Revista Internacional de Sociología (RIS)
Special issue on methodological challenges in participation research
Vol. 70, extra 2, 65-87, diciembre 2012
ISSN: 0034-9712; eISSN: 1988-429X
DOI:10.3989/ris.2012.04.04

METHODOLOGICAL CHALLENGES FOR THE LARGE *N* STUDY OF LOCAL PARTICIPATORY EXPERIENCES
Combining methods and databases

RETOS METODOLÓGICOS PARA EL ESTUDIO CUANTITATIVO DE LAS EXPERIENCIAS PARTICIPATIVAS LOCALES
Combinación de métodos y bases de datos

Carolina Galais carolina.galais@uab.cat
Universidad Autónoma de Barcelona (UAB). Spain
Joan Font jfont@iesa.csic.es
Pau Alarcón palarcon@iesa.csic.es
Dolores Sesma lsesma@iesa.csic.es
Instituto de Estudios Sociales Avanzados (IESA-CSIC). Spain

Abstract
In this article we analyse the effects of different data collection strategies in the study of local participatory experiences in a region of Spain (Andalusia). We examine the divergences and similarities between the data collected using different methods, as well as the implications for the reliability of the data. We have collected participatory experiences through two parallel processes: a survey of municipalities and web content mining. The survey of municipalities used two complementary strategies: an online questionnaire and a CATI follow-up for those municipalities that had not answered our first online contact attempt. Both processes (survey and data mining) were applied to the same sample of municipalities, but provided significantly different images of the characteristics of Andalusia's participatory landscape. The goal of this work is to discuss the different types of biases introduced by each data collection procedure and their implications for substantive analyses.

Keywords
Citizen participation; Data collection procedures; Internet data mining; Local participation; Participatory experiences; Survey administration mode.

Resumen
En este artículo analizamos los efectos de diferentes estrategias para la recolección de datos en el estudio de las experiencias participativas andaluzas. Examinamos para ello las diferencias y similitudes entre los datos recogidos mediante diferentes métodos, así como las implicaciones para la fiabilidad de los datos. Para ello, hemos utilizado dos procedimientos paralelos. En primer lugar, una encuesta a municipios y la minería de datos en Internet. La encuesta se realizó utilizando dos modos de administración diferentes, un cuestionario *online* y un cuestionario telefónico de seguimiento a los municipios que no respondieron al primer intento de contacto vía correo electrónico. Tanto la encuesta como la minería de datos fueron aplicados a la misma muestra de municipios, aunque arrojaron diferencias significativas en cuanto a las características del panorama participativo en Andalucía. El objetivo de este trabajo es discutir los diferentes tipos de sesgos introducidos por cada procedimiento de recogida de datos y sus implicaciones para posteriores análisis sustantivos.

Palabras clave
Experiencias participativas; Minería de datos online; Modo de administración de encuestas; Participación ciudadana; Participación local; Procedimientos de recogida de datos.

Introduction[1]

Most previous attempts at providing a general picture of local participation activity have used self-administered surveys sent to municipalities (DETR 1998; Birch 2002; FEMP 2002; Ajángiz and Blas 2008). Is this a reliable strategy that can provide a good overall picture of reality? Are there other alternatives that could provide better information? In this paper we try to answer these questions through an analysis of the data collection process regarding participatory experiences completed at a local level in Andalusia, a region of Spain. To do so, we examine the divergences and similarities that arise from the comparison of two different methods (one of them with two modes of administration, making three different data sources) of collecting and coding information regarding a few hundred participatory experiences.

The first main goal of this paper is to discuss the virtues and limitations of two contrasting strategies of data collection. The first strategy used was the more traditional one, a survey of municipalities. The second was a data mining strategy using the Internet. With this aim, we conducted two parallel data collection processes that tried to capture the same reality. In addition, the survey faced a common problem related to this methodology: dealing with refusals and with the resulting moderate response rates. To address this, the first mode of administration (Computer Assisted Self Interview, CASI) was complemented with a Computer Assisted Telephone Interview (CATI) survey. This allowed us to address two subsequent research questions. First, we asked whether the differences between our surveys were a product of comparing two different sets of municipalities (the larger and more engaged with the research topic, which answered our online survey in the first place versus the remaining ones that answered the telephone survey) or whether some of the differences were the result of using two different modes of administration (CATI vs. CASI). Second, once we aggregated these two sources of data, we were able to compare them with the results from our data mining approach to learn more about the biases each of them produced on the pictures of the reality obtained.

The structure of the paper is as follows. In the next section we justify why making these comparisons is important and present the research design and data collection procedures we have used. Section 3 makes the first comparison between the two stages of the survey (CASI vs. CATI). In this first comparison, our two universes were different and, as a result, we also expect to find important differences in the characteristics of the experiences collected. We discuss whether all differences were compositional (i.e., caused by the fact that we are measuring two different parts of our final universe). The complementary explanation is that some of these differences may be the result of the

[1] A previous version of this paper was presented at the Conference "Methodological challenges in participation research", IESA (CSIC), Córdoba, November 4-5, 2011. We thank the session discussant, the participants and Donatella della Porta for helpful comments.

two modes of administration used. Section 4 moves to the comparison of the final results of both data collection procedures (survey vs. Internet-collected information). We follow the same logic as in section 3, showing the differences in a few important variables and analysing to what extent they are due to the data collection mode. Section 5 briefly presents three potential future research strategies to continue exploring the causes of the remaining differences.

Theoretical framework and research design

This paper stems from the decisions and challenges faced when gathering information on participatory experiences at the local level for the MECPALO project[2]. One of the main goals of the project is to build several regional databases of participatory experiences developed at the sub-regional level[3]. This is aimed at making a description of the characteristics of these experiences, as well as answering a series of questions related to the origins, democratic qualities and attitudinal consequences of those experiences.

We have argued elsewhere about the need to build close to local participation realities that go beyond the prevalent case study strategy (Font et al. 2011; Font and Galais 2011)[4]. However, drawing such a picture is not an easy task. Three different approaches are found in previous research. First, the selection of a limited subset of experiences that share some common organisational or territorial characteristics ("focused mappings" e.g., Schattan 2006; Sintomer et al. 2008). Second, the gathering of several varied experiences that try to capture the maximum diversity regarding those processes (Subirats et al. 2001; Della Porta and Reiter 2009). Third, the development of a survey of municipalities to obtain a list of the municipalities' responses (DETR 1998; Birch 2002; FEMP 2002). Since the first approach allows building a more reliable but also more incomplete picture of reality, we wanted to assess the advantages and problems of the two remaining strategies.

