hypotheses is done at the level of concepts. These concepts are mental representations, i.e. entities that exist in the brain but are not directly observable. In order to test the hypotheses, it is necessary to operationalize these concepts by specifying empirical indicators or measures for each of them. In observational designs, the measures are often survey questions. If the concepts are simple, what Northrop (1947) calls "concepts by intuition", then a single question is enough to measure them. If the concepts are more complex, what Northrop (1947) calls "concepts by postulation", then more than one question is needed in order to measure them. Explicit definitions of the concepts are necessary in that case.

A good operationalization is one that selects the question that maximizes the strength of the relationship between the latent variable of interest (or concept) and the observed answer to the question (also called indicator or measure), i.e. one that maximizes the quality. The quality can be computed as the product of validity and reliability. The difference between one and the quality estimate corresponds to measurement errors. Therefore, if the quality is equal to one, it means there are no measurement errors at all. This is the ideal situation. Researchers should try to get as close as possible to that. Nevertheless, in practice, there are always at least some random errors, such that the quality is never equal to one.

Measurement errors may affect a lot the results of a research. Differences can be observed that have nothing to do with real differences, but are the consequences of using different measures of the concepts of interest.

An illustration of this is given by Saris and Gallhofer (2007). The authors report the correlations between three indicators of social trust and three indicators of trust in institutions, using data from the European Social Survey round 1. The first indicator for social trust is: “generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?” The first indicator for trust in institutions is: “how much do you personally trust the parliament?” Using a 4-point scale, the correlation in Great-Britain between these two indicators is -0.147 (significant). One may conclude that there is a negative relationship between trusting other people and trusting the parliament.

Nevertheless, using an 11-point scale to ask the same question to the same sample of respondents, the correlation becomes 0.291 (significant). One may conclude that there is a positive relationship between trusting other people and trusting the parliament. The same pattern is found using other indicators of social and/or trust in institutions. This example shows that the same questions asked in the same country in the same survey to the same people lead to opposite conclusions, just because the number of response categories in the scale changed. Since small variations in the choice of the format of the scales have such important consequences on the substantive conclusions, it is really crucial to study and take into account the quality of the questions.

Since this quality, in practice, is never perfect, correction for measurement error is always necessary. In order to do this correction for measurement error, one needs to know the size of the errors. Said differently, one needs to get an estimate of the quality of the questions (Saris and Revilla, 2013).

A lot of research has been done in that direction (e.g. Andrews, 1984; Scherpenzeel and Saris, 1997; Alwin, 2007; Saris and Gallhofer, 2007). Also, procedures have been developed in order to help researchers operationalizing their concepts of interest and maximizing the quality of questions. For instance, Saris and Gallhofer (2007) propose a three-step procedure in order to go from the concept to the request for an answer: distinguishing concepts by postulation and concepts by intuition, developing assertions for each concept by intuition and transforming the assertions in requests for an answer.

Most survey questions are closed questions where a specific scale is proposed to the respondents in addition to the request for an answer. Therefore, researchers also have to make decisions about the format of the scale they want to use for their indicators. Again, the literature provides information about the effects of the wording of survey questions on their responses (Belson, 1981; Schuman and Presser, 1981; Alwin and Krosnick, 1991; Tourangeau, Rips and Rasinski, 2000) and guidelines about which kind of scales to use (Sudman and Bradburn, 1983; Converse and Presser, 1986; Dillman, 2000).

Saris and Gallhofer (2007) propose a meta-analysis of many experiments and use the estimates to predict the impact of the different characteristics of a scale on the quality. The prediction can even be done now in a semi-automatic way using the program SQP 2.0 (Saris et al, 2011)[1].

However, previous research concentrates on the quality of questions asked in Europe and in the US. At the same time, previous research show that the quality varies across countries and languages. This can be because of cultural differences across respondents from different countries or languages, or because of language specific differences that do not allow translating some questions exactly in the same way in another language.

[1] Available for free at http://www.sqp.nl/

Therefore, we cannot extend the results from the US and Europe to other parts of the world. This means that very little is known about the quality of questions in Latin America or in Asia or in Africa. For Venezuela, Handlin (2012) evaluates several common measures of social class in terms of validity and reliability. For Brazil and Mexico, Nyitray et al (2009) use a test-retest approach to estimate the reliability of questions about sexual behaviors.

