COOPERATION ON MULTI-MODE DATA COLLECTION (MMDC)

MIXED MODE DESIGNS FOR SOCIAL SURVEYS - MIMOD

GRANT AGREEMENT FOR AN ACTION WITH MULTIPLE BENEFICIARIES

AGREEMENT NUMBER – 07112.2017.010-2017.786

Report containing the results of the applications of selected methods on mixed-mode social surveys

Experimenting methods to asses and adjust mode effect when a single mode control survey is available as a benchmark: a case study on the Italian "Aspects of daily life" survey

WP2 - Deliverable 3

Date: November 30th, 2018

Claudia De Vitiis
Francesca Inglese
Alessio Guandalini
Marco D. Terribili
Roberta Varriale

WP2: Mode bias/mode effects and adjustment for mode-effects

Contents

Summary

This report is the third deliverable of Work Package 2 (WP2) of the Grant Agreement on Mixed Mode Design for social surveys (MIMOD). In the context of the objective of the WP2, which focuses on mode effect, this report presents a set of analyses for assessing and adjusting mode effect in a specific survey context. The methods considered are framed in the review of the methodologies reported in the first deliverable of WP2 (Buelens, Van den Brakel and Schouten, 2018), containing an overview of the literature on mode effect in mixed mode surveys.

Mixed mode introduces several issues that must be addressed, both at the design phase and at the estimation phase, in order to ensure the accuracy of the estimates. The surveys based on mixed mode must be designed and realised, in fact, keeping in mind the constraints that the produced estimates must be consistent and comparable with the analogue ones obtained in the previous survey editions, for ensuring that changes in the time series are exclusively due to real changes of the observed phenomenon and not to changes in the data collection methods, suspected to be responsible of mode effect.

As described in Deliverable 1, mixed mode simultaneously generates nonresponse error (selection effects) and measurement error (measurement effects). Selection effects occur when different types of respondents choose different modes to complete the survey. The occurrence of a selection effect is in itself not a problem but it makes a mixed mode design valuable. Measurement effects refer to the influence of a survey mode on the answers respondents give, such that one person would give different answers in different modes. Put differently, measurement effects are caused by differences in measurement errors. The major problem of mixed mode designs is that selection and measurement effects are confounded and appropriate inference methods to evaluate mode effect are needed.

In particular, several methods to assess mode effect can be applied when experimental designs are planned for mixed mode surveys. This work focuses on the methods which can be applied for the assessment and adjustment of mode effect in a survey setting where an independent single mode survey is carried out together with a mixed mode survey.

The proposed analyses are applied to the experimental situation of ISTAT "Multipurpose Survey on Households - Aspects of daily life - 2017". In the 2017 edition, the mixed mode was used for the first time as a web technique was added to the traditional PAPI technique in a sequential design. A parallel single mode PAPI design was planned to allow for an assessment of mode effect on two independent samples collected with different techniques. This experimental design is different from the experimental context of Deliverable 2 of WP2 which is based on a re-interview design.

The experimental design of ISTAT survey allows for the application of some methods to disentangle selection and measurement effects on the basis of auxiliary information that is assumed to be mode insensitive, acquired from registers or collected by the survey itself. The goal of the present analyses is the evaluation of the impact of the switching to mixed mode on the estimates in a specific survey context which has to produce a variety of indicators to satisfy both national and European information needs.

For this purpose, methods to assess the impact of mixed mode on the accuracy of the estimates are applied aiming at evaluating different components of the total non-sampling error: the response and the representativeness of the two samples are evaluated through the analysis of the different nonresponse processes and representativeness indicators; models to disentangle and estimate the measurement error and selection effect in the mixed mode sample are experimented, also taking the single mode survey as a benchmark. Finally, a comparison is made between estimates obtained using different methods for adjusting mode effect (weighting/calibration and multiple imputation).

The results of the applied methods highlight that the mixed mode design catches better the overall population resulting more "representative" than the single mode design. When the assessment of mode effect is carried out for specific variables, the results can generally provide an explanation for breaks in the series of estimates due to both selection and measurement effect. The detection of measurement effects can provide an

useful advice for the planning of future edition of the survey, in order to exploit positively the coverage improvement deriving from the mixing of techniques.

The set of the analyses applied in this context can be considered as a possible list of subsequent steps, usable by researchers of other NSIs to carry out an assessment of mode effect in similar situations.

1. Introduction

This report is the third deliverable of Work Package 2 (WP2) of the Grant Agreement on Mixed Mode Design for social surveys (MIMOD). In the context of the objective of the WP2, which focuses on mode effect, this report presents a set of analyses for assessing and adjusting mode effect in a specific survey context.

Mixed mode has been adopted generally both to contrast declining response and coverage rates and to reduce the cost of the surveys. The use of different data collection techniques, including generally the cheapest ones as the web interview, helps, in fact, in contacting different types of respondents in the most suitable way for each of them.

Anyway, mixed mode introduces several issues that must be addressed, both at the design phase and at the estimation phase, by assessing and adjusting the bias effects (mode effect) due to the use of mixed mode (MM), in order to ensure the accuracy of the estimates. The surveys based on MM must be designed and realised, in fact, keeping in mind the constraint that the produced estimates must be consistent and comparable with the analogue ones obtained in the previous survey editions, for ensuring that changes in the time series are exclusively due to real changes of the observed phenomenon and not to changes in the data collection methods, suspected to be responsible of mode effect.

As described in Deliverable 1 of WP2 (Buelens, Van den Brakel and Schouten, 2018), mixed mode simultaneously generates nonresponse error (selection effects) and measurement error (measurement effects). Selection effects occur when different types of respondents choose different modes to complete the survey. The occurrence of a selection effect is in itself not a problem but it makes a mixed mode design valuable. Measurement effects refer to the influence of a survey mode on the answers respondents give, such that one person would give different answers in different modes. Put differently, measurement effects are caused by differences in measurement errors. The major problem of mixed mode designs is that selection and measurement effects are confounded and appropriate inference methods to evaluate mode effect are needed.

In particular, methods to assess mode effect can be applied when experimental designs are planned for mixed mode surveys. This work focuses on the methods which can be applied for the assessment and adjustment of mode effect in a survey setting where an independent single mode survey is carried out together with a mixed mode survey.

The proposed analyses are applied to the experimental situation of ISTAT "Multipurpose Survey on Households - Aspects of daily life - 2017". In the 2017 edition, the mixed mode was used for the first time as a web technique was added to the traditional PAPI technique in a sequential design. A parallel single mode PAPI design was planned to allow for an assessment of mode effect on two independent samples collected with different techniques. The experimental design of this survey allows somehow, to disentangle selection and measurement effects, on the basis of auxiliary information that is assumed to be mode insensitive, acquired from registers or collected by the survey itself.

In this context, methods to assess the impact of MM on the accuracy of the estimates were applied, following a study framework developed in subsequent steps: the response and the representativeness of the two samples are evaluated through the analysis of the different nonresponse processes and representativeness indicators; models to disentangle and estimate the measurement error and selection effect in the MM sample are experimented, also taking the single mode survey as a benchmark. Moreover some attempt of adjusting for mode effect are made, both through weighting/calibration and multiple imputation.

The set of the analyses applied can be considered as a possible list of subsequent steps, usable by researchers of other NSIs to carry out an assessment of mode effect in similar situations.

