

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Multi-mode Data Collection: What Can Still Be Expected?

by Marina Signore, Giovanna Brancato, Giorgia Simeoni

2009



Multi-mode Data Collection: What Can Still Be Expected?

Marina Signore, Giovanna Brancato, Giorgia Simeoni¹

Abstract

A major issue in official statistics is the availability of objective measures supporting the based-on-fact decision process. Istat has developed an Information System to assess survey quality. Among other standard quality indicators, nonresponse rates are systematically computed and stored for all surveys. Such a rich information base permits analysis over time and comparisons among surveys. The paper focuses on the analysis of interrelationships between data collection mode and other survey characteristics on total nonresponse. Particular attention is devoted to the extent to which multi-mode data collection improves response rates.

Key Words: Multi mode data collection, Response rate, Business surveys.

1. Introduction

The purpose of this work is to evaluate the impact on quality of Istat data collection strategy, taking into account for the confounding factors. The data for the analyses are taken from Istat centralised quality information system named SIDI and described in section 2. Using metadata from SIDI, a description of the data collection systems currently used at Istat is provided in section 3 and 4. The different combinations of data collection modes, pre-contact and follow-up modes are presented and commented together with an analysis of the changes occurred in the last decade. The SIDI system provides both metadata for a descriptive analysis and the quality indicators for quantitative analyses and modelling. As described in section 5, logistic regression models for correlated data were applied on the response probability. The results of our analyses allow us to draw general conclusions and some specific considerations for business surveys.

2. Istat centralised quality documentation system

Istat developed the SIDI centralised information system in order to enhance and support quality activities. It manages standard quality indicators (SQIs) and related process metadata for all Istat surveys (Brancato et al., 2004, Brancato et al., 2006). The integrated examination of SQIs and process metadata allow the user i) to better understand the SQIs of a single survey; ii) to make comparisons among different surveys with the same relevant features; iii) to assess the impact of process improvements as well as top management strategies for improving quality. Therefore, SIDI is a quality tool for a variety of users, such as survey managers, quality managers, Istat top management, other producers of official statistics (i.e. agencies of the Italian statistical system, other National Statistical Institutes, Eurostat and other international organisations), and external users.

In the last decade, Istat top management promoted the use of web surveys, particularly in addition to mail for business surveys. To this purpose, Istat developed a dedicated website called Indata which allows the respondent to fill in electronic forms online as well as to upload forms filled in offline (Balestrino and Fanfoni, 2006). The use of web data collection mode has proved to be effective for single surveys, especially with regard to the increased timeliness and quality of process operations and costs saving, and also for the reduction of response burden in terms of questionnaire completion time (Zeli, 2008).

¹ Marina Signore, Istat, Via Cesare Balbo 16, Rome, Italy, 00184, signore@istat.it; Giovanna Brancato, Istat, Via Cesare Balbo 16, Rome, Italy, 00184, brancato@istat.it; Giorgia Simeoni, Istat, Via Cesare Balbo 16, Rome, Italy, 00184, simeoni@istat.it

However, so far the effectiveness of Istat strategy has never been assessed. This paper aims at assessing the improvements in terms of quality obtained by using mixed data collection modes with particular regard to the advantages of web and mail mixed modes for business surveys. To this purpose, we exploited the qualitative and quantitative information managed in the SIDI system. The information in SIDI has been systematically collected for all Istat surveys since 2002 and provides objective and comparable measures for the based-on-fact analyses and decision process.

In particular 201 primary surveys (surveys from administrative sources; direct surveys; or partly from administrative sources and partly direct surveys) are documented in SIDI as well as 100 secondary studies or statistics compilations. The analyses presented in this work are restricted to the subset of direct surveys and partly from administrative sources and partly direct surveys for a total of 110 statistical processes. For our purposes, we focused on metadata and quality indicators related to the data collection phase. Survey characteristics of interest that are extensively documented are: the data collection mode, pre-contact and follow-up methods and their changes over time as well as the type of responding units, the data collection periodicity, and the length of the questionnaire. Such metadata, together with SQIs, were used to perform both exploratory and model-based analyses. In addition, the SIDI system manages a set of standard quality indicators that enables to assess the quality of the data collection phase in terms of response rates as well as to analyse the main reasons for nonresponse (e.g. refusal, no-contacts, and so on). SQIs on the data collection phase are available for approximately half of the surveys for which such a set of indicators is meaningful. For the analyses of this paper, data from 61 surveys were used, corresponding to a total of 460 different survey editions².