It should be noted that we lack a census of experiences. That is, there is no sampling frame with which to start. Bearing this in mind, we started by designing a representative sample of Andalusian municipalities. Andalusia has 770 municipalities, from which we selected a sample of 400. These 400 municipalities are representative of the municipa-

---

[2] MECPALO is the Spanish acronym for the project *Local participation processes in Southern Europe: causes and consequences*. The project's principal investigator (PI) is Joan Font and the research team includes researchers from three Spanish institutions, as well as a French team (PI: Yves Sintomer) and an Italian team (PI: Donatella della Porta).

[3] The universe of analysis is formed by any participatory process (from a 2 hour consultation to a stable and periodical mechanism) whose aim is to discuss local policies or issues and which has either been promoted or has gained recognition from local authorities.

[4] A similar argument has also been developed by other authors (e.g., Baiocchi et al. 2011).

lities with more than 1000 inhabitants[5]. The sample was stratified by province and city size (Font et al. 2011).

Our main unit of observation comprises experiences and not municipalities. Sampling Andalusian municipalities (those that develop participatory experiences as well as those that do not) allows us to answer additional research questions (e.g., why some councils conducted few or no participatory practices while others undertook quite a few), but this is not the aim of our particular research. In addition, our sampling strategy guarantees acceptable variability among the contextual explanatory factors.

We then designed a web-based questionnaire that addressed more than 50 questions on the existence and characteristics of participatory experiences, and sent a link to our survey to the public officials in charge of citizenship participation affairs in each of these 400 municipalities or —in the absence of this position— to the mayor. A total of 120 municipalities responded to the call after three follow-up messages, which means that the final response rate for the CASI survey was 30%. Higher response rates were obtained in municipalities with more than 20,000 inhabitants and slightly higher rates in those governed by the political party to the left of the social democrats (United Left, IU; response rate 37%). The municipalities included in the study were asked to provide up to two experiences. Some did not report any, some provided one, and a few completed the questionnaire twice, once per each experience. These experiences were transformed into our units of analysis. Considering that some municipalities had not developed a participatory process and some provided two, this made up a total of 156 experiences.

However, the response rate pointed to some of the limitations of our final sample. It is known that non-response bias may jeopardise the reliability of the portrait presented by the data, as well as the relationship between variables. In our case this is particularly true with regards to the link between city size and local government ideology. For instance, if municipalities ruled by left-wing parties were more prone to answer the survey regardless of their participatory performance, but those ruled by conservatives only answered if they had successful experiences to report, this could weaken our conclusions about the relationship between ideology and participatory initiatives.

The survey research literature has shown that when high non-response rates may jeopardise the representativeness and variability of data, a possible course of action is to switch the mode of administration (Dillman et al. 2009). Thus, we launched a second phase of the data collection process. This second phase consisted of contacting the remaining municipalities that did not answer our online survey, relying this time on telephone interviewing. Two main hypotheses about non-response drove this effort. First, for

[5] There were two reasons for choosing only a section of the 770 municipalities. First, the need to exclude the smallest municipalities that develop interesting participatory practices, but hardly formalise them and do not have the resources to publish them on a website. Second, given the large number of municipalities in the next strata (1,000 to 10,000 inhabitants) we preferred to make a sample of them and retain more resources to undertake a more intense follow-up that could lead to a higher response rate.

municipalities without an evident interest in the field of citizenship participation, dealing with a (relatively long) self-administered cuestionnaire could be a reason for skipping the survey. Using a different mode of administration that avoided any writing and a shorter questionnaire lead by an interviewer, which conveyed a sense of duty to respondents while resolving their doubts, could increase the response rate. Second, we suspected that municipalities that did not have personnel devoted specifically to participation (especially smaller municipalities) might have been more reluctant to find the appropriate person to answer the survey. Experienced interviewers could reach this person more easily. The CATI survey achieved a 62% response rate (174 municipalities of the 380 that had not answered the CASI survey). As a result, the combined CASI and CATI surveys represent a 73.5% response rate (see section 3 for more details).

The next complementary strategy was carried out in an attempt to improve two previous processes to collect participatory experiences that had been performed in relation with the MECPALO project (Della Porta and Reiter 2009; Font and Galais, 2011). We searched the net for websites of the same 400 municipalities using keywords following the common standard for web content mining (Cooley et al. 1997). We used a codebook that followed the survey questionnaire, including most of the same information. This effort resulted in a new database containing 125 experiences[6]. Previous research on web content mining agrees that the main pitfall of this data collection method is that "in the absence of a known population, a truly random sample [of relevant websites] is not possible" (Miller, Pole and Bateman 2010:4). Our work, however, avoids this flaw because it starts with a representative sample of Andalusian municipalities.

In this respect, we first compare the two parts of the survey data collected through the CASI and CATI methods. In this case, we expect to find important differences in the data since they correspond to two different subpopulations of the local universe. In addition, both CASI and CATI have strengths and weaknesses. Thus, we will probably also find differences that are the result of the two different modes of administration. In online self-administered surveys, there is no interviewer to enhance social desirability. However, this means that no one can either clarify the meaning of the answers or encourage responses (Bradburn et al. 2004). It is quite likely that CASI data contain much more item non-response (Diaz de Rada 2011) and possibly more measurement error due to the misunderstanding of more difficult questions. On the other hand, telephone interviewing may show more random measurement errors, more survey satisficing, and more social desirability response bias (Chang and Krosnick 2009). These effects may cancel each other out and result in similar data quality, thus justifying the decision to merge both surveys[7].

---

[6] Experiences where too limited information (less than 20% of the variables) was found were not included in the final database. Approximately 20 experiences fit in this category.

[7] A similar argument has been used in Diaz de Rada (2010) which shows that through the compensation of different sources of bias, the results of a combined personal and phone pre-election poll obtained better results than any of them alone.