The report is organised as follows: section 2 outlines the framework of the analyses carried out to assess and adjust the mode effect, while section 3 describe the survey context in which the analyses of mode effect were carried out. Section 4 focus on the comparison between the two samples (SM and MM), section 5 is

dedicated to the assessment of mode effect in the MM sample and section 6 describes some experiments of adjusting for mode effect. Section 7, finally, discuss the results and outlines some conclusions.

2. Outline of the analyses of the mode effect when a control survey is available

The analyses presented are useful for evaluating the impact on the quality of survey estimates deriving from switching to mixed mode design when an experimental setting is defined with a control single mode survey carried out independently.

The experimental context where the proposed analyses can be applied is summarised in the following scheme:

General Survey Contest	Experimental : Parallel independent samples (single mode SM, mixed mode MM)
Main goal of the analyses	Evaluation of the switching from single to mixed mode, Evaluation of total non-sampling (measurement) error components
Theoretical context	Counterfactual approach
Available auxiliary information	Demo-social covariates from Register
Phases of the analyses	– Comparison between the single mode and mixed mode samples – Evaluation of the mode effect in the MM design – Adjusting for mode effect

The objective of the analyses is to evaluate first the impact on the estimates of the survey of the introduction of mixed mode design with respect to the previous single mode design and, subsequently, to analyse in depth the reasons that determine significant differences in the estimates obtained with the two designs.

Following this scheme, a study on the ISTAT "Multipurpose Survey on Households - Aspects of daily life - 2017" (ADL survey) data was developed on several levels of analysis:

1. the first level is based on the comparison between the two samples SM and MM;
2. the second level addresses the evaluation of the mode effect (selection and measurement) in the samples of respondents web and PAPI in the MM design;
3. the third level carried out some experiments to adjust for mode effect.

In the first level of analysis, tests were performed on the differences in the estimates calculated on the two sample, SM and MM, for a set of relevant survey variables, with the aim of highlighting the variables for which a suspect of mode effect was significant (paragraph 4.1).

Subsequent analyses were conducted to study the bias caused by the total nonresponse in the two samples. To this end, auxiliary variables acquired from archives on individuals were redefined at the household level because the household as a whole is involved in the response and in the "choice" of the mode . The response processes were analysed (paragraph 4.2) and the indicators of representativeness were evaluated (paragraph 4.3) in order to identify differences (especially in terms of magnitude of the bias) that could explain the differences in the estimates of the survey produced with the SM and MM samples. The different composition of samples determined by the differences in the total nonresponse processes could contribute to generate differences in the estimates, due to selection effect (error of non-observation). In the analysis, in fact, a fundamental aspect taken into account is that estimates are affected by total response differently in the two samples, generating different selection effect. In general, the analysis and treatment of total nonresponse in MM survey is a complex operation due to the particular way in which the response process is developed. In fact, in a sequential design the distribution of the sample of respondents of the follow-up phase depends on the results of the response process that is realized in the first phase with the web technique.

Part of this first step of analysis was also the evaluation of the bias introduced by total non-response with respect to a benchmark estimate (paragraph 4.4). Moreover, to estimate the measurement and selection effects in the MM sample, a method that takes the single mode survey as a benchmark is experimented (paragraph 4.5).

In the step 2, the analysis of the mode effect in the MM sample was carried out taking into account the complexity of the problem and an appropriate theoretical reference context. Methods were used that make the samples of respondents to the web and PAPI techniques comparable. The propensity score (Rosenbaum and Rubin, 1983), has been applied to study the selection effect and the measurement effect of some target variables of the survey (paragraph 5.1).

The equivalence of the measurements in the MM survey is analyzed with the diagnostic method multi-group confirmatory factor analysis (MCFA). The correspondence of the measurement model used to represent a "behavioral model" for subjects who responded with web and PAPI techniques, and of the mean level of the latent factors useful for measuring the phenomenon with the two techniques was tested. The MCFA has been carried out after controlling for selection effect and after carrying out an exploratory analysis for the identification of the latent structure of the phenomenon (paragraph 5.2).

In step 3 some experiments of adjusting for mode effect have been made. In particular, the calibration on fixed proportions of web and PAPI responses has been applied in order to stabilize the total measurement error over time (Buelens 2015). Moreover, in a counterfactual perspective, a method of multiple imputation has been applied. Alternative estimates of the main parameters of the survey have been obtained and compared with those produced by the other methods of adjustment.

In the following scheme the phases and the methods considered in the study are listed.

	Method	**Objective**	**Assumptions/Conditions**
First phase	1) Tests on the differences in the estimates calculated on the two sample for a set of relevant survey variables	Highlighting the variables for which a suspect of mode effect was significant	Independence between the two samples
	2) Tests on the response rates in the SM and MM sample. 3) Indicators of representativeness 4) Tests on the differences on estimates of benchmark variables known for selected sample units	Analysis of the response processes and evaluation of the bias caused by the total nonresponse	Independence between the two samples; MAR assumption for the response models
	5) Instrumental variable approach	Disentangling measurement and selection effects	Representativity assumption
Second phase	6) Propensity score	Disentangling measurement and selection effects	MAR assumption for the response models; Balancing assumption
	7) Multi-group confirmatory factor analysis	Analysis of the equivalence of the measurements in surveys	Identification of the latent structure of the phenomenon
Third phase	8) Weighting methods as propensity score, calibration	To adjust selection effect	Ignorability of selection mechanism; Measurement error negligible
	9) Mode calibration	To stabilize the total measurement error	Invariance over time of measurement error
	10) Multiple imputation (standard)	To adjust measurement effect	MAR assumption

3. Survey context

The "Survey on Aspect of daily life" (ADL survey) is part of the integrated system of Multipurpose Surveys on households, which started in 1993 with the aim of producing information on individuals and households. Through this survey several thematic areas are investigated, from an individual and household point of view. Information content can be grouped into four large areas:

- family, home and area where people live;
- health conditions and lifestyles;
- culture, sociality and free time activities;
- interaction between citizens and services.

Among the information gathered about culture, sociality and leisure activities, there are those on the degree of satisfaction of individuals for certain aspects of life (family and friendship relationships, health, economic situation, free time and work), on subjective well-being (satisfaction for life on the whole) and on the degree of interpersonal trust. In the section dedicated to the family instead there are the questions on the perception of the economic situation and the main problems of the area in which they live.

The survey involves each year a selected sample of about 24.000 households (of which a set of around 18.000 respondent households are interviewed, corresponding to around 38.000 individuals), concentrated in nearly 850 Italian municipalities through a two stage sample design. The sample of households is selected from the centralized municipal register.

In the 2017 edition of the ADL survey, a mixed mode technique was introduced for the first time. A web technique has been added to the traditionally used PAPI technique in a sequential design: an online questionnaire that can be self-compiled by respondents (WEB technique) or, alternatively, direct interviews with a questionnaire on paper, administered by an interviewer (PAPI technique). To fill in the online questionnaire, sample households use the credentials given in the inviting letter sent by ISTAT. If the family did not complete the questionnaire on the web, at the end of the time period for online filling, a municipal interviewer address personally the same questionnaire to all its members.

Fort the first occasion of the MM design, in order to analyse the impact of the mixed mode on the estimation of the parameters of interest, a survey design was made to randomly divide the sample of each municipality into two sub-samples: to the first one, of larger size, the mixed web/PAPI technique has been administered sequentially (mixed mode, MM design); to the second, only the PAPI interview has been proposed (single mode SM, control sample).