However, very few studies have been done so far, and they are about very specific topics. Much more information is needed in this direction. One goal of this paper is to start filling in this gap by looking at the quality of questions asked using different scales in Mexico and Colombia. The quality is computed as the product of reliability and validity.

Another specificity of this paper is that we look at the quality of web survey questions. Most of the previous research studied face-to-face or telephone data-collection modes, even if some research used the Telepanel, in the Netherlands (Saris, 1991, 1998) which can be considered as an ancestry of the Web surveys. More recently, there were also a few studies looking at the impact on the quality of using web versus more traditional modes of data collection (e.g. Revilla and Saris, 2012; Revilla, 2013; Revilla, Saris, Loewe, Ochoa, 2013). But there is still little evidence about the quality of questions in web surveys.

So we want to get a first piece of information about the quality of web survey questions in Mexico and Colombia. At the same time, we want to test if the general tendencies about qualities of scales with different properties encountered in previous studies also apply in this new context.

First, we will present the different characteristics of the scales studied and our hypotheses about how these characteristics influence the quality. Then, the method used

to test the hypotheses will be explained, followed by a short presentation of the data used. Finally, the results will be shown and discussed.

Hypotheses

a. The use of agree-disagree (AD) scales versus item specific (IS) scales

Item specific (IS) scales are defined as scales where the categories used to express the opinion are exactly those answers we would like to obtain for this item (Saris, Revilla, Krosnick and Shaeffer, 2010). For instance, if one is interested in the degree of trust a person has in different institutions, a IS scale may be a scale using the labels "no trust at all" to "complete trust". By opposition, an agree-disagree (AD) scale can be used by asking the respondents how much they agree or disagree with the statement "I generally trust this institution". The answer categories can for instance go from "disagree totally" to "agree totally".

The impact on the quality of using AD versus IS scales has already been studied in various studies (Scherpenzeel and Saris, 1997; Saris and Gallhofer, 2007; Saris et al., 2010). The quality of IS scale is in almost all experiments and countries higher than the quality of AD scales. Over several topics and many countries, Saris et al. (2010) get an average difference in quality estimates of around 20% in favour of the IS scales. However, the data they use comes from face-to-face or self-completed paper and pencil interviews In European countries. Therefore, it is interesting to test if the pattern is maintained in web surveys in Latin American countries.

Even if the estimates of quality vary from country to country, the general trend that IS scales perform better appears to be the same in most of the countries previously studied.

Moreover, we do not have strong reason to think that the cognitive mechanisms they propose to explain the difference between IS and AD scales are interacting with the mode of data collection, even if it could be argued that the web survey, by giving more freedom in the pace of the interview to the respondents, may allow them to take more time to think about their answer and that this could allow them to achieve the extra step of the agree-disagree scales in a better way. But even if they have the possibility of doing it, we do not believe that they usually do it.

Thus, our first hypothesis is: in Mexico and Colombia too, the AD scales will lead to a lower quality than the IS ones (*H1*).

b. The number of answer categories

The theory of information (Garner, 1960) states that a scale with two response categories can assess only the direction of the respondents' opinion, attitude or behaviour, whereas if this number of response categories increases, the intensity of the opinion, attitude or behaviour, can also be assessed. If the scale has an odd number of response categories, a neutral position can be observed additionally. Thus, more information can be obtained by using longer scales and using middle points. However, the recommendations about how many points should be used vary in the literature (Likert, 1932; Alwin, 1992; Dawes, 2008).

The crucial question is the following: does more information means higher quality of the questions? The evidences from real data about the impact of the number of answer categories on the quality defined as the strength of the relationship between the observed answer and the latent construct of interest are not so clear (Andrews, 1984; Scherpenzeel, 1995; Alwin, 1997; Alwin, 2007).

Revilla, Saris and Krosnick (forthcoming) suggest that you need to distinguish between AD and IS scales. They found that for the AD scales, the quality decreases when going from 5 to 7 and from 7 to 11 responses categories. However, they do not study IS scales. But they assume that for IS the trend is opposite to the AD scales. This is one of their explanations for the mixed results in the literature.

We follow them to propose Hypothesis 2: the increase in the number of responses categories (till 11) positively affects the quality of IS scales (*H2*).

c. The use of fixed reference points

Following Saris and Gallhofer (2007), we call "fixed reference point" a response category that indicates without any doubt the position of this response category on the subjective opinion scale for all respondents. An example of a label that everybody understands without hesitation is the most extreme possible position, like "completely agree" (Saris and Rooij, 1988; Saris and Gallhofer, 2007).