In the following table, some characteristics of the 110 Istat statistical processes of interest are reported.

Table 2-1
Surveys distribution according to some characteristics

	Characteristics							
	Periodicity			Reporting unit			Regulations	
	Yearly	Infra-yearly	Multi-yearly	Household	Institution	Enterprise	Yes	No
Number	39	37	34	17	26	67	50	60
%	35.5	33.6	30.9	15.5	23.6	60.9	45.5	54.5

As can be noted, the majority of Istat surveys are conducted on enterprises and almost half are ruled by European regulations. Cross-tabulations of these variables show that the majority of household surveys are conducted on a multi-year basis, and that 86% of regulations concern businesses surveys (73% of short term business statistics and 57% of structural ones are ruled by regulations).

3. Shifts from single to multi data collection modes

In the last decade, several changes in data collection mode were observed. Similarly to other countries, also in Italy, there was a tendency to switch from single to multi-data collection modes. This occurrence concerned different types of data collection modes and was supported by the advantages offered by the computer assisted techniques.

Table 3-1
Shifts from single to multi mode (1998-2008)

From	To	Surveys
PAPI	PAPI, CAPI, CATI	1
	PAPI, questionnaire delivery and collection	1
CATI	CATI, CAWI	1
Mail	CAWI, mail	11
	Mail, e-mail	3
	Disk by mail, mail	2
	CAWI, mail, e-mail	1

² Each survey calculates the set of SQI for each survey edition according to its periodicity.

The majority of changes concerned mail business surveys as a result of the already mentioned Istat strategy that promoted the use of the web site Indata as primary data collection means. Thanks to a good feedback from enterprises 12 business surveys added the web mode to the mail and one survey uses it now as unique data collection mode while some new surveys chose the web as a single mode (see Table 4-2). More in detail, the shifts started on a regular basis from 2003 with a peak of shifts from mail to mail plus other modes in 2006.

4. Multi-mode systems at Istat

In order to study the effects on survey quality of mixed or multi data collection modes, several factors should be taken into account. On the one hand, the use of multi data collection modes permits to reduce costs and response burden and can improve coverage, survey participation and timeliness. On the other hand, mode-effect can be an additional source of measurement error. De Leeuw (2005) proposed a comprehensive approach introducing the concept of “multi-mode system” and a clear classification scheme that took into account not only the actual data collection mode, but also the pre-contacts and the follow-ups. For each combination identified, the strengths and weaknesses in term of effects on survey quality were also pointed out. Following this approach, single mode surveys that, before or after actual data collection, contact units with a means of communication that is different from the one used during data collection should be considered “multi-mode system” surveys. In general terms, the use of multi-mode systems exploits the benefits of using different means of contacting units without the risk of increasing measurement errors. Considering the actual data collection phase, different types of multi or mixed modes can be distinguished with different impacts on survey quality. For example, surveys offering respondents the choice among different means to answer the questionnaire (e.g.: mail and web) adopt a *concurrent* mixed data collection mode that presents some advantages as cost and burden reduction, but its effectiveness in term of response rate improvement should still be definitely proved and its impact on measurement error should be taken into account. In our analyses we focused on the evaluation of the effect of multi mode on non-response. Measurement error due to mode-effect has not been assessed here but it is considered a relevant issue for future research.

In the following the multi-mode system approach is applied to Istat situation. Advance letters are sent via mail by Istat to population or sample units for almost every survey. When municipalities are involved in the data collection, they are also in charge of sending advance letters. The advance letters inform the units on survey objectives, grant data confidentiality and provide contact information. Their positive effect on response rates is largely acknowledged (Biemer and Lyberg, 2003; Campoprese et al., 2000). In PAPI, CAPI and CATI surveys, in addition to the advance letters, telephone pre-contacts are carried on with the aim of setting appointments for the interview. As a consequence of these practices, all Istat surveys whose data collection mode is different from mail can be considered multi-mode systems.

As reported in Table 4-1, follow-ups are carried out in about 55% percent of current surveys. Focusing on CAWI and/or mail surveys, this percentage grows up to 76%. Almost all surveys performing follow-ups use several means (e.g. mail, telephone, fax), with a very high predominance of the telephone one. This leads to consider two third of single mode mail surveys as multi mode systems.