The overall response rate was 71% for the single mode sample and 74% for the mixed mode sample, for which the web response rate was 26.8%. A considerable regional variability of the response rates emerged, mainly for the web participation , as it is shown by the following table.

Table 1. Survey on Aspect of daily life: Response rates for the single mode e mixed mode samples

Geographical area	Response rate		
	Single mode (PAPI)	Mixed mode (web/PAPI)	
		web	Final
Nord west	65.9%	32.5%	71.2%
Nord est	70.2%	36.0%	73.6%
Center	68.6%	27.8%	70.2%
South	79.3%	17.7%	79.4%
Islands	71.3%	17.3%	74.2%
ITALY	71.0%	26.8%	74.0%

The first comparison of the estimates, obtained applying on the overall respondent sample the usual estimation procedures, with the estimates of the same survey referred to previous years, highlighted some relevant differences and the need of in-depth assessment and, possibly, some adjustments.

After the survey, administrative data were linked to the data set of selected sample individuals (respondent and non-respondent)in order to obtain external auxiliary information not affected by mode effect, to be used for the analysis of mixed mode effect (and nonresponse).

The administrative data base (DB) is that one of the Archimede Project, (Integrated archive of economic and demographic micro data, Garofalo, 2014, Ballabio *et al.*, 2018) built for expanding ISTAT information provided by administrative archives to producing longitudinal paths and cross-sectional collections of micro data to be made available to different users. This objective is achieved through the exploitation of administrative database information contents integrated into ISTAT platform SIM (Integrated Micro data System).

The linkage between the theoretical sample and the Archimede DB was performed through the individual code and allowed to get the following auxiliary variables not available from the selection register:

- Education level: below/equal/above high school diploma;
- Occupation type: employed, self-employed, not in labour age;
- Tax income.

From the selection register the following variables were derived:

- Citizenship: Italian/Foreign;
- Household type: one-component under 35, one-component 35-64, one-component over 64, two-components at least one under 35, two-components all over 34, more than two components at least one under 25, more than two components all of them over 24;
- Municipal type: Metropolitan cities, metropolitan area, other municipalities with <2000, 2000-10000, 10000-50000, >50000 inhabitants.

In the analyses described in the following paragraphs all the listed variables were used as mode insensitive covariates for the models, even if not all of them turned out to be useful. Moreover, we did not consider the occupation type, because of the large amount of missing values in the administrative file.

4. The comparison between the single and the mixed mode survey

4.1. Hypothesis testing on differences between SM and MM estimates

A preparatory analysis on MM in the Aspects of Daily Life (ADL) was carried out for testing the differences in the estimates of the main parameters of the survey under the two sample (SM and MM).

The considered questions are related to categorical answers. The following answers are collected through a self-compiled questionnaire on each individual:

- Life's satisfaction;
- Health conditions;
- Reading books in the last year;
- Habitual smoker;
- Use of internet;
- Trust in others;
- Main meal;
- Number of times in the last year you went to theatre;
- Number of times in the last year you went to cinema;
- Number of times in the last year you went to museums or exhibitions;

- Number of times in the last year you went to concerts of classical music, opera;
- Number of times in the last year you went to other music concerts;
- Number of times in the last year you went to sport events;
- Number of times in the last year you went to disco or clubs;

The following variables are collected through an interviewer. Some of them are observed on each individual:

- Frequency of seeing friends;
- Continuity in sport activity;

while these are observed at household level:

- Economic situation with respect to the previous year (household level);
- Household economic resources level.

Two tests for evaluating the differences in the estimates have been conducted:

- chi-squared-test to determine whether there is a significant difference between the distribution of the answer with respect to the data collection mode;
- t-test to determine whether the difference between proportions of individuals for each modality and for each categorical variables is significant with respect to the data collection mode.

Assume to have a binomial data[1] gathered with two different data collection modes, they can be presented in a contingency table that, in our case, might look like:

Variable *Y*	**Data collection mode**		
	SM	**MM**	**Total**
Successes	S_1	S_2	$S = S_1 + S_2$
Failures	F_1	F_2	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

where SM population is binomial (n_1, p_1), with S_1successes and F_1 failures, and MM population is binomial (n_2, p_2), with S_2successes andF_2failures.

The frequencies in the previous table can be compared with those in the expected frequency table:

Variable *Y*	**Data collection mode**		
	SM	**MM**	**Total**
Successes	$\frac{n_1 S}{n_1 + n_2}$	$\frac{n_2 S}{n_1 + n_2}$	$S = S_1 + S_2$
Failures	$\frac{n_1 F}{n_1 + n_2}$	$\frac{n_2 S}{n_1 + n_2}$	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

The statistic

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{\left(S_1 - \frac{n_1 S}{n_1+n_2}\right)^2}{\frac{n_1 S}{n_1+n_2}} + \cdots + \frac{\left(F_2 - \frac{n_1 F}{n_1+n_2}\right)^2}{\frac{n_1 F}{n_1+n_2}},$$

that measures the difference significance from the expected frequency when the answer is independent on the data collection mode, can be used for testing the independence with respect to the data collection mode. Expected frequencies are those ones we expect in the sample if the null hypothesis holds. Therefore, the chi-squared test is defined as

[1]It can be easily extended to polytomous data.

$$\begin{cases} H_o: & \chi^2 = 0 \\ H_1: & \chi^2 > 0 \end{cases},$$

that is the null hypothesis is equal to 0 because all the expected cell frequencies are equal to the observed ones. The null hypothesis is rejected when the statistic$\chi^2$is larger than the critical value of χ^2distribution at level$1 - \alpha$ with $(s-1)(t-1)$ d.f.. That is

$$\text{reject } H_o \text{ if } \chi^2 > \chi^2_{1-\alpha,(s-1)(t-1)}$$

where s and t are respectively the number of rows and columns in the contingency table and $1 - \alpha$is the confidence level. Larger is the value of χ^2 more is the evidence against H_o.

The t-test is used to compare two population proportions such as the proportion of individuals with a generic modality i (i $= 1, \ldots$ I) on Y variable with respect to the data collection mode. The test is defined as:

$$\begin{cases} H_o: & p_{1i} - p_{2i} = \mu_0 \\ H_1: & p_{1i} - p_{2i} \neq \mu_0 \end{cases},$$

where p_{1i} is the proportion of individuals with a generic modality i on the Y variable (i $= 1, \ldots$ I) interviewed with single mode (SM), p_{2i}that one interviewed with mixed mode (MM) and $\mu_0$is a fixed value, equal to 0 in this case. Note that $\sum_{i=1}^{I} p_{1i} = 1$ and $\sum_{i=1}^{I} p_{2i} = 1$.

The statistic

$$t = \frac{(\hat{p}_{1i} - \hat{p}_{2i}) - \mu_0}{\hat{S}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

under the null hypothesis is distributed as t-Student with $n_1 + n_2 - 2$ d.f.[2] and

$$\hat{S}_p = \sqrt{\frac{n_1[\hat{p}_{1i}(1-\hat{p}_{1i})] + n_2[\hat{p}_{2i}(1-\hat{p}_{2i})]}{n_1 + n_2 - 2}}.$$

Therefore, the null hypothesis is rejected when the value of the statistic is bigger than the critical value of t distribution at level $1 - \alpha$ with $n_1 + n_2 - 2$d.f.. That is

$$\text{reject } H_o \text{if } |t| > t_{1-a,\, n_1+n_2-2}.$$

The t-test is replicated for each modality i (i $= 1, \ldots$ I). It is important to point out that there is a relation between these two tests. In fact when, for all the modalities of the Y variable, the difference in proportion are not significant the Chi-squared test accepts the null hypothesis (it means there is no dependence with respect to the data collection mode). On the contrary, when at least one modality of the Y variable for which the difference is significant, the Chi-squared test rejects the null hypothesis (it means there is dependence with respect to the data collection mode).