Our second comparison is between (aggregated) survey results and our data mining search. To our knowledge, no comparisons between survey-obtained and web-collected data validity have been conducted to date. In this case, compositional effects should be more limited since the initial sampling for both strategies includes the same 400 municipalities. However, these differences should exist. First, because our survey strategy asked specifically about a maximum of two experiences per municipality, whereas the Internet search strategy would collect as many as were sufficiently documented on the web[8]. Second, both data collection procedures have their own potential problems. Surveys are affected by the most common sources of error: the questionnaire and the role the respondents play in answering them. In its turn, Internet data mining may also introduce biases coming from the search engine (visibility of the webs caused by the amount of inlinks and outlinks and user searches) and from the researcher, including the keywords selected and the interpretation and coding of the results (Hindman 2008).

We will proceed by comparing the distribution of several relevant variables in these databases, and then move forward by comparing the explanatory power of the data source (our main independent variable) in a series of multivariate analyses. For these analyses, we have selected a set of dependent variables that have been proven relevant in participation studies (Table 1). First, and as a way to approach the phenomenon of the impact of participation on politics, we will look at the number of policy phases that were actually accessible for citizens during the process. We selected this variable as a proxy for influence, i.e., the degree to which citizens were involved in the public decision (Arnstein 1971; Parés 2009). We will count the number of phases (diagnosis, programming, decision, implementation, evaluation) in which citizens had a say (Font et al. 2011). This produces a numerical variable that ranges from 0 to 5.

Next, we will consider whether the local government was the only driving force of those experiences. Some scholars have suggested that the direction of the driving forces (top-down vs. down-top) may affect the design and results and, in short, the qualities of the participatory process (Fung 2006; Della Porta 2008; Font and Galais 2011). Thus, we have generated a dichotomous variable that differentiates those experiences where civil society had played some role in proposing or organising the experience.

Inclusiveness measures the attempt to involve wide and diverse sectors of society in the process (Fung 2006; Della Porta 2008). Such inclusiveness may be pursued either by extensive mobilisation strategies oriented to achieving a large number of participants, or through the plural representation of views and opinions. Thus, we use two different variables to capture this idea. First, we use a dichotomous variable that distinguishes experiences with an open call or random selection of participants from the census vs. all possible forms of restricted call (i.e., personal invitations). Second, we include the number of participants since this is a traditional indicator of the legitimacy of the participatory component of a process.

---

[8] Seven municipalities include three or more experiences (six in one case). This means that these seven municipalities concentrate 32% of the total experiences included in the Internet-collected database.

Finally, we take into account whether we deal with a temporary experience (from a one-hour session to a two-year process) or with a stable mechanism that shows the will of the promoters to institutionalise citizens' participation.

Table 1.
Dependent variables: Dimensions, contents and response categories

Dimension	Content	Categories
Origin	Promoter/organiser of the experience	Dichotomous: Only local government vs. local government and civil society
Policies	Number of participatory policy phases	Continuous: 0 to 5
Inclusiveness	Plurality of actors	Dichotomous: Open to all or random selection vs. invitation
Inclusiveness	Number of participants	Continuous: Eight categories going from 1 (ten participants or less) to 8 (more than a thousand participants)
Stability	Temporary process or permanent mechanism	Dichotomous

A TWO-STEP SURVEY. MIXING MODES OF ADMINISTRATION

In order to offer a first glimpse of the similarities and differences of CATI and CASI data, this section begins by comparing the sampling differences between the two. It then looks for significant differences among the dependent variables mentioned above. Finally, multivariate analyses are used to test to what extent data source may bias results when explaining such participatory features.

Table 2 shows the virtues of the mix-mode administration of the final survey quite clearly. A comparison of the first and second column of the table shows that the CASI survey had significant biases in the response rates of the different categories (much higher for large municipalities, but also in the province of Cordoba or in municipalities governed by the leftist IU). In contrast, the comparison of the initial sample (first column) and the final survey results (last columns) are remarkably similar for all variables and categories. As a result, we can be quite confident that a survey with a high response rate (74%), which is distributed quite homogeneously among all sectors, does not contain significant biases regarding potential variables that need to be controlled.

On the other hand, precisely because the types of municipalities that have answered the CASI and the CATI survey are quite different, we should expect significant differences

in the type of participatory processes obtained through each of these procedures. As expected, the first important difference appears when comparing partial non-response, which tends to be much higher in the CASI method. To provide just a few examples, partial non-response was 10% versus 0% for easy questions such as having a department in charge of participation or not, and 49% versus 1% for more "difficult" questions such as the number of people working full time on participation-related activities for the CASI and CATI methods, respectively.

Graph 1 displays a pattern regarding the policy phases opened for participation that also appears in other questions. There are differences in the results found through both modes of administration, since participation in the diagnosis phase is significantly more common among the municipalities that answered the CASI survey. On the other hand,

Table 2.
Composition of initial and effective samples of Andalusian municipalities by province, city size and party

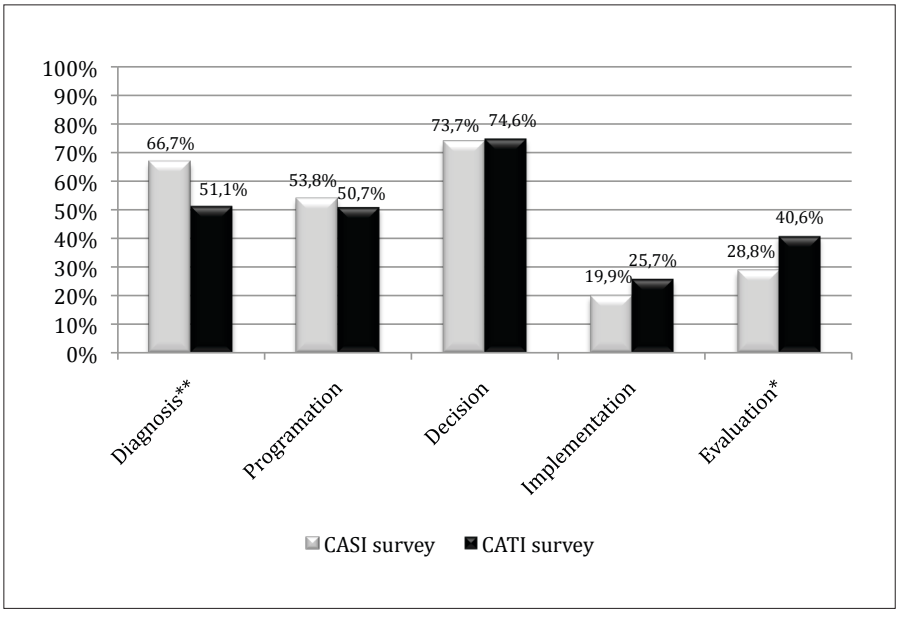
		Initial designed sample (%)	CASI Survey (%)	CATI survey (%)	Total Survey (%)
Province	Almeria	8.5	5.8	10.9	8.8
	Cadiz	8.5	10.0	8.6	9.2
	Cordoba	11	20.0	6.9	12.2
	Granada	18	14.2	20.1	17.7
	Huelva	9	5.8	9.8	8.2
	Jaen	13.5	13.3	14.4	14.0
	Malaga	12.5	12.5	12.6	12.6
	Seville	19	18.3	16.7	17.4
Inhabitants	1,000-5,000	44	34	49.4	43.1
	5,000-10,000	18.25	20	20.7	20.4
	10,000-20,000	18.25	15	16.1	15.7
	20,000-50,000	12.25	20	8.6	13.3
	+50,000	7.25	11	5.2	7.6
Political party of the mayor	PSOE	62.5	61.6	69.5	66.3
	PP	16.3	14.2	14.4	14.3
	IU	12.3	15.0	10.3	12.2
	PA	4.0	5.0	1.7	3.0
	Independent/ others	5.0	4.2	4.0	4.1
N		400	120	174	294