In the following we decided to focus on the actual data collection mode, rather than considering the complete data collection system. This choice is justified by the spread of different communication means before and after data collection in Istat surveys, that would not allow us to differentiate them.

Table 4-1
Follow-up use and means, Istat current surveys*

Follow-up use and means	Number of current surveys	Percent of current surveys
Telephone and other means	50	50.5
Mail alone or mail and e-mail	2	2.0
Fax alone	1	1.0
E-mail alone	1	1.0
No follow-up	45	45.5
Total	99	100

* Censuses and occasional surveys are excluded

Table 4-2 reports the data collection mode distribution for Istat current surveys. About 47% of single mode surveys are by mail, and a remarkable part (about 26%) uses CATI mode. This predominance is observed also taking into account only business surveys. Considering multi mode data collection surveys, the more common combinations (about 71%) are mail and CAWI techniques or mail and other computer assisted means in a concurrent way. These techniques are basically adopted for business surveys, whereas PAPI and questionnaire delivery and collection³ (about 17%) are usually adopted for household surveys.

Table 4-2
Istat current surveys* distribution by data collection mode

Single mode		Multi mode: 2 techniques		Multi mode: 3 techniques	
Types	Surveys	Types	Surveys	Types	Surveys
Mail	30	CAWI, mail	17	CAWI, mail, e-mail	2
CATI	17	PAPI, questionnaire delivery and collection	6	PAPI, CATI, CAPI	1
E-mail	6	Mail, e-mail	3	CAWI, mail, Disk by mail	1
CAWI	5	CATI, CAPI	1	Mail, e-mail, Disk by mail	1
PAPI	4	CAWI, CATI	1		
CAPI	1	CAWI, e-mail	1		
Questionnaire delivery and collection	1	Mail, Disk by mail	1		
Total	64	Total	30	Total	5

*Censuses and occasional surveys are excluded

5. Logistic regression modeling on response probability

The analysis of the effects of survey characteristics on unit non-response started with the evaluation of the trend over the last 10 years of the Istat surveys' response rate mean. The response rate was computed as the ratio of the *respondent units* to number of *in-scope* and *unresolved units* (Brancato et al., 2004; Hidiroglou et al., 1993). Overall, the trend resulted quite stable and not decreasing, differently from what observed in other statistical agencies (Petroni et al., 2004) and reported in de Leeuw (2005). This led us to investigate in detail the impact of other survey characteristics on response probability. To this aim, logistic regression models were applied to the data. The response rate was the outcome variable. As independent factors, the following variables were tested: the data collection mode, being a multi vs. a single data collection mode survey, the type of reporting unit (household, enterprise, institution), the survey periodicity, the size of the sample or the population to be investigated, the out-of-scope rate⁴, performing follow-ups or not, the calendar year, the survey design (total survey, random sampling, non random sampling) and the questionnaire length in terms of number of pages. The out-of-scope rate was considered because the indicator selected to represent the response probability does not take into account for the amount of these errors. In particular, an inverse correlation between the two quality measures was expected. Each survey contributed to the analysis with all its available indicators and metadata. However, it has to be mentioned, that the metadata, i.e. most of the independent variables included in the model, are rather stable over time. In any case, surveys shifting from single to multi data collection mode, contributed to the model with the response rates and data

³ In this mode, the interviewer performs a face to face interview, leaves an additional questionnaire or a diary to the unit and collects it back after some days.

⁴ Computed as the ratio of the number of *out-of-scope units* to the number of *resolved units* (Brancato et al., 2004; Hidiroglou et al., 1993).

collection modes for the two periods. This approach leads to the use of non-independent observations (within survey correlation). In order to derive correct inference when correlated responses are present, as it was the case, estimation of regression parameters was obtained by Generalized Estimating Equations (GEE) (Diggle et al., 2002). This approach adopts a *marginal model*, in which the average response of observations sharing the same covariates is modeled as a function of the covariates.

In GEE logistic regression model specification, although its robustness property to errors in the correlation specification, a relevant issue is the selection of an appropriate working correlation structure, to estimate the regression parameters more efficiently (Ballinger, 2004). In our application, an *exchangeable* correlation matrix was used, that reflects an absence of logical ordering for observations within clusters, since there was not evidence of the existence of a trend requiring an ordering over time. In this case, the correlation is estimated as a sum of the standardized residual products. In practice, it is assumed that the within-survey observations are equally correlated. Different correlation structures were also tested, providing similar results, and therefore addressing towards the hypothesis of a mild intra-survey correlation. The unstructured option, i.e. the estimate of all possible correlations within-survey responses, led to un-estimability of the model.