Figure 1 shows the synthesis of the obtained results, listing the variables for which a significant difference between the estimated distributions obtained from two samples is evidenced with the two tests. The details of the analysis are reported in the Appendix A.

[2] For large n_1 and n_2, t is distributed as a standardized Normal.

Figure 1. Result on "Survey on Aspect of daily life" (ADL survey) for testing the differences in the estimates of the main parameters under the two sample single mode (SM) and mixed mode (MM).

Variable	Significant [a]
Life's satisfaction	YES
Health conditions	YES
Reading books in the last year	YES
Habitual smoker	NO
Use of internet	YES
Trust in others	YES
Main meal	YES
Number of times in the last year you went to theatre	YES
Number of times in the last year you went to cinema	YES
Number of times in the last year you went to museums or exhibitions	YES
Number of times in the last year you went to concerts of classical music, opera	YES
Number of times in the last year you went to other music concerts	YES
Number of times in the last year you went to sport events	YES
Number of times in the last year you went to disco or clubs	YES
Number of times in the last year you went to archaeological sites or monuments	YES
Frequency of seeing friends	YES
Continuity in sport activity	NO
Economic situation with respect to the previous year (household level)	NO
Household economic resources level	YES

[a] YES, if the tests show a significant difference in the answer under the two data collection mode; NO, otherwise.

The tests applied to the considered set of questions show that the two samples (MM and SM) seem to refer to two different populations with strongly different features. In fact, just for "Habitual smoker" the differences are no significant for the two samples.

For "Trust in other" and "Main meal", without considering the non-response (NR) as modality, the two sub-sample are equal. Instead, considering NR as modality, the difference are significant. This is probably due to the structure of the web questionnaire that provides the modality "Non response" for skipping the question, while in the PAPI questionnaire a blank is classified as "Non response".

With respect to the section of the questionnaire compiled by the interviewer there are no significant differences for "Continuity in sport activity" and the "economic situation with respect to the previous year between the two samples. While there are difference in "Friends frequentation" and "Household economic resources level". Therefore, the interviewer effect, probably mainly due to the social-desirability, depends to the considered question. In the considered case, the two questions on social life and economic resources portraits a worst situation in MM than in SM due exactly to social-desirability.

4.2. Test of differences between response rates

Having so highlighted some significant differences between estimates obtained from MM and SM surveys, we moved to analyse the impact of non-response in the two samples in order to understand the role of the different response processes.

Response or non-response can be considered as an household behaviour. In fact, no member of an household is respondent if the household is non-respondent.

The following auxiliary variables defined at household level, are used in the response rate analysis:

- Higher education level: below/equal/above high school diploma
- Income class: 5 quintiles (10.508, 20281, 29778, 46079 euros)

- Citizenship: Italian/Foreign household
- Household type: one-component under 35, one-component 35-64, one-component over 64, two-components at least one under 35, two-components all over 34, more than two components at least one under 25, more than two components all of them over 24.
- Municipal type: Metropolitan cities, metropolitan area, other municipalities with <2000, 2000-10000, 10000-50000, >50000 inhabitants.

The differences between response rates in the two samples is tested using a *z*-test for comparing two population proportions (π_{SM}, π_{MM}) with independent samples (*SM, MM*) , with the following formula:

$$H_0: \pi_{SM} = \pi_{MM}$$

$$H_1: \pi_{SM} \neq \pi_{MM}$$

$$Z_c = \frac{p_{SM} - p_{MM}}{\sqrt{p(1-p)\left(\frac{1}{n_{SM}} + \frac{1}{n_{MM}}\right)}},$$

where: $p = (x_{SM} + x_{MM})/(n_{SM} + n_{MM})$.

Considering a confidence level 1-α=0.99, then $Z_{\alpha/2}$ will be equal to 2.576.

So we will reject the null hypothesis H_0 if $|Z_c| > 2.576$.

In the following table the response rate observed in the two independent samples is shown, according to the categories of the auxiliary variables, which are reported in bold if the difference between response rates in the two sample is statistically significant.

Table 2. Observed response rates and sample size in the two samples, according to the categories of the main auxiliary variables

Variable	Category	SM (PAPI)			MM (web/PAPI)			p	Z_c
		n	n.risp	resp. rate	n	n.risp	resp. rate		
		n_{SM}	x_{SM}	p_{SM}	n_{MM}	x_{MM}	p_{MM}		
Household nationality	Foreign	336	216	64.29%	706	497	70.40%	68.43%	-1.984
	Mixed	**7241**	**5318**	**73.44%**	**17630**	**13540**	**76.80%**	**75.82%**	**-5.619**
	Italian	437	224	51.26%	926	479	51.73%	51.58%	-0.162
Income class (euro)	**<10508**	**1452**	**909**	**62.60%**	**3619**	**2489**	**68.78%**	**67.01%**	**-4.226**
	10508-20281	**1646**	**1148**	**69.74%**	**3818**	**2853**	**74.72%**	**73.22%**	**-3.814**
	20281-29778	1582	1179	74.53%	3960	2992	75.56%	75.26%	-0.802
	29778-46079	1675	1246	74.39%	3917	3034	77.46%	76.54%	-2.481
	>46079	1659	1276	76.91%	3948	3148	79.74%	78.90%	-2.365
Municipal type	**Metropolitan cities**	**1216**	**674**	**55.43%**	**2814**	**1846**	**65.60%**	**62.53%**	**-6.124**
	Metropolitan area	681	485	71.22%	1606	1192	74.22%	73.33%	-1.485
	Other municipalities under 2000 inhab.	618	487	78.80%	1536	1193	77.67%	77.99%	0.574
	Other municipalities between 2000 and 10000 inhab.	1984	1554	78.33%	4830	3870	80.12%	79.60%	-1.673
	Other municipalities between 10000 and 50000 inhab.	2132	1603	75.19%	5116	3959	77.38%	76.74%	-2.017
	Other municipalities over 50000 inhab.	**1383**	**955**	**69.05%**	**3360**	**2456**	**73.10%**	**71.92%**	**-2.816**
Household type	one-component under 35	280	167	59.64%	623	394	63.24%	62.13%	-1.031
	one-component between 35 and 64	**1182**	**756**	**63.96%**	**2773**	**1965**	**70.86%**	**68.80%**	**-4.289**
	one-component over 64	1243	851	68.46%	3066	2213	72.18%	71.11%	-2.438
	two-components at least one under 35	91	63	69.23%	237	156	65.82%	66.77%	0.587
	two-components all over 34	2100	1594	75.90%	4920	3858	78.41%	77.66%	-2.312
	more than two components, at least one under 25	**2264**	**1663**	**73.45%**	**5539**	**4255**	**76.82%**	**75.84%**	**-3.151**
	more than two components, all over 24	854	664	77.75%	2104	1675	79.61%	79.07%	-1.126
Geographic Area	**North-West**	**1900**	**1274**	**67.05%**	**4369**	**3187**	**72.95%**	**71.16%**	**-4.734**
	North-East	**1741**	**1236**	**70.99%**	**4028**	**3001**	**74.50%**	**73.44%**	**-2.771**
	Center	1616	1128	69.80%	3735	2680	71.75%	71.16%	-1.447
	South	1781	1421	79.79%	5314	4272	80.39%	80.24%	-0.555
	Islands	976	699	71.62%	1816	1376	75.77%	74.32%	-2.395
Higher Educational Level	**Below high school diploma**	**1680**	**1239**	**73.75%**	**4006**	**3104**	**77.48%**	**76.38%**	**-3.024**
	High school diploma	**2749**	**1971**	**71.70%**	**6612**	**5069**	**76.66%**	**75.21%**	**-5.066**
	Above high school diploma	**3585**	**2548**	**71.07%**	**8644**	**6343**	**73.38%**	**72.70%**	**-2.606**
Total		**8014**	**5758**	**71.85%**	**19262**	**14516**	**75.36%**	**74.33%**	**-6.048**

