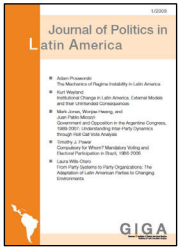


Journal of Politics in Latin America

Revilla, Melanie, and Carlos Ochoa (2015),
Quality of Different Scales in an Online Survey in Mexico and Colombia, in:
Journal of Politics in Latin America, 7, 3, 157–177.

URN: http://nbn-resolving.org/urn/resolver.pl?urn:nbn:de:gbv:18-4-9035

ISSN: 1868-4890 (online), ISSN: 1866-802X (print)

The online version of this article can be found at: <www.jpla.org>

Published by
GIGA German Institute of Global and Area Studies, Institute of Latin American Studies and Hamburg University Press.

To subscribe to the print edition: <ilas@giga-hamburg.de>
For an e-mail alert please register at: <www.jpla.org>

The *Journal of Politics in Latin America* is part of the GIGA Journal Family, which also includes *Africa Spectrum*, *Journal of Current Chinese Affairs* and *Journal of Current Southeast Asian Affairs*: <www.giga-journal-family.org>.

Research Note

Quality of Different Scales in an Online Survey in Mexico and Colombia

Melanie Revilla and Carlos Ochoa

Abstract: The formulation of theories and hypotheses is done at the level of concepts. These concepts are often tested by operationalizing them using survey questions. However, measurement errors make it impossible for survey questions to measure the concepts of interest perfectly. In order to correct for measurement errors, information is needed 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, but this has not yet been studied in some other parts of the world. In this paper, we use a multitrait-multimethod approach to estimate the quality of 27 questions in Mexico and Colombia. These initial results on quality for Central and Latin American countries show quality estimates that are relatively similar in terms of their relationships with the scale characteristics to what has been observed in the USA and Europe.

■ Manuscript received 28 March 2014; accepted 14 October 2015

Keywords: Latin America, Central America, Mexico, Colombia, quality, measurement errors, multitrait-multimethod (MTMM) experiments

Melanie Revilla, PhD, is a researcher at RECSM, Universitat Pompeu Fabra. Her current research focuses on measurement errors and quality issues in web surveys. Her publications and more information can be found on the following website: <www.upf.edu/survey/members/melanie.html>.
E-mail: <melanie.revilla@hotmail.fr> or <melanie.revilla@upf.edu>

Carlos Ochoa is Marketing and Innovation Director at Netquest, the leading provider of online panels and advanced survey software in Latin America, Spain and Portugal. As responsible for the data delivered by an Online Panel Company, he is fostering innovation projects mainly in the quality data collection area. Publications: <https://es.linkedin.com/in/carlosochoa1050>. Website: <www.netquest.com>
E-mail: <cochoa@netquest.com>

Introduction

Research usually starts with the formulation of theories, based on observations. From these theories, hypotheses are derived and tested to determine whether they should be accepted or rejected. The formulation of theories and hypotheses is done at the level of concepts. These concepts are mental representations; that is, entities that exist in the brain, but are not directly observable.

In order to test the hypotheses, these concepts are operationalized 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 they are complex – what Northrop (1947) calls "concepts by postulation" – then more than one question is needed to measure them and explicit definitions are necessary.

A good operationalization selects a 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). In other words, the question maximizes the quality, which can be computed as the product of validity and reliability. The measurement errors are equal to one minus the quality. Therefore, there are no measurement errors when the quality is equal to one. Researchers should try to get as close as possible to this ideal situation.

In practice, however, there are always errors, and these errors may affect many of the results of a study. Differences can be observed that are not real, but are in fact the consequences of using different measures of the concepts of interest. Saris and Gallhofer (2007) provided an illustration using the European Social Survey round 1. They asked the following questions to measure social trust and trust in institutions, respectively:

- "Would you say that most people can be trusted or that you can't be too careful in dealing with people?"
- "How much do you personally trust the parliament?"