In Table 5-1, the results in terms of estimate, standard error, significance level, Odds Ratio (OR) and 95% confidence limits of the OR are reported. At this stage of the analysis, all the factors were tested in simple logistic regression models. When considering the detailed data collection mode, the model estimates an almost ten times higher response probability in PAPI as compared to mail surveys (OR = 9.93). CAWI or e-mail single mode surveys are associated with much higher response probability respect to mail. In our data, adding other computer assisted modes to the mail data collection mode is not associated to better performance in terms of responses. However, it has to be noticed that there is not evidence in the literature on nonresponse reduction associated to offering more options (de Leeuw, 2005). If the data collection variable is summarised into a more aggregate one, considering only being a multi mode versus being a single data collection mode survey, the regression parameter is still significant, reflecting an almost two times higher response probability in the first group. As expected, institutions and households are associated with better response rates. Unexpectedly, infra-yearly surveys are those with the highest response probability. The greatest is the sample/population to be surveyed, the hardest is to obtain its participation. As expected, the out-of-scope rate is inversely correlated to the response probability, as not expected, performing follow-ups is not associated to better response rates. Other variables that were tested and did not reach the significance level are: year of the survey, survey design and number of pages of the questionnaire.

Table 5-1
Logistic models on response probability, fitting one factor at time (all units, GEE)

Parameters	Estimate	S.E.	p	OR	95% Confidence limits	
Data collection mode: ref. Mail						
PAPI	2.30	0.63	<.005	9.93	2.79	35.33
CAWI or e-mail alone	2.03	0.14	<.0001	7.65	5.80	10.10
CATI, CAPI	1.54	0.18	<.0001	4.67	3.24	6.72
PAPI, delivery/collection	1.24	0.19	<.0001	3.47	2.37	5.08
CATI	0.15	0.20	0.44	1.17	0.79	1.73
Mail and other CA modes	0.10	0.10	0.32	1.10	0.91	1.33
Data collection mode: ref. Single						
Multi	0.55	0.27	<.05	1.73	1.01	2.97
Reporting unit: ref. Enterprise						
Institution	1.27	0.19	<.0001	3.57	2.43	5.25
Household	0.98	0.27	<.005	2.66	1.55	4.57
Periodicity: ref. Yearly						
Infra-yearly	1.43	0.44	<.005	4.19	1.74	10.10
Multi-yearly	0.34	0.30	0.26	1.41	0.77	2.58
Sample/population size: ref. N ≥ 10000						
N < 1000	1.02	0.34	<.005	2.79	1.40	5.53
1000 ≤ N < 5000	0.58	0.28	<.05	1.78	1.01	3.14
5000 ≤ N < 10000	0.30	0.29	0.30	1.35	0.76	2.42
Out-of-scope rate	-0.04	0.01	<.01	0.96	0.94	0.99
Follow-up: ref. Yes						
No follow-up	0.75	0.31	0.01	2.11	1.15	3.89