Response rates in the two samples are significantly different when the household presents some characteristics, such as mixed nationality, lower income class, is composed by just one component between 35 and 64 years old or by more than two components, at least one of them under 26 years old; moreover, when the household is located in the North area or in a big city (metropolitan or not).

These results can be confirmed testing the independence between the household response behaviour and the belonging to each sample, taking into account the mentioned modalities of the auxiliary variables already studied.

Table 3. Chi squared test to evaluate the independence between response behaviour and samples belonging, according to categories of the main auxiliary variables

VARIABLE	CATEGORY	CHI-SQUARED	P-VALUE
Household nationality	Foreign	0.606	0.436
	Mixed	**29.911**	**<.0001**
	Italian	0.147	0.701
Income class (euro)	**<10508**	**9.188**	**0.002**
	10508-20281	**12.173**	**0.001**
	20281-29778	0.697	0.404
	29778-46079	6.531	0.011
	>46079	5.920	0.015
Household type	one-component under 35	0.102	0.749
	one-component between 35 and 64	**13.246**	**0.000**
	one-component over 64	5.878	0.015
	two-components at least one under 35	0.300	0.584
	two-components all over 34	4.508	0.034
	more than two components, at least one under 25	**8.119**	**0.004**
	more than two components, all of them over 24	1.019	0.313
Geographic Area	**North-West**	**19.263**	**<.0001**
	North-East	**7.175**	**0.007**
	Center	1.464	0.226
	South	0.008	0.930
	Islands	2.927	0.087
Municipal type	**Metropolitan cities**	**33.791**	**<.0001**
	Metropolitan area	1.971	0.160
	Other municipalities under 2000 inhab.	0.422	0.516
	Other municipalities between 2000 and 10000 inhab.	2.991	0.084
	Other municipalities between 10000 and 50000 inhab.	1.556	0.212
	Other municipalities over 50000 inhab.	5.505	0.019
Highest Educational Level	**Below high school diploma**	**9.147**	**0.003**
	High school diploma	**25.668**	**<.0001**
	Above high school diploma	**6.793**	**0.009**

Studying the categories of the auxiliary variables separately it can be observed that independence hypothesis is rejected when response rates in the two samples were significantly different, according to the parametric z-test.

This two tests confirm that some household covariate pattern lead to different response behavior in the two samples.

4.3. Analysis of total nonresponse bias

4.3.1. Indicators of representative response

To assess the overall quality of respondents' samples in terms of bias, indicators of representative response, known as *R*-indicators and partial *R*-indicators, were used. These indicators are based on a measure of the response propensity variability and, in general, they describe how the sample of survey respondents reflects the population of interest with respect to certain characteristics; essentially, they measure how much the sample of respondents in a survey deviates from the representative response. Furthermore, partial *R*-

indicators allow to analyze if the different designs systematically influence the response composition of the subpopulations, when some subpopulations are under or over represented in the respondents' samples; they can be consideredas a measure of the contribution of each variable to the representative response.

The overall variance of the response propensity and its between and within (strata-subpopulations) components are the measures of variability considered in the *R*-indicators and the partial *R*-indicators.

In this context of study unconditional *R*-partial indicators were adopted to make comparisons between SM and MM surveys (Shlomo et al., 2009). For both independent theoretical samples, the response propensity was estimated through a response model (logistic regression model).

The following indicators of representative response were calculated for both SM and MM samples:

R-indicator:

$$R(\rho_X) = 1 - 2S(\rho_X)$$

where

- ρ_X is the response propensity
- $S(\rho_X)$ is the standard deviation ofρ_X

Estimate of *R*-indicator:

$$\hat{R}(\hat{\rho}_X) = 1 - 2\hat{S}(\hat{\rho}_X) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i \left(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_X\right)^2}$$

where:

$$\hat{\bar{\rho}}_X = \frac{1}{N}\sum_{i=1}^{n} d_i \hat{\rho}_X(x_i)$$

d_i sampling weight.

Unconditional partial *R*-indicator for auxiliary variable Z:

$$P_u(Z, \rho_X) = S_B(\rho_X | Z)$$

The between variance of the response propensity

$$S_B^2(\rho_X | Z) = \frac{1}{N-1}\sum_{k=1}^{K} N_k \left(\bar{\rho}_{X,k} - \bar{\rho}_{X,}\right)^2 \cong \sum_{k=1}^{K} \frac{N_k}{N}\left(\bar{\rho}_{X,k} - \bar{\rho}_{X,}\right)^2$$

is estimated as:

$$\hat{S}_B^2(\rho_X | Z) = \sum_{k=1}^{K} \frac{\hat{N}_k}{N}\left(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_{X,}\right)^2$$

where

$$\hat{\bar{\rho}}_{X,k} = \frac{1}{N_k}\sum_{s_k} d_i \rho_X(x_i)$$

$\hat{N}_k = \sum_{s_k} d_i$ - estimated population size in the stratum *k*,

s_k - sample of respondents in the stratum *k*.

For Z=k, the unconditional partial *R*-indicator

$$P_u(Z, k, \rho_X) = S_B(\rho_X | Z = k)\frac{\left(\bar{\rho}_{X,k} - \bar{\rho}_X\right)}{\left|\bar{\rho}_{X,k} - \bar{\rho}_X\right|} = \sqrt{\frac{N_k}{N}}\left(\bar{\rho}_{X,k} - \bar{\rho}_X\right)$$

is estimated as:

$$\hat{S}_B(\hat{\rho}_X | Z = k) = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X\right)$$

4.3.2. Some results on response representativeness

The auxiliary variables utilized in the response model are: household typology, income class, higher educational level and geographical area. The table below shows the values of the *R*-indicator and its estimate for single mode (SM) and mixed mode (MM) samples.

Table 4. R-indicators in SM and MM samples

***R*-Indicator**	**SM sample**	**MM sample**
$R(\rho_X)$	0.81195	0.85227
$\hat{R}(\hat{\rho}_X)$	0.81397	0.85376

Response is defined to be representative if all the response propensities in the sample are equal, that is when the *R*-indicator is equal to 1, from table 4 emerges that the MM sample of respondents deviates less from the representative response with respect to the SM sample, 0.85376 in the first and to 0.81195 in the second.