Saris and Gallhofer (2007) reported that the correlation between these two measures in Great Britain, using a four-point scale, was -0.147, which is significant. One may conclude that there is a negative relationship between trusting others and trusting the parliament. Nevertheless, when an 11-point scale was used to ask the same question to the same

sample, the correlation was 0.291 (significant). One may conclude that there is a positive relationship.

The same pattern was found using other indicators of social trust and/or trust in institutions. The above example shows that opposite conclusions can be drawn using the same questions asked in the same country in the same survey to the same people, just because the number of response categories changed. Since small variations in the choice of the format of the scales have such important consequences on the substantive conclusions, it is crucial to study and take into account the quality of the questions. It is also necessary to correct for measurement error (Saris and Gallhofer 2007; Saris and Revilla 2015).

In order to do this correction, one needs to know the size of the errors. In other words, it is necessary to have an estimate of the quality of the questions.

Research has been done in this 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 operationalize their concepts of interest. For instance, Saris and Gallhofer (2007) proposed a three-step procedure for moving from the concept to the request for an answer.

In most survey questions, a specific scale is proposed to the respondents in addition to the request for an answer. Therefore, researchers must also have make decisions about the format of the scale. The literature provides information about the effects of the wording of questions on the responses (Belson 1981; Schuman and Presser 1981; Alwin and Krosnick 1991; Tourangeau, Rips, and Rasinski 2000) and guidelines about which scale to use (Sudman and Bradburn 1983; Converse and Presser 1986; Dillman 2000).

Saris and Gallhofer (2007) used the estimates of a meta-analysis of many experiments to predict the impact that the different characteristics of a question would have on the quality. Even now, such predictions can be made in a semi-automatic way using the program SQP 2 (Saris et al. 2011), which is available for free at <www.sqp.nl/>.

However, previous research has concentrated on the quality of questions asked in Europe and in the USA, but has also shown that the quality varies across countries. This could be due to cultural differences across respondents from different countries or to language-specific differences that do not make it possible for some questions to retain their exact meaning once translated.

Therefore, we cannot extend the results from the USA and Europe to other parts of the world. Accordingly, very little is known about the

quality of questions in Latin America. Handlin (2013) evaluated several common measures of social class in terms of validity and reliability in Venezuela. Nyitray et al. (2009) used a test–retest approach to estimate the reliability of questions about sexual behavior in Brazil and Mexico. However, the total number of studies conducted so far is quite small and their topics are quite specific.

The main goal of the present paper[1] is to start filling this gap by providing initial information about the quality (computed as the product of reliability and validity) of questions asked using different scales in Mexico and Colombia.

These quality estimates can be used while designing future questionnaires. They can help to decide which scale to put in the survey. They can also be used after data collection in order to correct for measurement errors and achieve proper standardized relationships across the different variables (DeCastellarnau and Saris 2014).

We start by presenting the different characteristics of the scales studied and our hypotheses regarding how these characteristics influence the quality. We then explain the method used to test the hypotheses, followed by a short presentation of the data. Finally, the results will be shown and discussed.

Hypotheses

A The Use of Agree–Disagree (AD) Scales versus Item-Specific (IS) Scales

In IS scales the categories used to express the opinion are exactly the answers that the researcher would like to obtain for this item (Saris et al. 2010). For instance, if one is interested in the degree of trust a person has in different institutions, an IS scale may be a scale that ranges from "no trust at all" to "complete trust". An AD scale asks respondents how much they agree or disagree with a specific statement; for example: "I generally trust this institution". The answer categories can range, for instance, from "disagree totally" to "agree totally".