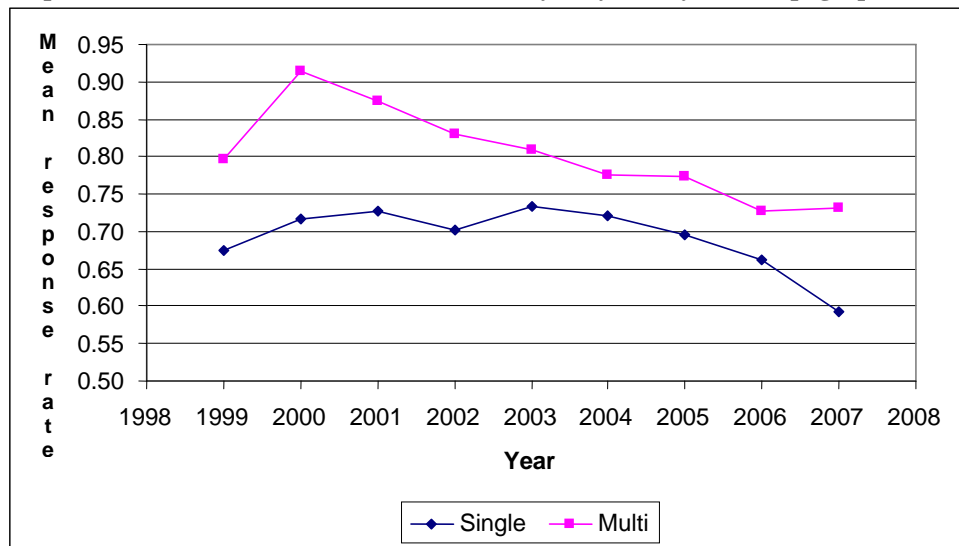
Some of these preliminary results were expected, others were suspicious, as for example that on the follow-up variable or the non significance of the length of the questionnaire. This might be due to the heterogeneity of the population under study. For example, the follow-up is not performed by surveys on households, that are those with higher response rates. Similarly, household surveys have in general long questionnaires, and this may dilute the overall effect on the response rate. This was as well confirmed by the unsuccessful attempts to obtain a good multiple regression model, where a great number of interaction terms was required to improve the fitting. We came out with the conclusion that, studying together such different populations - enterprises, households, and institutions - would require complex models, with a great number of interactions, increasing difficulties for interpreting the results and need for a considerably great sample size. Therefore we decided to concentrate our attention to the specific subpopulation to which Istat data collection strategy was more oriented, represented by the enterprises.

Table 5-2 shows the results of the best model fitted on the subgroup of the enterprises. The model includes: to be a multi or single data collection mode survey, the survey calendar year, the interaction between these two variables, the survey periodicity and the length of the questionnaire. First of all, the calendar year, that previously did not reach the significance level, is now significant. As a result, whereas it cannot be said that the response rates are decreasing over time for all Istat surveys, a slight decreasing trend is observed in the subset of the enterprises. The multi/single variable is very significant, telling us that, still, after adjusting for the other factors, the response probability in multi mode is considerably higher than in single data collection mode surveys. Also the interaction between the calendar year and the multi-single data collection was significant, reflecting different slopes in the two groups. Differently from the previous model on the total of the units, the infra-yearly surveys do not have a greater response probability than the yearly ones, whereas the multi-yearly are those with the best results. When studying the enterprises, the length of the questionnaire, that previously was not significant, became significant. The variable population/sample size was also almost significant, and in particular having to survey more than ten-thousand-units, resulted in an OR of 0.39 respect to smaller sample sizes (less than one-thousand units). However, the introduction of this factor did not change the size of the other parameters of the model, therefore, to keep the model parsimonious, it was not included in the final one.

Table 5-2
Logistic models on response probability, best model (enterprises, GEE)

Parameters	Estimate	S.E.	p	OR	95% Confidence limits	
Data collection mode: ref. Single						
Multi	0.63	0.10	<.0001	1.89	1.55	2.29
Year: ref. 1999						
2000	0.20	0.04	<.0001	1.23	1.13	1.33
2001	0.25	0.05	<.0001	1.28	1.16	1.41
2002	0.13	0.05	0.01	1.14	1.03	1.26
2003	0.28	0.07	<.0001	1.33	1.15	1.53
2004	0.21	0.05	<.0001	1.24	1.12	1.38
2005	0.10	0.04	0.02	1.10	1.01	1.20
2006	- 0.06	0.04	0.18	0.94	0.86	1.03
2007	- 0.36	0.05	<.0001	0.70	0.63	0.78
Interaction Data collection mode*Year (not reported)						
Periodicity: ref. Yearly						
Infra-yearly	0.06	0.29	0.84	1.06	0.59	1.91
Multi-yearly	0.37	0.11	<.001	1.45	1.16	1.81
# questionnaire pages	- 0.13	0.01	<.0001	0.88	0.86	0.91

Figure 5-1
Response rates (model estimates) for two infra-yearly surveys with 1-page questionnaire



Finally, in Figure 5-1, the mean response rates by year for two hypothetical infra-yearly surveys having a 1-page questionnaire, being the first single-mode and the second multi-mode, are shown, according to the estimated model. It can be noticed that, also adjusting for the other factors, still the multi mode survey consistently obtains better response rates than the single mode one, although for both surveys the trend is decreasing.