In the following analysis, the *R*-indicator computation is based on the response propensity, estimated through response models defined for each geographical area (North, Center, South and Islands).

Table 5. R-indicators in SM and MM samples in the geographical area

***R*_Indicator**	**SM sample**		**MM sample**	
	$R(\rho_X)$	$\hat{R}(\hat{\rho}_X)$	$R(\rho_X)$	$\hat{R}(\hat{\rho}_X)$
North	0.84654	0.84977	0.84043	0.84295
Center	0.75239	0.74822	0.84160	0.83563
South and Islands	0.83956	0.84012	0.90717	0.91357

Table 5 shows that while for the North the values of the R-indicators are similar for the two samples, for the other geographical areas are very different. The response in these cases is more representative when MM survey is adopted. Moreover, although the web response rates are much lower in the South and Islands, in the MM survey the sample of respondents better reflects the population of interest with respect to certain characteristics used in the models.

In the tables below, the contribution of the variables household typology, income class and geographical area to the representativeness of the response is analysed through unconditional partial R-indicators.

If the variability of the propensity to respond between the subpopulations (strata) is high, then the contribution of the variable to non-representativeness is higher. This indicator is non-negative and assumes values less than or equal to 0.5. At the subpopulation level assumes values between -0.5 and 0.5: a negative value indicates that a subpopulation is under-represented, while a positive value indicates that a subpopulation is over-represented, the value zero (0) means that it is represented (Schouten et al., 2011).

Table 6.Unconditional partial R-indicator for the income class variable and subpopulations (strata)

			SM sample	**MM sample**
Variable	Income class	$\hat{P}_u(Z,\hat{\rho}_X)$	0,0032	0,0020
Strata	<10508	$\hat{P}_u(Z,k,\hat{\rho}_X)$	-0,0449	-0,0359
	10508-20281		-0,0077	-0,0016
	20281-29778		0,0142	0,0035
	29778-46079		0,0141	0,0123
	>46079		0,0261	0,0238

The unconditional partial R-indicator for the "income class" assumes a higher value in the SM sample rather than in the MM sample, this means that the variable contributes more to the non-representativeness of the response in the SM sample. If the subpopulations (strata) are analysed, it should be noted that for the first

and second quintiles of the distribution (lower incomes) there is an under-representation of the two samples but more marked for the respondents SM sample. In addition, there is a greater over-representation of household with higher incomes in the SM sample (Tab. 6).

Table7.Unconditional partial R-indicator for the household typology and subpopulations (strata)

			SM sample	**MM sample**
Variable	Household typology	$\hat{P}_u(Z,\hat{\rho}_X)$	0,0029	0,0019
Strata	one-component<= 34	$\hat{P}_u(Z,k,\hat{\rho}_X)$	-0,0267	-0,0276
	one-component 35 - 64		-0,0328	-0,0199
	one-component>= 65		-0,0126	-0,0114
	two-components at least one <= 35		-0,0015	-0,0080
	two-components all > 34		0,0219	0,0166
	more than two components at least one <= 24		0,0096	0,0091
	more than two components all >24		0,0206	0,0155

The "household typology" contributes, according to the indicator, more to the non-representativeness of the response in the SM sample (0.00291) than in the MM sample (0.00195) . When the subpopulations are analysed, it should be noted that for households with one-component and two components aged less than or equal to 35 years there is an under-representation of the two respondents samples but more accentuated for the SM survey design except for the last subpopulation. There is still more over-representation for households with two components over 34 and households with more than two components in the SM sample compared to the MM sample (Tab. 7).

Table 8. Unconditional partial R-indicator for the geographical area and subpopulations (strata)

			SM sample	**MM sample**
Variable	Geographical area	$\hat{P}_u(Z,\hat{\rho}_X)$	0,0024	0,0012
Strata	North-West	$\hat{P}_u(Z,k,\hat{\rho}_X)$	-0,0267	-0,0094
	North-East		0,0092	0,0067
	Center		-0,0159	-0,0214
	South		0,0374	0,0238
	Islands		-0,0001	0,0024

The "geographical area" contributes more to the non-representativeness of the response in the SM sample than in the MM sample (Tab. 8).

Table 9.Unconditional partial R-indicator for the "higher educational level" and subpopulations (strata)

			SM sample	**MM sample**
Variable	Higher educational level	$\hat{P}_u(Z,\hat{\rho}_X)$	0,0002	0,0003
Strata	below high school diploma	$\hat{P}_u(Z,k,\hat{\rho}_X)$	0,0111	0,0120
	equal high school diploma		0,0049	0,0111
	above high school diploma		-0,0034	-0,0085

The contribution to the representative response of "higher educational level" variable is, on the whole, very low in both samples, but the over- and under- representation of the samples is, always, greater in the MM sample respect to SM sample (Tab. 9).

4.4. Total bias due to nonresponse on benchmark variable

In order to show an example of measure of the total bias due to nonresponse, in table 10 the estimated frequencies of one auxiliary variable, the income class, in both SM and MM samples (direct estimates on respondents) are reported in comparison with the estimates obtained on the theoretical sample as a benchmark value (SM+MM). The numbers highlight that the SM sample is overall slightly more biased than the MM sample: this example confirms that the nonresponse bias is different in the two samples and therefore that for an assessment of the mode effect based on a control sample (SM) it is necessary to make the two samples comparable.

Table 10. Income class bias in SM and MM samples

Income class	Benchmark estimate	SM sample estimate	MM sample estimate	SM Absolute bias	MM Absolute bias
I	17.5%	15.1%	15.4%	2.4%	2.1%
II	16.1%	15.6%	15.8%	0.5%	0.3%
III	18.5%	18.6%	19.0%	0.1%	0.5%
IV	22.1%	23.1%	22.5%	1.0%	0.4%
V	25.8%	27.5%	27.3%	1.8%	1.5%
Total				5.8%	4.9%

4.5. Analysis of selection and measurement effects with SM survey as benchmark

The design of the MM sample in the ADL survey involves selection and measurement effects; these are effects that are confusing and difficult to evaluate. Comparing the respondents of MM sample with the "comparable" respondents of SM sample allows distinguishing the PAPI respondents from the web/PAPI respondents and disentangling measurement and selection effects (Vannieuwenhuyze et al., 2010).

The analyses conducted above regarding the response rates and the representative response show that the respondents in the SM and MM samples (households) are different.The social-demographical composition of the respondents in the two samples is not the same. This suggests that it is possible that the MM design attracts a more hard-to-reach population than the PAPI mode.

In order to make the SM and MM samples comparable, a calibration system was adopted separately for the two samples starting from the sampling weights.

For both samples the population distributions of the socio-demographic variables were used. These variables are: region (21 categories), age class (8 categories: <=5, 6-13, 14-24, 25-34, 35-44, 45-54, 55-64 >=65), sex, citizenship (Italian/foreign), municipal type as above defined.

The method proposed by Vannieuwenhuyze et al. (2010) is based on the probability distributions of the target survey variables(categorical) estimated from the two respondent's samples (SM and MM).Using the PAPI survey (SM sample) as a benchmark, selection and measurement effects can be partially evaluated.