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

The impact on the quality of using AD versus IS scales has already been studied (Scherpenzeel and Saris 1997; Saris and Gallhofer 2007; Saris et al. 2010). In almost all experiments and countries, the quality of IS scales is higher than that of AD scales. Over several topics and many countries, Saris et al. (2010) obtained an average difference in quality estimates of around 20 percent in favor of the IS scales. One of the main problems that Saris et al. identified with AD scales is that they can elicit high levels of acquiescence bias; that is, a high tendency of some respondents to agree with any item, regardless of its content. Such behavior is usually explained by a tendency to avoid social friction (Leech 1983) or to defer to people with higher social status (Lenski and Leggett 1960). Harzing (2006) shows significant differences in acquiescence across countries from different continents. Mexico has a higher level of acquiescence than the USA and all European countries (Harzing 2006: 253). Among the possible cultural causes (cf. Hofstede 2001) for these differences are the high power distance existing in Mexico, the medium uncertainty avoidance, and the high degree of collectivism (Johnson et al. 2005). Higher acquiescence can lead to more systematic errors when AD scales are used and, therefore, to lower quality. The general trend for Western countries was that IS scales perform better. We expect this to be the case to an even greater degree in Latin American countries. Thus, our first hypothesis is as follows: The AD scales will lead to a much 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 only assess the direction of the respondents' opinion, attitude or behavior; if the number of response categories increases, the intensity of the opinion, attitude, or behavior can also be assessed. Additionally, if the scale has an odd number of response categories, a neutral position can be observed. Therefore, more information can be obtained by using longer scales and middle points. However, the recommendations about how many points should be used vary in the literature (Likert 1932; Alwin 1992; Dawes 2008).

The question is whether more information means that the questions will have higher quality. The evidence 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, 2007).

Revilla, Saris, and Krosnick (2013) suggested that you need to distinguish between AD and IS scales. They found that, for the AD scales, the quality decreases when going from five to seven response categories and from seven to 11 response categories. They did not study IS scales, but they assumed that the trend is opposite for IS. This is one of their explanations for the mixed results in the literature.

The theory of information is not country-specific. Therefore, also for Latin American countries, we propose Hypothesis 2: Increasing the number of response categories (up to 11) positively affects the quality of IS scales (*H2*).

C The Use of Fixed Reference Points

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

A 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. If respondents interpret the labels of the response categories differently, they might choose different answers even if they have the same opinion. This is the problem of variation in response functions that was observed in practice by Saris and De Rooij (1988).

Saris and De Rooij (1988) showed that using fixed reference points reduces the potential variations by giving a clear meaning, shared by all the respondents, to the answer categories. With one fixed reference point, quite large variations can still be observed, whereas with two fixed reference points at the two endpoints of the scale, the response functions of the different respondents become much more similar.

We can expect the impact of using fixed reference points to vary depending on the level of extreme response style (ERS) in a country, which is defined as the tendency of some respondents to select the endpoints categories of a scale. ERS varies in terms of the function of the social need for clarity and precision (Johnson et al. 2005) and of the country-level extraversion (Harzing 2006). For Spanish-speaking countries, Harzing found high ERS, with Mexico having the highest ERS of the 26 countries she studied (cf. Harzing 2006: 253).

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

How Can We Test These Hypotheses?

A Method

We tested the hypotheses by comparing the quality estimates of scales with different characteristics: AD versus IS scales, scales with different numbers of answer categories, and scales using fixed-reference points versus those that did 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 modeling (SEM). The approach used is the multitrait-multimethod (MTMM, Campbell and Fiske 1959). More specifically, we use the true score MTMM model proposed by Saris and Andrews (1991), which explicitly distinguishes reliability and validity coefficients, as can be seen in the system of equations below or in the graphical representation of Appendix 1.

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad (1)$$
$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad (2)$$

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, and 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 is itself the sum of the trait and the effect of the method used to assess it. By substituting (2) into (1), we obtain the more common MTMM model, which does not differentiate between 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 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 repetition if the same respondents have

to answer all forms. In order to reduce the cognitive burden of the respondents and to limit possible memory effects (van Meurs and Saris 1990), the MTMM approach is combined with a split-ballot approach (Saris, Satorra, and Coenders 2004). Using this approach, respondents are split randomly into several groups and each group is given a combination of two methods for a given set of three traits, instead of getting all the three 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, using a three-group design (that is, 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) solves most of the non-convergence and improper solutions problems. On the other hand, differences in quality can be obtained depending on whether the method is used at the beginning or at the end of the survey: respondents can learn (in which case the quality will increase) or they can get tired of answering (in which case the quality will decrease).