Sources: Survey E1107 (IESA)

participation in the decision, implementation and the evaluation phases in particular is more common in the phone survey. Such discrepancies may be due to genuine differences among the municipalities that answered the surveys and their practices, but are more likely related to response order effects (Krosnick and Alwin 1987; Tourangeau and Smith 1996). Indeed, established research states that respondents of self-administered, visual-presented questionnaires are more prone to check off the first response option presented, what is known as the "primacy effect". On the contrary, when respondents are asked questions orally, such as in face-to-face or telephone interviews, they are more prone to agree with the final option offered, a phenomenon called the "recency effect". These tendencies and likely biases that run in opposite directions are, however, likely to cancel each other out if we gather together the data collected through different administration modes.

Nevertheless, not all the differences between CASI and CATI surveys will be due to administration modes. As argued before, it is quite normal that some differences will

Graph 1.
Differences between administration modes for phases of the policy process

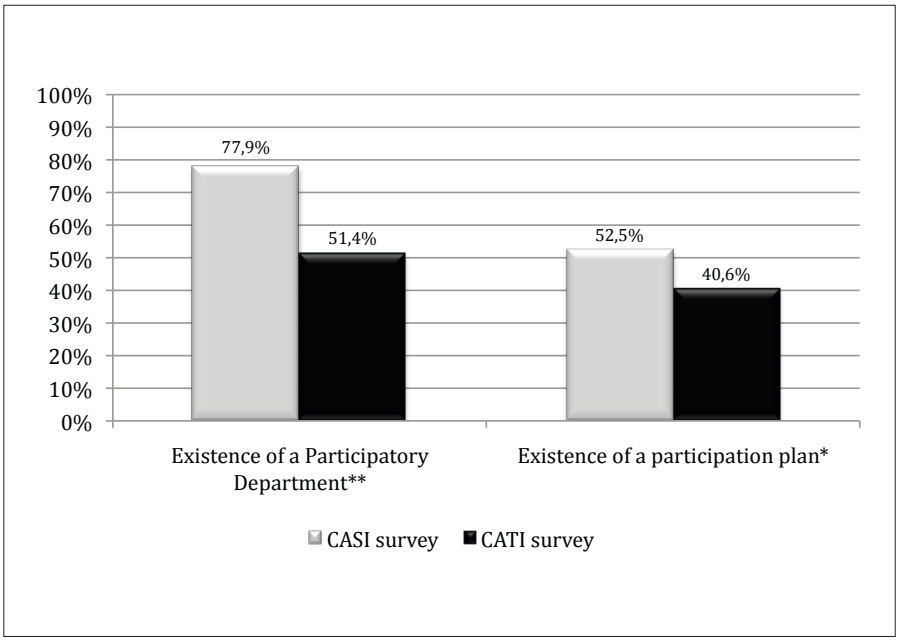


* Denotes significant differences between the averages with a significance level of 0.05
** Denotes significant differences between the averages with a significance level of 0.01
Sources: Font (2001) and Survey E1107 (IESA)
N=432

arise since the municipalities that both surveys cover are not the same. To make a first step towards disentangling the effects of the administration mode and of the composition of both samples, we have conducted a series of regression analyses of the dependent variables justified above. Regarding the relevant controls, we have taken into account city size, since it is one of the most important variables that distinguishes both samples and which is likely to have an effect on the type of participatory processes developed. The variable "inhabitants" takes values between 1 and 5, whose meaning is consistent with the categories displayed in Table 2.

Graph 2 justifies the inclusion of additional controls. As it becomes evident, the municipalities that answered the CASI survey are more likely to have a participation department and a local participation plan. These two variables are likely measuring quite a different level of a city council's engagement with citizens' participation and the resources available to deal with it. We will include them in further multivariate analyses as

Graph 2.
Institutional resources by administration mode



*Denotes significant differences between the averages with a significance level of 0.05
**Denotes significant differences between the averages with a significance level of 0.01
N=416 and 415, respectively.

dichotomous variables where 1 denotes having a participation plan or having a participation department[9].

The results of the regression analysis are displayed in tables 3 (logistic regressions) and 4 (OLS regressions). In both tables, the first column for each of the dependent variables shows the explanatory power of the mode of administration alone, whereas the second column shows the effect of the administration mode once we control for some of the important compositional variables that distinguish both populations.

Table 3.
Explanatory factors of participation characteristics: Logistic regressions

	Government as single organiser				Participation open to everyone				Stability			
	Only data source		With other variables		Only data source		With other variables		Only data source		With other variables	
	B	p	B	p	B	p	B	p	B	p	B	p
Data source: CASI	1.27	**	1.27	**	-.440	*	-.37	-	-.022	-	-.021	-
Inhabitants	-	-	.005	-	-	-	-.122	-	-	-	-.048	-
Participation plan	-	-	.007	-	-	-	.34	-	-	-	.584	**
Participation department	-	-	.1	-	-	-	.16	-	-	-	-.009	-
Constant	-.628	**	-.727	*	0.44	**	.42	-	.517	**	.396	-
R2 Nagelkerke	.111		.112		.015		.032		0		.026	
N	418		418		420		420		425		425	

*<0.05; **<0.01

Sources: Survey E1107 (IESA)

[9] In this latter case, we have also coded as 1 those municipalities that do not have a department under this designation but where another department is in charge of participatory affairs. By doing so we do not penalise small towns.