One basic assumption in survey research is that all respondents have the same response function. This means that two persons with the same opinion will select the same answer category. But if respondents interpret the labels of the response categories differently, then, they might choose different answers even if they have the same opinion. This is the problem of variation in response functions which has been observed in practice by Saris and Rooij (1988).

These authors show that using fixed reference points help to reduce the potential variations by giving a clear meaning, shared by all the respondents, to the answer categories. With one fixed reference point, the authors observe still quite large variations, whereas with two fixed reference points at the two end of the scale, the response functions of the different respondents are becoming much more similar. This is

expected to happen in web surveys as well as in more traditional modes and over different countries.

Therefore, we assume the following: the use of fixed reference points for the two end points of the scale increases the quality (*H3*).

How can we test these hypotheses?

a. Method

The hypotheses can be tested by comparing the quality estimates of scales with different characteristics: AD versus IS scales, scales with different number of answer categories and scales using fixed-reference points or not.

However, we first need to compute the quality estimates. For a given question *i* (also called "trait") and a given scale *j* (also called "method"), the quality, denoted q_{ij}^2, can be computed as the product of the reliability r_{ij}^2 and the validity v_{ij}^2. The reliability coefficient r_{ij} and the validity coefficient v_{ij} can be estimated using Structural Equation Modelling (SEM). The approach used is called MultiTrait-MultiMethod (MTMM, Campbel and Fiske, 1959). More exactly, we use the true score MTMM model proposed by Saris and Andrews (1991). This model explicitly distinguishes reliability and validity coefficients, as can be seen in the system of equations below or in the graphical representation of Appendix 1.

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \qquad (1)$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \qquad (2)$$

Where F_i is the i^{th} trait, M_j is the j^{th} method, Y_{ij} is the observed answer for the i^{th} trait and the j^{th} method, T_{ij} is the true score or systematic component of the response, e_{ij} is the random error associated with Y_{ij}.

Equation (1) defines each observed variable as the sum of the associated systematic component and random errors. Equation (2) says that each systematic component itself is the sum of the trait and the effect of the method used to assess it. By substituting (2) into (1), we get to the more common MTMM model which does not differentiate reliability and validity.

As usual, the random errors are assumed to be uncorrelated with each other and with the independent variables in the different equations. On the contrary, the different traits are assumed to be correlated. The method factors are assumed to be uncorrelated between them and with the traits. Also, the impact of the method factor on the different traits measured with a common scale is assumed to be equal.

In order to be identified, such a true score MTMM model usually requires at least three correlated traits, each measured with three different methods. This means a lot of repetitions if the same respondents have to answer all forms. In order to reduce the cognitive burden of the respondents and to limit the possible memory effects (van Meurs and Saris, 1990), the MTMM approach can be combined with a split-ballot approach (Saris, Satorra and Coenders, 2004). The split-ballot approach consists in splitting respondents randomly into several groups. Each group receives a different "treatment". Since the assignment is random, we expect the groups to be similar, except for sampling variations. Therefore, significant differences between groups are interpreted as coming from the "treatments".

In the split-ballot MTMM design, each split-ballot group gets a combination of two methods for a given set of three traits, instead of getting all the three methods. The different "treatments" consist in different combinations of two methods.

The model is still identified under quite general conditions (Saris, Satorra and Coenders, 2004), even if in practice many non convergence problems and improper

solutions occur (Revilla and Saris, 2013). However, by using a three-group design, most of the non-convergence and improper solutions problems are solved. A three-group design means that the respondents are randomly assigned to three groups. For instance, Group 1 gets methods 1 and 2, group 2 gets methods 2 and 3, and group 3 gets methods 3 and 1. On the other hand, in a three-group design, we can get differences in quality depending if the method is used at the beginning or at the end of the survey: respondents can learn, in that case, the quality will increase, or they can get tired of answering, in that case, the quality will decrease.

The split-ballot true score MTMM model can be estimated with any SEM software. We use the Maximum Likelihood multiple-group estimation procedure of LISREL (Jöreskog and Sörbom, 1991) with the Pearson[2] correlation matrices, means and standard deviations as input data (see Appendix 2 for an example of the initial LISREL input). The different groups correspond to the different split-ballot groups. However, each country is analyzed separately. This choice was done because our goal is not to compare the countries but to compare the quality of different scales within each country. Therefore, doing a combined analysis would not add essential information. But it would make the testing of the model more delicate.