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

In SEM, before looking at the estimates it is crucial 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, and Satorra 2008). This software takes into account the power of the test, the modification indices, and the expected parameter change, to test at the level of a single parameter whether there is a misspecification.[2] If the model is more complex, it is more difficult to know which parameters should be freed first.

Starting from the initial model described above, the model is corrected step by step when misspecifications are found until an acceptable

2 Default values of the software are used so that a misspecification is defined as a deviation larger than .4 for the standardized loadings and larger than .1 for the causal effects and correlations.

fit is achieved (see Appendix 3 for a list of the extra parameters freed). The reliability and validity coefficients of the final models are then used to compute the quality estimates: $q^2_{ij} = r^2_{ij} * v^2_{ij}$.

B Data

We test the hypotheses using data from a survey completed by respondents from the Netquest online panel (<www.netquest.com>) in Mexico and Colombia. Approximately 1000 panelists responded in each country. Quotas for age and gender were used in order to obtain similar distributions in the sample as in the general population on these two variables.

The use of web data collection was driven by practical reasons. This method may affect the quality estimates slightly, but previous research shows that it is possible to get similar quality estimates for web and face-to-face surveys, both for a probabilistic-based online panel (Revilla and Saris 2012) and for an access online panel (Revilla et al. 2015). However, the Internet penetration is higher in the countries studied in previous research (92.9 percent in the Netherlands, 67.2 percent in Spain[3]) than in Colombia (59.5 percent) and even more than in Mexico (36.5 percent). Therefore, the samples may not be completely representative of the general population. However, Revilla (2012) did not find any effect on the quality estimates of respondents' main background characteristics, so the general results should not be disturbed.

The survey is a shortened version of the European Social Survey (ESS) round 4. The survey contains three split-ballot MTMM experiments regarding satisfaction, social trust, and trust in institutions.

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 democracy works (trait 3). The experiment about social trust asks whether the respondents would say that most people can be trusted or that you cannot be too careful in dealing with people (trait 1), whether the respondents think that most people would try to take advantage of them or would try to be fair (trait 2), and whether 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 the degree to which the respondents personally trust the country's parliament (trait 1), legal system (trait 2), and the police (trait 3).

3 See <www.internetworldstats.com/stats4.htm#europe> (9 November 2015).

Each of the traits is measured with three methods. Table 1 gives their main characteristics. The complete questionnaire can be found online.[4]

Table 1. The Main Differences across Methods

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

The satisfaction experiment makes it possible to test the difference between AD and IS scales (*H1*) and the effect of fixed-reference points (*H3*). We expect the quality of M_3 to be the lowest (*H1*) and the quality of M_1 to be higher than that of M_2 (*H3*). Overall, therefore, we expect the satisfaction experiment to have: $q^2_{M1} > q^2_{M2} > q^2_{M3}$.

The social trust and trust in institution experiments makes it possible to look 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 in Mexico and Colombia. It gives the quality for each trait and method separately, together with the average quality across the three traits. 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).

4 The version for Mexico is available at <http://test.nicequest.com/surveys/global_glacier/eb5e4c34-e56e-4f1c-be7d-7354febeb01f> (it was adapted to Colombia just by changing the name of the country) (9 November 2015).

Table 2. Quality Estimates q^2_{ij} in Mexico and Colombia for the Different Traits (t_i) and Methods (M_j)

Experiment	Method	Mexico			
		t_1	t_2	t_3	Mean
Satisfaction	M_1 = 11-pt compl.	.63	.70	.78	.70
	M_2 = 11-pt	.57	.68	.70	.65
	M_3 = 5-pt AD	.50	.66	.57	.58
Social trust	M_1 = 11-pt (B)	.68	.75	.73	.72
	M_1 = 11-pt (E)	.81	.85	.81	.83
	M_2 = 2-pt (B)	.42	.52	.66	.53
	M_2 = 2-pt (E)	.42	.42	.55	.46
	M_3 = 6-pt (B)	.67	.63	.80	.70
	M_3 = 6-pt (E)	.67	.63	.80	.70
Trust in institutions	M_1 = 11-pt	.78	.85	.85	.83
	M_2 = 6-pt battery	.68	.70	.53	.64
	M_3 = 11-pt score (B)	.85	.85	.76	.82
	M_3 = 11-pt score (E)	.78	.83	.81	.81