Table 4.
Explanatory factors of participation characteristics: OLS regressions

	Number of policy phases opened for participation				Number of participants (categories)			
	Only data source		With other variables		Only data source		With other variables	
	B	p	B	p	B	p	B	p
Data source: CASI	.002	-	-0.08	-	-.074	-	-.322	-
Inhabitants	-	-	0.03	-	-	-	.253	**
Participation plan	-	-	0.57	**	-	-	.528	**
Participation department	-	-	0.13	-	-	-	-.079	-
Constant	2.43	**	2.07	**	3.67	**	3.03	**
R^2	0		0.039		0		.053	
N	432		432		418		418	

*<0.05; **<0.01
Sources: Survey E1107 (IESA)

Tables 3 and 4 show different estimation models for the five dependent variables analysed. In three of the five regressions, the coefficient for our main independent variable is not significant. For instance, when data are collected through CASI surveys, the government tends to be the principal organiser of the experience, and this relationship between the administration mode and this trend of participatory experiences does not disappear after controlling for organisational resources or the size of the municipality. The fact that the experience was open to everyone also seems to be affected by the administration mode in the sense that municipalities that administered our CASI survey tend to hold restricted processes. Nevertheless, the difference is not significant once we control for city size and resources.

In summary, the use of a combined mix-mode strategy resulted in a substantially higher response rate which would spare us some biased conclusions regarding the relationship between variables. This is probably a result of combining two different administration modes as the reduction in the bias is related to the size of the municipality since large cities were keener to answer our first CASI survey. Some differences between both databases (especially regarding the role of different actors) are still present when analysing those data and considering the administration mode as an explanatory factor. Nevertheless, in some cases these differences disappear when we control for the impact

of the administration mode on some factors that determine different populations (i.e., size of municipality and organisational resources related to participation). In most cases, the differences are not significant and this becomes an encouraging starting point that reinforces the strategy of merging both datasets.

Comparing the results from two data gathering procedures. Survey versus Internet search

Our next step is to compare the survey dataset, where CASI ans CATI data have been merged, with data gathered through web mining. From now on, we will not distinguish between administration modes regarding surveys. In this case, the populations of the two datasets (survey vs. internet mining) should be more similar, since we are covering the same 400 municipalities. If differences are found, they should be the result of three main factors. Firstly, not every participatory process makes its way to being published on a website. This is evidenced by the fact that we have collected 432 experiences through the survey[10] and only 125 through the data mining strategy. Moreover, even if some experiences can be tracked through the web, not all of them are equally visible. Some of them may lack the keywords or links that allow search engines to identify and present them among the first results. Second, the coders are very different. In one case, the respondent also plays the role of coder: she must retrieve her memories and subjective perceptions and then attempt to find a correspondence with the categories available in the questionnaire. In contrast, a data mining strategy implies that researchers act as coders or instruct coders about how to translate the information provided by municipal websites into final, meaningful values. Third, the survey allowed for a maximum of two experiences per municipality, whereas the data mining strategy put no limit on this number, thus resulting in a more limited number of municipalities in the latter database[11] and a larger number of experiences per municipality[12].

The Internet data content mining process began with a careful search for key terms in the websites of our 400 sampled municipalities. The keywords successively searched

[10] If we give full credence to the survey results, the full number would be much larger since the average number of experiences acknowledged by the municipalities was around four, but we only asked them to report the details of two of them. There is also a potential bias in the selection they made of their two experiences. We therefore developed a complementary survey to analyse this possible effect, but the results are not yet available.

[11] A total of 120 municipalities were sampled by survey and around 70 by the data mining strategy. Several experiences in the latter strategy correspond to supra-local processes that cannot be attributed to a single municipality.

[12] 1.3 experiences per municipality in the survey and 1.8 in data mining.

were: *citizen participation, participatory budget, local agenda 21, citizen forum, assembly, survey, local democracy, e-government, e-democracy, strategic plan, and citizen participation department*[13]. Additionally, general web searches (Google) were made using the same keywords combined with the region's name and selecting those processes that corresponded to our 400 municipalities. In each of these searches we searched five Google pages to find information about new participatory experiences or additional information about those already collected. When we could not obtain any new information amid the first five Google results, we stopped and moved to the next concept. Information was also researched on the website of the Andalusian Federation of Municipalities (FAMP, in Spanish), which hosts a database containing a few dozen participatory experiences.

Once an experience was found, we *googled* for additional information about the experience to complete as many fields of our database as possible. If we located an experience but the information available was too limited (less than 20% of the variables), we did not include it in the final database. We then coded the information obtained using the same variables and concepts reflected in the survey, excluding subjective questions as well as a few more questions that were of limited analytical interest.

To proceed with the analysis we will compare the results of both datasets and analyse the differences between them using the same strategy as in section 2. Table 5 shows that, even if the sampling frame for both data gathering procedures (surveys and web content mining) are identical, the final results show very different landscapes for the Andalusian municipalities. Surveys provide a more similar picture to the initial sample design, in which small municipalities were prevalent. On the contrary, Internet data provide very few experiences in small municipalities. A large number of them are found in large cities of more than 50,000 inhabitants. This is probably the result of differences regarding resources, which would have an impact on the efforts devoted to the online diffusion of these experiences[14]. This triggers a serious underrepresentation of the experiences emerging from small municipalities.

These enormous differences could point to two radically different realities in all aspects. Nevertheless, the discrepancies among data collection procedures are not found for all the relevant variables in this study. In fact, the other two variables whose real distribution we know for sure are quite similar in both datasets. As Table 6 shows, only one province is significantly overrepresented by web data mining (Cadiz). Regarding the political party of the mayor, we find the exact translation of the actual party shares by city size amid the web mining data. Since the PSOE (social democratic

[13] Participación ciudadana, presupuestos participativos, agenda 21 local, foro ciudadano, asamblea, encuesta, democracia local, e-gobierno, e-democracia, e-participación, plan estratégico, concejalía participación ciudadana.

[14] Most small municipalities have a website, but they are extremely simple and contain limited information. Most often this information deals more with the locality and its attractiveness rather than with local policy.