Indeed, in SEM, it is crucial before to look at the estimates to test the fit of the model. Following Saris, Satorra and Van der Veld (2009), we test the models using the JRule software (Van der Veld, Saris, Satorra, 2009). This software takes into account the power of the test, as well as the modification indices and the expected parameter

[2] "If the researcher is interested in measurement-quality altogether (including the effects of categorization), or in assessing the effects of categorization on measurement quality, the Pearson correlations should be used" (Coenders and Saris, 1995, p.141)

change, to test at the level of a single parameter if there is or not a misspecification[3]. If the model is more complex, more possible misspecifications can appear and it becomes more difficult to know which parameters should first be freed.

Starting from the initial model described before, the model is corrected step by step when misspecifications are found, till an acceptable fit is achieved (see Appendix 3 for a list of the extra parameters freed). Then, the reliability and validity coefficients of the final models are used to compute the quality estimates: $q^2_{ij} = r^2_{ij} * v^2_{ij}$.

b. Data

In order to test the hypotheses, we use data from a survey completed by respondents from the Netquest[4] online panel in Mexico and Colombia. In each country, around 1000 panellists answered. Quotas for age and gender were used to get similar distributions in the sample as in the general population on these two variables. The survey includes questions similar to the ones of the European Social Survey (ESS) round 4. It was simply shortened (focusing on the core modules) and adapted to online survey (keeping the design of the web version as similar as possible to the show-cards of the ESS). This survey contains three split-ballot MTMM experiments, about satisfaction, social trust and trust in institutions.

Each experiment is looking at three traits. The satisfaction experiment asks how satisfied the respondents are with the present state of the economy in the country (trait 1), with the way the government is doing its job (trait 2) and with the way the democracy works (trait 3). The experiment about social trust asks if the respondents

[3] A misspecification is defined as a deviation larger than .4 for the standardized loadings and than .1 for the causal effects and correlations (default values of the software; they can be adjusted by the researchers if they wish to do a test more or less strict but we kept the standard thresholds).

[4] More information at: www.netquest.com

would say that most people can be trusted or that you cannot be too careful in dealing with people (trait 1), if the respondents think that most people would try to take advantage of them or would try to be fair (trait 2), and if they would say that most people deserve their trust or that only very few deserve it (trait 3). The experiment about trust in institutions asks how much the respondents personally trust the country's parliament (trait 1), the legal system (trait 2) and the police (trait 3).

Each of the traits is measured with three methods. Table 1 gives the main characteristics of the methods used to measure the traits of each experiment. The complete questionnaire can be found online[5].

Table 1: The main differences across methods

Experiment	***Characteristics of the methods***
Satisfaction	M_1 = 11 points IS (completely in/satisfied) M_2 = 11 points IS (in/satisfied) M_3 = 5 points AD
Social trust	M_1 = 11 points IS M_2 = 2 points IS M_3 = 6 points IS
Trust in institutions	M_1 = 11 points IS M_2 = 6 points battery IS M_3 = 11 score IS

[5] The version for Mexico is available at http://test.nicequest.com/surveys/global_glacier/eb5e4c34-e56e-4f1c-be7d-7354febeb01f (the questions where adapted to Colombia just by changing the name of the country)

The satisfaction experiment allows testing the difference between AD and IS scales (*H1*) and the effect of fixed-reference points (*H3*). Hypothesis 1 implies that we expect the quality of M_3 to be the lowest. Hypothesis 3 implies that we expect the quality of M_1 to be higher than the one of M_2. All together, we therefore expect for the satisfaction experiment to have: $q^2_{M1} > q^2_{M2} > q^2_{M3}$

The social trust and trust in institution experiments allow looking at the quality for different numbers of response categories when focusing on IS scales (*H2*). For the social trust experiment, we expect: $q^2_{M1} > q^2_{M3} > q^2_{M2}$. For the trust in institutions experiment, we expect: $q^2_{M1} = q^2_{M3} > q^2_{M2}$.

Results: the quality estimates

Table 2 presents the quality estimates for each experiment both in Mexico and Colombia. It gives the quality for each trait and method separately, together with the average quality for the three traits together. When the quality varies depending on the position of the method, both are indicated: the estimate when the method is at the beginning (with a "B" in parentheses) and when the method is at the end (with an "E" in parentheses).