Experiment	Method	Colombia			
		t_1	t_2	t_3	Mean
Satisfaction	M_1 = 11-pt compl.	.79	.85	.88	.84
	M_2 = 11-pt	.67	.81	.80	.76
	M_3 = 5-pt AD	.41	.47	.44	.44
Social trust	M_1 = 11-pt (B)	.63	.61	.67	.64
	M_1 = 11-pt (E)	.72	.67	.73	.71
	M_2 = 2-pt (B)	.41	.56	.61	.53
	M_2 = 2-pt (E)	.41	.56	.61	.53
	M_3 = 6-pt (B)	.60	.77	.87	.75
	M_3 = 6-pt (E)	.90	.77	.98	.89
Trust in institutions	M_1 = 11-pt	.78	.80	.89	.82
	M_2 = 6-pt battery	.75	.70	.67	.71
	M_3 = 11-pt score (B)	.73	.85	.81	.80
	M_3 = 11-pt score (E)	.73	.85	.81	.80

Note: Pt = 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 social trust experiment, the initial model in both countries led to improper solutions, with a negative variance for the third method factor. By allowing some parameters to vary for a given method depending on whether the method was asked at the beginning of the survey or at the end, we obtained 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. Therefore, we should be careful about the conclu-

sions we draw from this experiment. Replication of the results would be needed in order to achieve greater confidence.

For the two other experiments, the initial models led to proper solutions and the results were less sensitive to corrections. When the introduction of parameters misspecified in JRule did not change the results, 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 endpoints (M_1) is the highest. It is followed by the one 11-point scale that does not have fixed reference endpoints (M_2), and finally that of the five-point AD scale (M_3). The differences are generally larger between M_2 and M_3 than between M_1 and M_2. Using AD scales led to lower quality. In particular, the difference in Colombia is huge. We should note that the number of points also varies. However, Revilla, Saris, and Krosnick (2013) found that AD scales with 11 points generally have lower quality than those with five points. Finally, using fixed reference points also increases the quality, but in a lower proportion. This is in line with *H3*.

With regard to the trust in institution experiment, we expected the quality of the 11-point scales (M_1 and M_3) to be equal and higher to that of the six-point scale (M_2). Indeed, we found that M_2 has the lowest quality. The quality estimates for M_1 and M_3 are very similar in general. In Mexico, we also found a difference for M_3 depending on the position of the method within the questionnaire, but taking the average over the three traits erased this difference. Overall, using 11-point scales with separate questions (either with a radio button scale or by asking respondents to write a score between 0 and 10) leads to a better quality than using a six-point scale with all questions 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, the shortest scale (M_2) has the lowest quality. However, the order between the 11-point (M_1) and the six-point scale (M_3) is different depending on the country. In Mexico, M_1 has the highest quality (as expected), whereas in Colombia it is M_3. This finding suggests that using a two-point scale results in lower quality than a six- or 11-point scale; however, which of the six- and 11-point scales is better varies across countries. Nevertheless, these results should be confirmed by further research, for the limits mentioned earlier.

What Can We Conclude?

In conclusion, this study used 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.

Overall, we found support for two hypotheses:

> (*H1*): AD scales lead to a much lower quality than IS ones, especially in Colombia.
> (*H3*): The use of fixed reference points for the two end points of the scale increases the quality, but only slightly.

Hypothesis 2 is also generally supported:

> (*H2*): Increasing the number of response categories (up to 11) positively affects the quality of IS scales.

It was only in Colombia in the social trust experiment that the results were not completely in line with *H2*, since the six-point scale has a higher quality than the 11-point scale. However, this may be linked to the problems encountered during the analyses of the model for this experiment. In general, therefore, 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 USA or Europe (for example, compared with results in Saris and Gallhofer 2007).