Table 5.
Number of inhabitants across data source

	Survey (%)	Internet data mining (%)	Differences
<5,000 inhabitants	41.1	10.5	-30.6*
5,000-10,000 inhabitants	21.1	18.1	-3
10,000-20,000 inhabitants	14.8	14.3	0.5
20,000-50,000 inhabitants	14.6	22.9	8.3*
>50,000 inhabitants	8.6	34.3	-26*
Total	100	100	

*Denotes significant differences between categories based on a two-sided test with a significance level of 0.05
N=537 (432 surveys; 105 data mining). Experiences involving more than one municipality are excluded (N=20).
Source: Font (2001), Survey E1107 (IESA) and Internet data mining Andalusia 2011 (IESA).

Table 6.
Political party of the mayor and province by data source

		Survey (%)	Internet data mining (%)	Differences
Political party	PSOE	68.8	60	-8.8
	PP	12.3	21	8.7*
	IU	13.2	17.1	3.9
	PA	1.6	0	-1.6
	Independent	4.2	1.9	-2.3
	N	432	105[1]	
Province	Almeria	8.3	6.6	-1.7
	Cadiz	9.0	19.7	10.7*
	Cordoba	12.7	13.1	0.4
	Granada	17.8	13.9	-3.1
	Huelva	7.2	4.1	-3.1
	Jaen	14.4	21.3	6.9
	Malaga	13.4	9.8	-3.6
	Seville	17.1	11.5	-5.6
	N	432	122[2]	

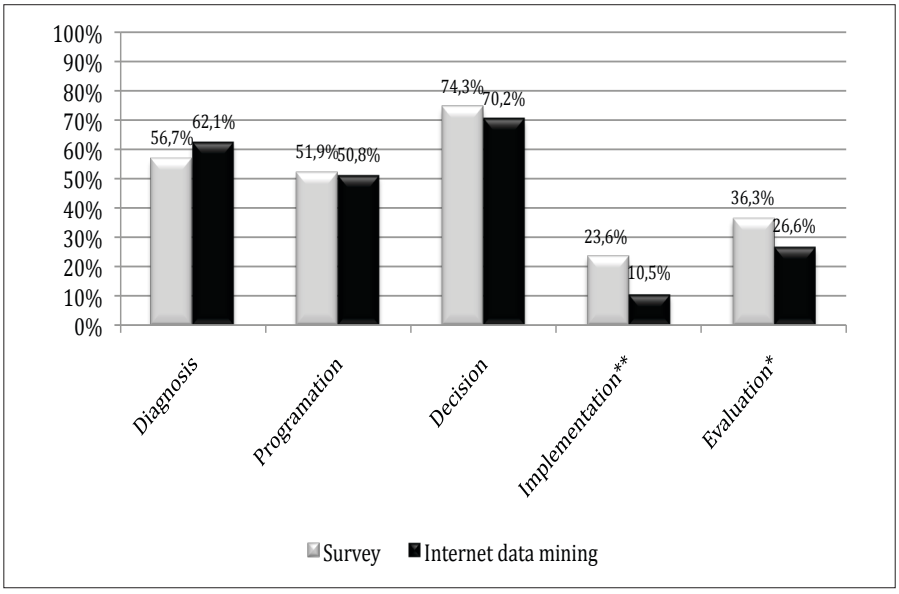
* Denotes significant differences between categories based on a two-sided test with a significance level of 0.05
Sources: Font (2001), Survey E1107 (IESA) and Internet data mining Andalusia 2011 (IESA).
[1] Excludes supra-local experiences, where no single government party can be identified (N=20).
[2] Excludes experiences that affected more than one province (N=3).

party) prevails in small municipalities, the experiences developed by this centre-left party are more important in the survey-collected database, where small municipalities are also more prevalent. We find exactly the opposite pattern regarding the conservative PP party (Table 6). Thus, even if we could apparently be facing an ideology bias, this is probably due to the underrepresentation of small municipalities when an Internet data mining strategy is adopted.

Graph 3 displays the differences between the two sources regarding the same variable shown in the previous section, namely the policy phases opened to participation. As can be observed, the implementation and evaluation phases are significantly less present among experiences collected through Internet data mining.

This shows again that data sources are potentially relevant and, as a result, we will proceed with the same analytical strategy: to predict the same five dependent variables using the data source as our main independent variable and city size and resources as

Graph 3.
Differences between administration modes for phases of the policy process



*Denotes significant differences between the averages with a significance level of 0.05
**Denotes significant differences between the averages with a significance level of 0.01
Sources: Font (2001), Survey E1107 (IESA) and Internet data mining Andalusia 2011 (IESA)
N=557

controls. Tables 7 and 8 show the regression coefficients for these variables. As can be seen, there are substantive differences between Internet-collected experiences and survey data for all our dependent variables. Experiences collected by means of web mining are less prone than survey-collected experiences to crystallise as permanent mechanisms and to open this kind of process to several policy phases. Moreover, they are also more likely to be driven exclusively by local governments. However, they are more inclusive as they are more open to everyone and able to mobilise more participants. Controlling for the number of inhabitants and/or the most important local resources does not make these effects disappear.

Table 7.
Explanatory factors of participation characteristics. Survey vs. Internet mining differences (logistic regressions)

	Government as single organiser				Participation open to everyone				Stability			
	Only data source		With other variables		Only data source		With other variables		Only data source		With other variables	
	B	p	B	p	B	p	B	p	B	p	B	p
Data source: Internet mining▲	3.6	**	3.6	**	0.5	*	1.2	**	-1.4	**	-1.2	**
Inhabitants	-	-	0.1	-	-	-	-0.1	*	-	-	-0.28	-
Participation plan	-	-	0.1	-	-	-	0.3	-	-	-	0.64	**
Participation department	-	-	-0.05	-	-	-	0.2	-	-	-	-0.02	-
Constant	-0.18	**	-0.4	-	0.3	**	0.3	-	0.52	**	0.35	-
R2 Nagelkerke	0.29		0.27		0.01		0.06		0.11		0.09	
N	543		523		543		523		550		530	

▲ The reference category for the variable "Internet mining" is "survey".
*<0.05; **<0.01
Sources: Font (2001), Survey E1107 (IESA) and Internet data mining Andalusia 2011 (IESA)

Table 8.
Explanatory factors of participation characteristics. Survey vs. Internet mining differences (OLS regressions)