Table 2: Quality estimates q^2_{ij} in Mexico and Colombia for the different traits (t_i) and methods (M_j)

Experiment	*Method*	*Mexico*				*Colombia*			
		t_1	t_2	t_3	Mean	t_1	t_2	t_3	Mean
Satisfaction	M_1=11pts compl	.63	.70	.78	.70	.79	.85	.88	.84
	M_2=11pts	.57	.68	.70	.65	.67	.81	.80	.76
	M_3=5pts AD	.50	.66	.57	.58	.41	.47	.44	.44
Social trust	M_1=11 pts (B)	.68	.75	.73	.72	.63	.61	.67	.64
	M_1=11 pts (E)	.81	.85	.81	.83	.72	.67	.73	.71
	M_2=2 pts (B)	.42	.52	.66	.53	.41	.56	.61	.53
	M_2=2 pts (E)	.42	.42	.55	.46				
	M_3=6 pts (B)	.67	.63	.80	.70	.60	.77	.87	.75
	M_3=6 pts (E)	.67	.63	.80	.70	.90	.77	.98	.89
Trust in institutions	M_1= 11pts	.78	.85	.85	.83	.78	.80	.89	.82
	M_2= 6pts battery	.68	.70	.53	.64	.75	.70	.67	.71
	M_3= 11score (B)	.85	.85	.76	.82	.73	.85	.81	.80
	M_3= 11score (E)	.78	.83	.81	.81				

Note: Pts = number of response categories; compl = labels of the end points start with "completely"

Before looking at these estimates, we should mention some limits encountered during the analyses. First, even if a three group design was used, in the experiment social trust, the initial model in both countries led to improper solutions (also referred to as Heywood cases), with a negative variance for the third method factor. By freeing some parameters, in particular allowing that some parameters can vary for a given method depending if the method was asked at the beginning of the survey or at the end, we could get a proper solution. However, the results are very sensitive to corrections. It

is difficult to be sure that the corrections we made are all adequate and that we did not miss any other correction that would be necessary, even more when many corrections cannot be done without getting again an improper solution (negative variances). Therefore, we should be very careful about the conclusions we can draw from this experiment. Replications of the results would be necessary to get more confidence.

For the two other experiments, the initial models led to proper solutions and the testing was a bit less delicate. The results were less sensitive to corrections. Sometimes, introducing some of the parameters misspecified in JRule does not really seem to be helpful. In these cases, we chose not to introduce them, even if the general fit measures of the model were not so good.

Keeping this in mind, Table 2 shows that in the satisfaction experiment, the quality for the 11-point scale with fixed reference end points (M_1) is the highest. It is followed by the one of the 11-point scale without fixed reference end points (M_2) and finally the one of the 5 AD scale (M_3). The differences are generally larger between M_2 and M_3 than between M_1 and M_2. This suggests that using AD scales, as in previous research for other countries and modes of data collection, leads to a much lower quality also for panellists of a web survey in Mexico Colombia than using IS scales. We should notice that the number of points also varies. However, previous research (Revilla, Saris and Krosnick, forthcoming) found that the quality of 11-point AD scales is in general lower than the one of 5-point AD scales. Therefore, we expect the lower quality to be due to the fact that the scale is AD and not to the number of points. Finally, using fixed-reference points also increases a bit the quality, but in a lower proportion.

About the trust in institution experiment, Table 2 shows also support for our hypotheses. We expected the quality of the two 11-point scales (M_1 and M_3) to be equal and higher than the one of the 6-point scale (M_2). We find that indeed, the lower quality

is the one of M_2 and that even if not exactly equal, the quality estimates for M_1 and M_3 are very similar in general. In Mexico, we also find a difference for M_3 depending on the position of the method within the questionnaire. But taking the average over the three traits erases this difference. Overall, the results indicate that using 11-points scales with separate questions (either with a radio button scale or asking to write a score between 0 and 10) leads to a better quality than using a 6-point scale with all questions combined in a battery. This can be a combined effect of the number of points and presentation in a battery.

Finally, for the social trust experiment, it seems that indeed the shorter scale (M_2) has the lower quality. However, the order between the 11-point (M_1) and the 6-point scale (M_3) is different depending on the country: in Mexico, as we expected, M_1 has the highest quality, whereas in Colombia it is M_3. This suggests that we get a lower quality by using a 2-point scale than a 6- or 11-point scale, but which of the 6- and 11-point scale is better varies across countries. Nevertheless, these results should be confirmed by further research, for the limits mentioned earlier.