For a generic variable, A, with *t* categories (i = 1, ..., *t*), two sub-variables, A_{SM} and A_{MM}, measured respectively on the respondents of the SM (PAPI) and MM (web / PAPI) samples, are defined. The A_{SM} variable has a multinomial distribution with parameters vector $\pi_{SM} = \pi_{SM,1}, \pi_{SM,2}, \pi_{SM,3}, \ldots, \pi_{SM,t}$, where $\pi_{SM,i}$ is the probability that $A_{SM=}i$, $0 \le \pi_{SM,i} \le 1$ and $\sum \pi_{SM,i} = 1$. The A_{MM} variable has a similar distribution to A_{SM}.

Furthermore, the variable M identifies the response with the two mode in the MM sample (1 for PAPI and 0 for web). M has a distribution with a parameter τ, $0 \le \tau \le 1$ that represents the probability that respondents will choose PAPI in the MM sample.

On the basis of the observed data, the following distributions are defined:

$P(A_{SM})$ - from SM, distribution of the respondents sample with PAPI mode;

$P(A_{SM}|M=1)$ - from MM, distribution of the respondents sample with PAPI mode in MM;

$P(A_{MM}|M=0)$ - from MM, distribution of the respondents sample with web mode in MM;

$P(M=1)$, $P(M=0)$ distribution from all MM.

Applying the total probability formula it is possible to derive:

$$P(A_{SM}) = P(A_{SM}|M=0)P(M=0) + P(A_{SM}|M=1)P(M=1)$$

from which

$$P(A_{SM}|M=0) = P(A_{SM})\frac{1}{P(M=0)} - P(A_{SM}|M=1)\frac{P(M=1)}{P(M=0)}$$

This equation can be used to derive the probabilities $\pi_{SM,i}|M=0$ for each category of the A_{SM} variable for respondents (PAPI or web) in MM. These probability distributions can be compared:

- with the probability distributions observed for respondents who choose PAPI in MM. The selection effect is given by:

  $$(\pi_{SM,i}|M=1) - (\pi_{SM,i}|M=0)$$

- with the probability distributions observed for respondents who choose PAPI or web in MM. The measurement effect is given by the difference between the probability of response measured with PAPI or web (MM) minus the probability of response measured with PAPI (SM):

  $$(\pi_{MM,i}|M=0) - (\pi_{SM,i}|M=0)$$

The interpretation of the probabilities for each response category and the mode effect can be facilitated by combining the parameters vector $\boldsymbol{\pi}_{SM}$ in the single parameters, the mean and the variance

$$\mu_{SM} = (1 * \pi_{SM,1} + 2 * \pi_{SM,2} + 3 * \pi_{SM,3} + \ldots + t * \pi_{SM,t}),$$

$$\sigma^2_{SM} = ((1-\mu_{SM})^2 * \pi_{SM,1} + (2-\mu_{SM})^2 * \pi_{SM,2} + (3-\mu_{SM})^2 * \pi_{SM,3} + \cdots + (t-\mu_{SM})^2 * \pi_{SM,t}).$$

The selection and measurement effects can be evaluated on the estimated parameters. Measurement effects on the mean may indicate over- or under-estimation of one mode over the other. The measurement effects on variance may indicate that the responses are very heterogeneous for one mode in comparison with the other mode (Vannieuwenhuyze *et al.*, 2010).

4.5.1. Results obtained on comparable samples

The following tables show the results of the application of the method described above for some variables of the Aspects of Daily Life (ADL) survey, such as "reading books in the last 12 months", "use of pc in the last year" and "life's satisfaction". The first three columns of the tables respectively represent the distributions of the web group in MM sample, PAPI group in MM sample and PAPI group in SM sample.

Table 11.Estimate of the selection and measurement effects for "Reading books in the last 12 months"

Category	$A_{MM}\|M=0$	$A_{SM}\|M=1$	A_{SM}	$A_{SM}\|M=0$	Selection effect	Measurement effect
NO	0,4086	0,6291	0,5755	0,4813	0,1478	-0,0727
YES	0,5531	0,3348	0,3989	0,5115	-0,1767	0,0416
NR	0,0383	0,0361	0,0256	0,0073	0,0288	0,0311
mean	1,6297	1,4070	1,4502	1,5260	**-0,1190**	**0,1038**
variance	0,3098	0,3136	0,2988	0,2639	**-0,3275**	**0,3518**

Table 11 shows the presence of both selection and measurement effects. With respect to the category 'NO' the measurement effect is negative and selection effect is positive, while the opposite occurs with respect to the category "YES". The positive selection effect for the NO category may be indicative of the fact that the PAPI respondents in the MM sample are more likely to respond NO than the web respondents.

Mean and variance are negative for selection effect and positive for the measurement effect. A positive measurement effect on the mean indicates that the web survey mode over-estimate the overall variable in comparison with the other mode. The negative selection effect on the mean expresses that the PAPI respondents in the MM sample are on average less interested in reading books than the web respondents.

Table 12. Estimate of the selection and measurement effects for "Use of internet"

Category	$A_{MM}\|M=0$	$A_{SM}\|M=1$	A_{SM}	$A_{SM}\|M=0$	Selection effect	Measurement effect
yes, in the last 3 months	0,6430	0,4331	0,5279	0,6945	-0,2614	-0,0515
yes, from 3 months to 1	0,0321	0,0265	0,0260	0,0251	0,0014	0,0070
yes, more than 1 year-	0,0595	0,0501	0,0484	0,0455	0,0045	0,0140
never	0,2293	0,4559	0,3734	0,2285	0,2273	0,0008
NR	0,0360	0,0345	0,0243	0,0063	0,0282	0,0297
mean	1,9834	2,6323	2,3400	1,8271	**0,8052**	**0,1563**
variance	1,9440	2,2175	2,1727	1,6811	**2,8304**	**0,8096**

Concerning the "Use of internet" (Table 12), with respect to the first category both measurement and selection effects are negative, while with respect to the other categories are positive. More individuals respond to other categories when this question is asked by web. The positive selection effect for the all categories, except to the first category, indicates that the PAPI respondents in the MM sample are more likely to provide these answers than the web respondents.

Table 13. Estimate of the selection and measurement effects for "Life's satisfaction"

Category	$A_{MM}\|M=0$	$A_{SM}\|M=1$	A_{SM}	$A_{SM}\|M=0$	Selection effect	Measurement effect
00	0,0083	0,0054	0,0052	0,0047	0,0007	0,0035
01	0,0048	0,0048	0,0047	0,0046	0,0002	0,0002
02	0,0106	0,0089	0,0056	-0,0002	0,0091	0,0108
03	0,0151	0,0175	0,0113	0,0007	0,0168	0,0144
04	0,0243	0,0322	0,0262	0,0158	0,0164	0,0085
05	0,0875	0,0986	0,0821	0,0533	0,0453	0,0342
06	0,1548	0,1853	0,1725	0,1503	0,0350	0,0045

07	0,2645	0,2361	0,2478	0,2682	-0,0321	-0,0037
08	0,2651	0,2388	0,2806	0,3532	-0,1143	-0,0881
09	0,0875	0,0754	0,0827	0,0954	-0,0199	-0,0078
10	0,0355	0,0583	0,0547	0,0486	0,0097	-0,0131
NR	0,0420	0,0387	0,0266	0,0056	0,0331	0,0364
mean	8,1117	8,0592	8,1636	8,3453	**-0,2861**	**-0,2336**
variance	3,4032	3,4906	2,9841	2,0502	**-3,4164**	**-2,6004**

Concerning "Life's satisfaction" (Table 13), with respect to the categories '07, 08 and 09' the measurement and selection effects are negative, while with respect to the other categories are positive. It appears that the PAPI mode measures a higher mean Life's satisfaction compared to a web, which becomes a measurement effect. This result supports the hypothesis that respondents provide answers by assuming social desirable behavior.