	Number of policy phases opened for participation				Number of participants (categories)			
	Only data source		With other variables		Only data source		With other variables	
	B	p	B	p	B	p	B	p
Data source: Internet mining	-.247	-	-0.6	**	.787	**	.620	*
Inhabitants	-	-	0.05	-			.234	**
Participation plan	-	-	0.49	**			.565	**
Participation department	-	-	0.13	-			-.061	-
Constant	2.4	**	1.97	**	3.644	**	2.92	**
R2	.003		0.06		.023		.075	
N	556		536		490		472	

*<0.05; **<0.01

Sources: Font (2001), Survey E1107 (IESA) and Internet data mining Andalusia 2011 (IESA)

Both the leading role of the local government and the permanent character of a participatory process are probably highly correlated with its likelihood to become noticeable and, as a result, to reach the Internet. Some stable mechanisms, such as participatory budgeting, have been introduced in recent decades and their attractiveness and intensity have produced considerable media and web visibility. On the contrary, other permanent mechanisms that follow the logic of sectoral or territorially-based consultation councils have existed for years and do not attract much media attention. Even when they make their way to municipal websites, the documentation they provide tends to be insufficient to consider them a valid case[15].

[15] Very often only the internal regulation of their composition can be found, but no information on their real dynamics. This pattern is only broken in very large cities like Madrid or Barcelona that offer much richer information on these mechanisms.

The policy phases are very different and this has implications in the coding process. For instance, it would be difficult that two external observers or coders disagree on whether to qualify a particular participatory process as stable or as limited in time. However, the number of policy phases open to citizen participation is not so clear. Coding them becomes difficult even for qualified coders and their contents are difficult to understand for respondents that are not accustomed to thinking about this issue. We cannot claim that we know why these differences arise, but it is clear that these two questions are extremely different and that policy phases are a good candidate to expect a high respondent/coder influence.

Finally, the two aspects where the Internet processes score better (the two dimensions of inclusiveness) probably highlight the filters that help participatory mechanisms to make their way onto the web. The less interesting or spectacular processes, where politicians only try to avoid conflicts or to give legitimacy to sectoral policies through consultation with a limited network of actors would not reach local websites. Alternatively, we could consider that these practices are not interesting enough to give them visibility or to produce a participation section on the local website. However, when a municipality is surveyed, they will look at any practice they have developed to avoid appearing to be a passive administration. In the next section we present further strategies that could be pursued in order to confirm or discard these hypotheses.

Prospects for further research[16]

Throughout the paper we have pointed to several explanations for these differences between data collected by means of surveys and data gathered through internet data mining. How important are each of them? Are there possible research strategies to disentangle their relative weight? Basically, we have pointed to two main reasons why the results could be different: because the universes they reflect are not identical or because the people who have translated the reality into codes have used different criteria. We want to briefly sketch four alternative strategies (two dealing with the different universes and two dealing with the role of the coder) that could contribute to understanding where these differences come from and their methodological and substantive implications. The first one is a strategy based on automated data content mining. In order to overcome the biases introduced by the professional zeal of researchers, several scholars suggest the use of automated tools such as indexing software (Zafarani et al. 2008). These tools automatically harvest all the information from a text or website and count the words and

[16] Some of these strategies have been partially developed and have contributed to checking data mining quality. Others have been developed, but a detailed analysis cannot be presented here due to reasons of space.

semantic roots. It is necessary to subsequently classify this information into meaningful categories, but this avoids missing some information due to fatigue, for instance. *Issue discovery* software, for example, lists the words and lexemes mentioned more than once in a website. Of course for this or any other indexing outcome to be meaningful, it is still necessary that the researcher read and process all the information available, but these tools may help overcome some coding biases.

A second alternative would be to have an in-depth, qualitative look at both raw databases, identify the cases that correspond to the same experiences and compare the coding reached by each of these methods, paying special attention to which kind of variables offer greater differences and why it may happen. In our case, 28 experiences are "repeated" in our primary raw records. That is, they were found through the Internet and also through the survey. Even if this does not constitute a large sample, it may be plausible to perform some tests to identify the variables that show larger discrepancies between the two kinds of coders (survey respondents/researcher).

These two strategies focus on how the coding procedures could have produced different results, whereas our third and fourth alternatives would explore the reasons why certain experiences are more likely to reach the databases using one or another data collection procedure. The third alternative would focus on the potential bias introduced by the survey respondents when they choose their two experiences, since they could have chosen the most interesting ones or those where they were more intensely involved. To explore this possibility, we selected those municipalities that had indicated in the original CASI survey that they had developed more than two experiences, but had only given the details about two of them. In a new telephone survey we asked them to give details about two additional cases, now selected through a more objective criterion: the most recent ones. A comparison with the originally provided experiences will allow us to examine this potential source of bias.

In addition, our fourth strategy could deal with the gatekeeper role of people in charge of local websites and how they can bias which experiences will or will not be published. For example, in comparing both databases we could check whether the processes that have been developed by the mayor and not by a specific sectoral department have a higher probability of reaching the web or whether alternative patterns of selection could be identified.

Conclusion

There is no single perfect method for capturing the reality of local participation processes. Using a large *N* strategy to analyse this reality may be fruitful and necessary, but it implies selection and standardisation problems that are not easy to solve.

Surveys addressed to institutions also have problems related to non-response and social desirability. In this case, the use of a mixed administration mode strategy allowed us to achieve a much higher and less biased response rate than the single initial usage

of an Internet survey alone. However, problems related to social desirability and others such as the potential effects of the order of response categories or the difficulty of understanding response categories remain.

The alternative strategy of data mining has various flaws. In this case, the most crucial difference is that only a small part of the participatory experiences that appear in the survey have made their way to the Internet and some of them are so poorly documented that they cannot be studied. As a result, a data mining strategy produces a significantly different picture, where experiences developed in large cities that devote more resources to their websites are largely overrepresented. This problem may be less important in other countries where very small municipalities are less common than in Spain.

In any case, differences between the two strategies are not solely due to the bias introduced by Internet visibility. Both procedures start from different available information (the memories of respondents and the documents they want to consult when they answer versus the official reports that are found on the Internet), but in particular, they are interpreted and coded by different types of individuals that give different meaning to the variables handled in our study.

The use of careful comparisons, double-checking or alternative processes of data researching may help to understand where the differences come from and contribute to making data from different sources more comparable. Nonetheless, they will continue to provide different pictures of reality and we should be aware of the limitations and biases that each of them introduces. Even with these limitations, the picture provided by the data from any of the sources/methods of data collection discussed is more accurate than the one that is provided by the prevalent research design in the literature, namely, selecting a few case studies because they are the most successful available experiences.