What can we conclude? What is next?

In conclusion, this paper uses a split-ballot MTMM approach to get estimates of the quality, defined as the strength of the relationship between the latent variable of interest and the observed answers, in two countries for which this had not been done before: Mexico and Colombia. Also, it studies the quality in a web survey instead of in more traditional modes.

Overall, the analyses suggest that the trends discovered for other countries and data-collection modes also apply in this new context. Support is found for two hypotheses:

(*H1*): in Mexico and Colombia, AD scales lead to a lower quality than IS ones

(*H3*): the use of fixed reference points for the two end points of the scale increases the quality

Hypothesis 2 is also generally supported:

(*H2*): the increase in the number of responses categories (till 11) positively affects the quality of IS scales.

Only in Colombia in the social trust experiment, the results are not completely in line with *H2*, since the 6-point scale has a higher quality than the 11-point scale. But this may be linked to the problems encountered during the analyses and testing of the model for this experiment. So in general, this study shows support for the three hypotheses. Moreover, the quality estimates are quite similar in our analyses to what has been found in the US or Europe (e.g. comparing with results in Saris and Gallhofer, 2007).

Nevertheless, more MTMM experiments would need to be done in these new geographical areas, because the quality estimates are not exactly equal in the different countries and languages. To be able to correct for measurement errors in surveys done in different places, it is necessary to get estimates of the size of the errors or of their complement: the quality estimates. This is a crucial first step to be able to get correct estimates of the relationships of interest. It is even more crucial in the frame of comparative research: standardized relationships cannot be compared across countries if the quality estimates are not similar, except if we first correct for these differences in quality. More MTMM experiments are therefore really necessary.

References

Alwin, D.F. (1992). ‘‘Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement.’’ *Sociological Methodology*, 22: 83-118, edited by Peter V. Marsden. Washington, DC: American Sociological Association

Alwin, D.F. (1997). ‘‘Feeling Thermometers versus 7-point Scales: Which Are Better?’’ *Sociological Methods and Research* 25:318.

Alwin, D.F. (2007). *Margins of Errors: A Study of Reliability in Survey Measurement*. Wiley and Sons, Inc, Hoboken, New Jersey.

Alwin D.F., and J.A. Krosnick (1991). “The reliability of survey attitude measurement. The influence of question and respondent attributes”. *Sociological Methods and Research*, 20, 139–181.

Andrews, F.M. (1984). “Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach.” *Public Opinion Quarterly* 48(2):409-442.

Belson, W. (1981). *The design and understanding of survey questions*. London , Gower

Campbell, D.T. and D.W. Fiske (1959). ‘‘Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.’’ *Psychological Bulletin* 6:81-105

Coenders, G., and W.E. Saris (1995). “Categorization and measurement quality. The choice between Pearson and Polychoric correlations”. In W.E. Saris, *The MTMM approach to evaluate measurement instruments* (1995), Chapter 7, 125-144. Citeseer.

Converse J.M., and S. Presser (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage.

Dawes, J. (2008). ‘‘Do Data Characteristics Change According to the Number of Points Used? An Experiment Using 5-point, 7-point and 10-point Scales.’’ *International Journal of Market Research* 50:61-77.

Dillman D.A. (2000). *Mail and Internet Surveys. The Tailored Design Method*. New York: Wiley.

Garner, W.R. (1960). ‘‘Rating Scales, Discriminability, and Information Transmission.’’ *Psychological Review* 67:343-52.

Handlin, S. (2012). “Survey Research and Social Class in Venezuela: Evaluating Alternative Measures and their Impact on Assessments of Class Voting”. *Latin American Politics and Society*, 54(4)

Jöreskog, K.G. and D. Sörbom (1991). *LISREL VII: A Guide to the Program and Applications*. Chicago: SPSS

Likert, R. (1932). ‘‘A Technique for the Measurement of Attitudes.’’ *Archives of Psychology* 140:1-55.

Northrop, F.S.C. (1947). *The Logic of the Sciences and the Humanities*. New York: World Publishing Company.