Therefore, a few recommendations can be made for future questionnaire design in Mexico and Colombia. First, researchers should prefer the use of IS scales and avoid the AD ones, which lead to much lower quality. Second, when using IS scales, researchers might prefer using 11-points, even if there is one experiment in Colombia where the six-point scale performs better. Third, fixed-reference points should be used for the two endpoints of the scales.

Nevertheless, more MTMM experiments need to be conducted in these new geographical areas. It is possible that the impact of the scale characteristics depends on the topic, so more concepts should be tested. Also, different modes of data collection should be used to check the robustness of the results, particularly in Mexico where Internet coverage is quite low. Besides, 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 obtain esti-

mates of the size of the errors or of their complement: the quality estimates. This is a crucial first step in order to be able to obtain correct estimates of the relationships of interest, even when the questionnaires have been designed very carefully. As Table 2 shows, even the best methods are far from having a quality of one. Therefore, researchers should always correct for the remaining measurement errors. 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. Therefore, more MTMM experiments are necessary in Central and Latin America. If a large number are conducted, a meta-analysis similar to that of Saris and Gallhofer (2007) could be made and Latin American countries could be included in a program like SQP. This would create the attractive situation whereby researchers could obtain estimates of the quality without having to do MTMM experiments.

References

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

Alwin, D. F. (1997), Feeling Thermometers versus 7-point Scales: Which Are Better?, in: *Sociological Methods and Research*, 3, 25, 318–340.

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

Alwin, D. F., and J. A. Krosnick (1991), The Reliability of Survey Attitude Measurement. The Influence of Question and Respondent Attributes, in: *Sociological Methods and Research*, 20, 1, 139–181.

Andrews, F. M. (1984), Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach, in: *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, in: *Psychological Bulletin*, 6, 81–105.

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, in: *International Journal of Market Research*, 50, 61–77.

DeCastellarnau, A., and W. E. Saris (2014), *A Simple Procedure to Correct for Measurement Errors in Survey Research*, European Social Survey Education Net (ESS EduNet), online: <http://essedunet.nsd.uib.no/cms/topics/measurement/> (9 November 2015).

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, in: *Psychological Review*, 67, 343–352.

Handlin, S. (2013), Survey Research and Social Class in Venezuela: Evaluating Alternative Measures and their Impact on Assessments of Class Voting, in: *Latin American Politics and Society*, 55, 1, 141–167, online: <doi:10.1111/j.1548-2456.2013.00187.x>.

Harzing, A.-W. (2006), Response Styles in Cross-National Survey Research: A 26-Country Study, in: *International Journal of Cross Cultural Management*, 6, 2, 243–266, online: <doi:10.1177/1470595806066332>.

Hofstede, G. (²2001), *Culture's Consequences, Comparing Values, Behaviors, Institutions and Organizations across Nations*, Thousand Oaks, CA: Sage Publications.

Johnson, T., P. Kulesa, Y. Cho, and S. Shavitt (2005), The Relation between Culture and Response Styles: Evidence from 19 Countries, in: *Journal of Cross-Cultural Psychology*, 36, 264–277, online: <doi:10.1177/0022022104272905>.

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

Leech, G. N. (1983), *Principles of Pragmatics*, New York: Longman.

Lenski, G. E., and J. C. Leggett (1960), Caste, Class, and Deference in the Research Interview, in: *American Journal of Sociology*, 65, 463–467.

Likert, R. (1932), A Technique for the Measurement of Attitudes, in: *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., J. Kim, C. H. Hsu, M. Papenfuss, L. Villa, E. Lazcano-Ponce, 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, in: *American Journal of Epidemiology*, 170, 965–974.

Revilla, M. (2012), Impact of the Mode of Data Collection on the Quality of Answers to Survey Questions Depending on Respondents' Characteristics, in: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 116, 44–60, online: <doi:10.1177/0759106312456510>.

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

Revilla, M., and W. E. Saris (2012), A Comparison of the Quality of Questions in a Face-to-face and a Web Survey, in: *International Journal of Public Opinion Research*, Summer, 25, 2, 242–253, online: <doi:10.1093/ijpor/eds007>.