6. Conclusions

There is not evidence, at Istat, of a general decrease of response rates in surveys, with the exception of business surveys, where a decline was observed over the last decade. Multi mode data collection is always associated with better response rates, even when some relevant nonresponse factors are taken into consideration, such as the reporting unit or the size of the population or sample that is surveyed. The most interesting results are obtained in

the analyses performed on the enterprises, where the use of mixed modes is helping in limiting response rates decline. In this area, the multi data collection mode is mainly represented by the mail and web.

It can be concluded that in Istat business surveys, the mixed mode represents a way to improve survey quality not only in terms of participation, but also with respect to accuracy and timeliness. In fact, it is generally assumed that electronic questionnaires, that can incorporate edit rules, provide better quality data than paper questionnaires. Indeed, the use of web surveys contributes to reducing the costs of printing, mailing, data entry and editing and imputation as well as saving time and resources in data processing.

A very important issue, that has received growing attention at European level, is the reduction of response burden in particular to enterprises. Following Tarnai and Paxson (2004), we can state that “giving respondents multiple ways of responding can help to reduce response burden”, since they offer more options to the respondents. Istat experience gives an empirical confirmation in the sense that multi mode data collection began as a concurrent choice of mail or web mode but after one or two survey replications, the enterprises were provided only with the preferred mode (either mail or web). In particular, an important Istat survey, the *Survey on Enterprise Accounting System*, directed to all enterprises with more than 100 employees, shifted in the 2003 edition from mail to mixed mail and web mode with very positive results, and since 2006, adopted exclusively the web data collection mode.

In the literature, several studies have analysed the relationships between data collection modes and response rates even though no general conclusions were achieved. The availability of time series of quality indicators related to different surveys, allowed us to draw conclusions for the strategic sector of mail and mail plus web business surveys at Istat and to assess the validity of Istat policy in promoting the use of Indata. The extensive use of multi data collection modes might give rise to some concern on mode effect, i.e. the measurement error associated to the use of different concurrent data collection modes and this can represent an area for future research investigations.

References

- Balestrino, R. and Fanfoni, L. (2006). The Quality of Production Processes and the Electronic Survey System: The Experiences of ISTAT's Structural Business Statistics, *Proceedings of the European Conference on Quality in Survey Statistics*.
- Brancato, G., Carbini, R., Pellegrini, C., Signore, M. and Simeoni, G. (2006). Assessing Quality through the Collection and Analysis of Standard Quality Indicators: The Istat Experience *Proceedings of the European Conference on Quality in Survey Statistics*.
- Brancato, G., Pellegrini, C., Signore, M. and Simeoni, G. (2004). Standardising, Evaluating and Documenting quality: the implementation of Istat information system for survey documentation – SIDI, *Proceedings of the European Conference on Quality and Methodology in Official Statistics*
- Ballinger, G.A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis, *Organizational Research Methods*, 7, 127-150.
- Biemer, P.P. and Lyberg, L.E (2003). *Introduction to Survey Quality*, Hoboken, NJ: Wiley.
- Camporese, R., Quattrociochi L. and Sabbadini, L.L. (2000). Strategie di Prevenzione delle Mancate Risposte e degli Errori Non Campionari nelle Indagini Multiscopo, *Scritti di Statistica Economica, Quaderni di discussion n. 19*.
- De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys, *Journal of Official Statistics*, 21, 233-255.
- Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of longitudinal data*, 2nd edition, Oxford, UK: Oxford University Press.

- Hidiroglou, M.A., Drew, J.D. and Gray, G.B. (1993). A Framework for Measuring and Reducing Nonresponse in Surveys, *Survey Methodology*, 19, 81-94.
- Petroni, R., Sigman, R., Willimack, D., Cohen, S. and Tucker C. (2004). Response Rates and Nonresponse in BLS and Census Bureau Establishments Surveys, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 4159-4166.
- Tarnai, J. and Paxson, M.C. (2004). Survey Mode Preferences of Business Respondents, *Proceedings of American Association for Public Opinion Research*, American Statistical Association
- Zeger, S.L. and Liang, K.-Y. (1992), An Overview of Methods for the Analysis of Longitudinal Data, *Statistics in Medicine*, 11, 1825-1839.
- Zeli, A. (2008), A Mixed-mode Survey using Mail and Web Questionnaires: the SCI case, unpublished report, Rome, Italy: Istat.