Nyitray, A.G, Kim, J., Hsu, C.H, Papenfuss, M., Villa, L., Lazcano-Ponce, E. and A.R. Giuliano. (2009). “Test-retest reliability of a sexual behavior interview for men residing in Brazil, Mexico, and the United States: the HPV in Men (HIM) study”. *American Journal of Epidemiology* 170: 965–974.

Revilla, M. (2013). “Measurement invariance and quality of composite scores in a face-to-face and a web survey” *Survey Research Methods* 7.1(2013): 17-28.

Revilla, M., and W.E. Saris (2012). "A comparison of the quality of questions in a face-to-face and a web survey". *International Journal of Public Opinion Research.* Summer 2013, 25 (2): 242-253. First published online April 17, 2012. doi: 10.1093/ijpor/eds007

Revilla, M., and W.E. Saris (2013). "The Split-ballot Multitrait-Multimethod Approach: Implementation and Problems" *Structural Equation Modeling: A Multidisciplinary Journal*, 20:1, 27-46.

Revilla, M., Saris, W.E., and J.A. Krosnick (forthcoming) "Choosing the number of categories in agree-disagree scales". *Sociological Methods and Research*

Revilla, M., Saris, W.E., Loewe, G, and C. Ochoa (2013). "Can an online panel oriented to marketing surveys get similar SEM estimates of question quality as the face-to-face European Social Survey?" *RESCM Working Paper 33*. Available at: http://www.upf.edu/survey/working/working.html

Saris, W.E. (1991). *Computer Assisted Interviewing*. Newbury Park CA: Sage.

Saris, W.E. (1998). "Ten years of interviewing without interviewers: The telepanel". In M.P. Couper , R.P. Baker, J. Bethlehem, C. Clark, J. Martin, W.L. Nicholls II, and J.M. O'Reilly (eds.), *Computer-assisted Survey Information Collection*, New York: Wiley, 409–431.

Saris, W.E., and F.M. Andrews (1991). "Evaluation of Measurement Instruments using a Structural Modeling Approach." *Measurement Errors in Surveys* 575-599.

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

Saris W.E, D. Oberski, M. Revilla, D. Zavalla, L. Lilleoja, I. Gallhofer and T. Grüner (2011). "The development of the Program SQP 2.0 for the prediction of the

quality of survey questions". *RECSM Working paper 24.* Available at: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf

Saris, W.E., and M. Revilla (2013). "Correction for measurement errors in survey research: necessary and possible". *RECSM Working Paper 31*. Available at: http://www.upf.edu/survey/working/working.html

Saris, W.E., Revilla, M., Krosnick, J.A., and E.M. Shaeffer (2010) "Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options" *Survey Research Methods*, 4.1(2010): 61-79.

Saris, W. E. and K. d. Rooij (1988). "What kind of terms should be used for reference points? Variation in Response Functions: a Source of Measurement Error in Attitude Research". W.E. Saris. Amsterdam, *Sociometric Research Foundation*: 199-218

Saris,W. E., A. Satorra, G. Coenders (2004). "A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design." *Sociological Methodology* 34 311-347.

Saris, W.E., A. Satorra, and W. Van der Veld (2009). "Testing Structural Equation Models Or Detection of Misspecifications." *Structural Equation Modeling: A Multidisciplinary Journal* 16:561-82.

Scherpenzeel, A. (1995). *A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies*. Amsterdam, the Netherlands: Nimmo.

Scherpenzeel, A., and W.E. Saris (1997). "The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies." *Sociological Methods and Research* 25 (3):341.

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

Sudman S., and N. M. Bradburn (1983). *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey Bass.

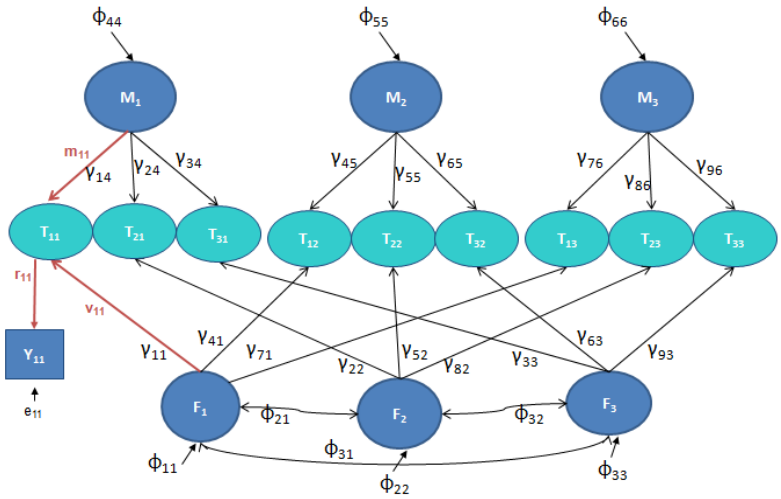
Tourangeau R., L.J. Rips, and K. Rasinski (2000). *The Psychology of Survey Response*. Cambridge MA: Cambridge University Press.