References

Ajángiz, R. and A. Blas. 2008. *Mapa de mecanismos y experiencias de participación ciudadana en el País Vasco*. Vitoria: Servicio de Publicaciones del Gobierno Vasco.

Arnstein, S. 1971. "A ladder of citizen participation in the USA." J*ournal of the Royal town Planning Institute* 57:176-1982.

Baiocchi, G., P. Heller and M. K. Silva. 2011. *Bootstrapping Democracy. Transforming Local Governance and Civil Society in Brazil*. Stanford: Stanford University Press.

Birch, D. 2002. *Public participation in local government. A survey of local authorities*. London: Office of the Deputy Prime Minister.

Bradburn, N.M., Sudman, S.,and B. Wansink. 2004. *Asking questions: The definitive guide to questionnaire design*. San Francisco: Jossey-Bass.

Chang, L. and J.A. Krosnick. 2009. "National Surveys via RDD Telephone versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73:641-78.

Cooley, R., Mobasher, B., and J. Srivastava. 1997. "Web mining: Information and pattern discovery on the World Wide Web." *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, Newport Beach, CA, November 1997.

Della Porta, D. 2008. "La partecipazione nelle istituzioni: concettualizzare gli esperimenti da democrazia deliberativa e partecipativa." *Partecipazione e Conflitto* 0:15-42.

Della Porta, D. and H. Reiter. 2009. *Gli esperimenti di democrazia deliberativa e partecipativa in Toscana*. Istituto Universitario Europeo e Regione Toscana.

DETR. 1998. *Enhancing public participation in local government –a research report*. London: DETR. Retrieved December (12), 2011 (http://www.local.dtlr.gov.uk/research/particip.htm.)

Díaz de Rada, V. 2010. "Comparación entre los resultados proporcionados por encuestas telefónicas y personales: el caso de un estudio electoral." *Colección Opiniones y Actitudes* 66. Madrid: CIS.

Diaz de Rada, V. 2011. "Encuestas con encuestador y autoadministradas por internet ¿Proporcionan resultados comparables?" *Revista Española de Investigaciones Sociológicas* 136:49-90.

Dillman, D. A., Phelps, G. Tortora, R. D., Swift, K., Kohrell J. and J. Berck 2009. "Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response, and the Internet." *Social Science Research* 38:1-18.

FEMP 2002. "La participación ciudadana en los ayuntamientos." *Memoria de la investigación sobre la implantación del área sus recursos, su desarrollo orgánico, los niveles y ámbitos de participación y su entorno asociativo.* Zaragoza: Departamento de participación ciudadana. Departamento de estudios. Retrieved October 20, 2011 (http://aragonparticipa.aragon.es/attachments/218_Participacion%20Ciudadana%20en%20los%20Ayuntamientos%20%28FEMP%29.pdf)

Font, J. and C. Galais. 2011. "The qualities of local participation: The explanatory role of ideology, external support and civil society as organizer." *International Journal of Urban and Regional Research* 35:932-48.

Font, J. et al. 2011. *Democracia local en Andalucía. Experiencias participativas en los municipios andaluces.* Sevilla: Junta de Andalucía.

Fung, A. 2006. "Varieties of participation in complex governance." *Public Administration Review* 66: 66-75.

Hindman, M. 2008. *The Myth of digital democracy*. Princeton, NJ: Princeton University Press.

Holbrook, A.L., Green, M.C. and J.A. Krosnick. 2003. "Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability." *Public Opinion Quarterly* 67:79-125

Krosnick J.A *and D. Alwin. 1987. "*An evaluation of a cognitive theory of response-order effects in survey measurement*." Public Opinion Quarterly* 51:201-*219.*

Miller, E.A., A, Pole,. and C. Bateman. 2010. "Variation in health bog features and elements by gender, occupation, and perspective." *Paper prepared at the 106 th Annual Meeting of the American Political Science Association*, September 2-5, 2010. Available at SSRN: http://ssrn.com/abstract=1649644.

Parés, M. (comp.) 2009. *Participación y calidad democrática. Evaluando las nuevas formas de democracia participativa.* Barcelona: Ariel.

Schattan, V. 2006. "Democratization of Brazilian health councils: the paradox of bringing the other side into the tent." *International Journal of Urban and Regional Research* 30:656-671.

Sintomer, Y., C. Herzberg and A. Röcke. 2008. "Participatory budgeting in Europe: potentials and challenges." *International Journal of Urban and Regional Research* 32:164-178.

Subirats, J., I. Blanco, J. Brugué, J. Font, R. Gomà, M. Jarque and L. Medina. 2001. *Experiències de participació ciutadana en els municipis catalans.* Barcelona: Escola d'Administració Pública de Catalunya.

Tourangeau R, and T.W. Smith. 1996. "Asking sensitive questions: the impact of data collection mode, question format, and question context." *Public Opinion Quarterly* 60:275-304.

Zafarani, R., M. Jashki, H. Baghi, B. Hamidreza and A. A. Ghorbani. 2008. "A novel approach for social behavior analysis of the blogosphere." Pp. 356-367 in *Canadian* Ai edited by S. Bergler. Berlin: Springer-Verlag.

Carol Galais is a postdoctoral fellow at the Université de Montreal. She holds a PhD in political science (2008, UPF) and is a member of the Research group Democracy, Elections and Citizenship at the Universitat Autonoma de Barcelona (UAB).

Joan Font is a senior researcher at the IESA/CSIC working on citizen participation in public policies. He was the research director at the major public survey institution (CIS) in 2004-2008. He has been a senior lecturer at the Political Science department of UAB (Barcelona) and a visiting scholar at the EUI (Florence) and UCD (Dublin).

Dolores Sesma is a research assistant at the Institute for Advanced Social Studies (IESA-CSIC). She completed a Master degree in public policies (2008, UCM) and a Master in Social Science and Health (UB). She is also specialized in data analysis (CIS, 2007).

Pau Alarcón holds a predoctoral scholarship from the Ministry of Science and Innovation and is researching at the IESA/CSIC. He holds a Bachelor of Sociology (UCM), a Diploma of Statistics (UMH) and a Master in Contemporary Latin-American Studies (UCM). He has been a visiting researcher at the University of Sydney.

RECEIVED: 04 April 2012
ACCEPTED: 27 June 2012