Revilla, M., W. E. Saris, and J. A. Krosnick (2013), Choosing the Number of Categories in Agree-Disagree Scales, in: *Sociological Methods and Research*, February, 43, 73–97, first published online December 2013, online: <doi:10.1177/0049124113509605>.

Revilla, M., W. E. Saris, G. Loewe, and C. Ochoa (2015), Can a Non-Probabilistic Online Panel Achieve Question Quality Similar to that of the European Social Survey?, in: *International Journal of Market Research*, 57, 3, 395–412, online: <www.mrs.org.uk/ijmr_article/article/104501> (9 November 2015).

Saris, W. E., and F. M. Andrews (1991), Evaluation of Measurement Instruments using a Structural Modeling Approach, in: P. B. Biemer et al. (eds), *Measurement Errors in Surveys*, New York: Wiley, 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., and M. Revilla (2015), Correction for Measurement Errors in Survey Research: Necessary and Possible, in: *Social Indicators Research*, 17 June, online: <doi:10.1007/s11205-015-1002-x> (9 November 2015).

Saris, W. E., and K. De Rooij (1988), What Kind of Terms Should Be Used for Reference Points?, in: W. E. Saris (ed.), *Variation in Response Functions: A Source of Measurement Error in Attitude Research*, Amsterdam: Sociometric Research Foundation, 199–218.

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

Saris, W. E., A. Satorra, and W. Van der Veld (2009), Testing Structural Equation Models or Detection of Misspecifications, in: *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 561–582.

Saris, W. E., M. Revilla, J. A. Krosnick, and E. M. Shaeffer (2010), Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options, in: *Survey Research Methods*, 4, 1, 61–79.

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, online: <www.upf.edu/survey/_pdf/RECSM_wp024.pdf> (28 October 2015).

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

Scherpenzeel, A., and W. E. Saris (1997), The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies, in: *Sociological Methods and Research*, 25, 3, 341–383.

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.

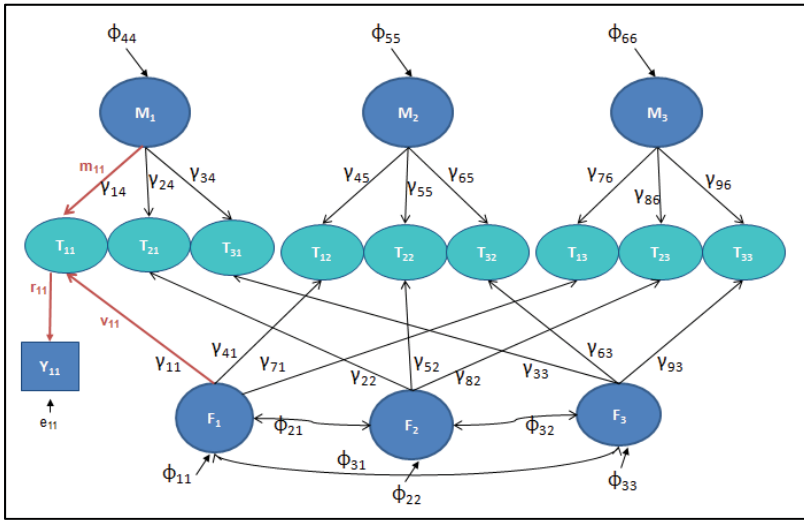
Tourangeau, R., L. J. Rips, and K. Rasinski (2000), *The Psychology of Survey Response*, Cambridge, MA: Cambridge University Press.

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: W. E. Saris and L. van Meurs (eds), *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod Studies*, Amsterdam: 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 ph 4 4 ph 5 5 ph 6 6
```

```
    value 1 ph 1 1 ph 2 2 ph 3 3
    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

La calidad de diferentes escalas en una encuesta online en México y Colombia

Resumen: La formulación de teorías e hipótesis se realiza 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 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 todavía. 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 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.

Palabras claves: América Latina, Mexico, Colombia, calidad, errores de medición, experimentos multirasgos-multimétodos