Van der Veld, W., W.E. Saris, and A. Satorra (2008). *Judgment Aid Rule*. Jrule 2.0: User manual (Unpublished Manuscript, Internal Report). Radboud University Nijmegen, the Netherlands.

Van Meurs, L. and W.E. Saris (1990). ‘‘Memory Effects in MTMM Studies.’’ In *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod Studies*, edited by W.E. Saris and L. van Meurs. Amsterdam, the Netherlands: North Holland. 134-146.

Appendices

Appendix 1: Path diagram of the true score MTMM model using LISREL's notations



Appendix 2: Initial Model, LISREL input

```
Analysis of Netquest satisf group 1 Colombia
Data ng=3  ni=9 no=640 ma=cm
km file=sb-group-1.corr
mean file=sb-group-1.mean
sd file=sb-group-1.sd
model ny=9 ne=9 nk=6  ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi
ph=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 1 te 7 7 te 8 8 te 9 9
value 0 ly 7 7 ly 8 8 ly 9 9
free ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3
value 1 ga 1 4 ga 2 4 ga 3 4 ga 4 5 ga 5 5 ga 6 5 ga 7 6 ga 8 6 ga 9 6
free ph 2 1 ph 3 1 ph 3 2
value 1 ph 1 1 ph 2 2 ph 3 3
fr ph 4 4 ph 5 5 ph 6 6
out mi iter= 300 adm=off sc
Analysis of group 2
Data ni=9 no=668 ma=cm
km file=sb-group-2.corr
mean file=sb-group-2.mean
sd file=sb-group-2.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
va 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6
value 1 te 1 1 te 2 2 te 3 3
value 0 ly 1 1 ly 2 2 ly 3 3
```

```
out iter= 300 adm=off sc
Analysis of group 3 Netquest
Data ni=9 no=694 ma=cm
km file=sb-group-3.corr
mean file=sb-group-3.mean
sd file=sb-group-3.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in
fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
va 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9
value 1 te 4 4 te 5 5 te 6 6
value 0 ly 4 4 ly 5 5 ly 6 6
pd
out mi iter= 300 adm=off sc
```

Appendix 3: List of corrections from the initial model, indicators of fit

The variables are in the following order: first, method 1 trait 1, trait 2, trait 3, then, method 2 trait 1, trait 2, trait 3, and finally, method 3 trait 1, trait 2 and trait 3.

Satisfaction experiment

- Mexico:
  - Free phi 5 4 in group 1
  - χ^2=152.66 with df=38
  - JRule: 5 possible misspecifications left
- Colombia:
  - No corrections
  - χ^2=179.11 with df=39
  - JRule: 8 possible misspecifications left

Social Trust experiment

- Mexico:
  - Analyze Correlation matrix and not covariance
  - Free theta 4 1 in group 1
  - Free thetas 7 4, 8 5, gammas 5 5, 6 5 in group 2
  - Free thetas 1 1, 2 2, 3 3, 8 2 in group 3
  - χ^2=81.56 with df=30
  - JRule: 8 possible misspecifications left
- Colombia:
  - Free gamma 8 6 group 1

- Free thetas 1 1, 2 2, 3 3, 7 7, 8 8, 9 9, gammas 7 1, 8 2, 9 3 in group 3
- χ^2=89.55 with df=29
- JRule: 8 possible misspecifications left

Trust in institutions experiment

- Mexico:
  - Free gammas 9 6, 3 4, 5 5, theta 6 3 in group 1
  - Free theta 9 6 in group 2
  - Free gammas 7 1, 8 6, 9 6 in group 3
  - χ^2=130.96 with df=31
  - JRule: no possible misspecifications left
- Colombia:
  - Free gammas 3 4, 9 6 in group 1
  - χ^2=97.68 with df=37
  - JRule: 2 possible misspecifications left