5. The analysis of mode effect in the MM survey

5.1. The application of Propensity Score for the assessment of mode effect

In MM sample difference in the estimates of the parameters of interest of the survey - calculated on the samples of web and PAPI respondents - can be determined either by the different composition of the samples or by differences in measurement errors (Hox et al., 2015).

Moving to get an assessment of selection and measurement effect in the MM sample, a Propensity score stratification adjustment methods was used (Rosenbaum and Rubin, 1983). Propensity score (PS) approach is adopted in observational studies by achieving a balance of covariates between comparison groups. In the mixed mode context propensity score can be interpreted as the probability of mode assignment conditional on observed covariates. With adjustments based on PS, the confounding effects of the selection mechanism are mitigated.

The application of this approach to the ADL survey implied:

- an estimation of the propensity score model parameters;
- the definition of sub-classification (strata) of web and PAPI respondents based on propensity score;
- the validation of the balancing assumption, through a chi-square test of the independence between the mode choice and each of the covariates;
- for each balanced group, the calculus of weights that equate the weighted proportion of web respondents with the proportion of PAPI respondents in the same stratum.

A logit regression model was used where the binary response variable is the mode choice web/PAPI. The parameters resulted significant for the following auxiliary variables: geographic region, type of municipality, household typology, income class and higher educational level.

For eight out of ten of the deciles of the distribution of the predicted probabilities the independence hypothesis was accepted for all variables. For each balanced group k, a correction factor, or weight, of the selection effect (Vandenplas et al., 2016) has been calculated as

$$w_k = \frac{n_{k,papi}/n_{papi}}{n_{k,web}/n_{web}}.$$

This weight allows an overall evaluation of the mode effect in the balanced classes: the selection effect for the variable y, $S_{web}(y)$, is obtained, following Vandenplas (2016), as the difference between the weighed and unweighted estimates of the respondents to the web mode, while the measurement effect, $M_{web}(y)$, is obtained as the difference between the weighted estimate of the web respondents and the unweighted estimate of PAPI respondents:

$$S_{web}(y) = \frac{\sum_{i=1}^{n_{web}} y_{i,web}}{n_{web}} - \frac{\sum_{i=1}^{n_{web}} w_{k,i}\, y_{i,web}}{n_{web}}$$

$$M_{web}(y) = \frac{\sum_{i=1}^{n_{web}} w_{k,i}\, y_{i,web}}{n_{web}} - \frac{\sum_{i=1}^{n_{papi}} y_{i,papi}}{n_{papi}}$$

The results in the table 14 show values of selection and measurement effects for the main variables of ADL survey resulted with significant difference between SM and MM design (see Figure 1).

Table 14. Selection and measurement effect estimated through PS for some variables

Variable	Category	Weighted Web mean	Web mean	PAPI mean	Selection effect	Measurement effect
Reading books, in last 12 months	No	0.485	0.451	0.618	0.034	-0.132
	Yes	0.432	0.508	0.347	-0.075	0.085
	NR	0.043	0.041	0.035	0.002	0.007
Frequency of seeing friends	Everyday	0.098	0.088	0.183	0.010	-0.085
	Sometimes a week	0.235	0.250	0.266	-0.015	-0.030
	Once a week	0.199	0.210	0.192	-0.011	0.007
	Sometimes a month	0.196	0.216	0.180	-0.021	0.015
	Sometimes a year	0.138	0.146	0.086	-0.008	0.052
	Never	0.050	0.048	0.060	0.001	-0.010
	No friends	0.020	0.018	0.022	0.002	-0.003
	NR	0.025	0.024	0.011	0.000	0.014
Use of Personal Computer	yes, in the last 3 months	0.528	0.604	0.441	-0.076	0.087
	yes, from 3 months to 1 year-ago	0.035	0.035	0.026	0.000	0.009
	yes, more than 1 year-ago	0.059	0.061	0.048	-0.002	0.011
	never	0.298	0.262	0.451	0.037	-0.153
	NR	0.040	0.038	0.034	0.002	0.006
Internet access	NO	0.212	0.180	0.350	0.032	-0.137
	YES	0.708	0.783	0.621	-0.075	0.087
	NR	0.040	0.036	0.029	0.004	0.011
Use of internet	Yes, in the last 3 months	0.608	0.675	0.559	-0.067	0.049
	Yes, between 3 months and one year ago	0.029	0.029	0.020	-0.001	0.009
	Yes, more than one year ago	0.046	0.047	0.029	-0.001	0.017
	Never	0.236	0.210	0.358	0.026	-0.122
	NR	0.041	0.039	0.034	0.002	0.007
Life's satisfaction	0	0.008	0.008	0.006	0.001	0.003
	1	0.006	0.005	0.004	0.001	0.001
	2	0.010	0.011	0.008	-0.001	0.002
	3	0.017	0.016	0.017	0.001	0.001
	4	0.026	0.026	0.028	0.000	-0.002
	5	0.100	0.094	0.094	0.006	0.006
	6	0.158	0.159	0.178	-0.001	-0.019
	7	0.245	0.257	0.239	-0.012	0.007
	8	0.233	0.256	0.250	-0.023	-0.017
	9	0.075	0.086	0.079	-0.010	-0.004
	10	0.036	0.036	0.060	-0.001	-0.025
	NR	0.047	0.046	0.037	0.001	0.010
Trust in others	in the majority of people	0.183	0.214	0.166	-0.031	0.017
	you have to be careful	0.732	0.740	0.799	-0.008	-0.067
	NR	0.047	0.046	0.035	0.001	0.012

The table shows that, for some of these variables, the estimated effects seem to lead to similar conclusions as the results of paragraph 4.5.1, deriving from the analyses based on a benchmark mode. Generally, both selection and measurement effects emerge, especially for some items of the selected variables.

5.2. The diagnostic method - multi-group confirmatory factor analysis

When data are collected with different techniques, it cannot be assumed that the responses to a certain item can be directly compared. In fact, the comparability of the answers depends on their degree of equivalence. The equivalence of the measurements in surveys using mixed data collection modes involves the same magnitude and direction of the measurement error.

The diagnostic method "multi-group confirmatory factor analysis" (MCFA) was used to analyze the phenomenon related to the behavior of respondents regarding to both the use of free time (cinema, theatre, sport performances etc.) and the habits of collecting waste. The aim is to assess (i) whether the measurement model used to represent a "behavioral model" is the same for subjects who responded with web and PAPI techniques, and (ii) if the mean level of the latent factors useful for measuring the phenomenon is the same between the two techniques.

The MCFA has been carried out after controlling for selection effect and after carrying out an exploratory analysis for the identification of the latent structure of the phenomenon. The software used to perform the analyses is Latent GOLD syntax version 5.0 (Vermunt and Magidson, 2013).

5.2.1. Multi-group confirmatory factor analysis

In the first application, the analysis of the behavior of respondents regarding to the use of free time is described. The MCFA has been applied in order to identify which are the latent constructs related with the behavioral model regarding the frequency of going to the cinema, theatre, etc. The aim is to analyze both the measurement invariance (configural, metric, scalar) of the factor model across groups (web/PAPI) and whether the mean levels of the latent factors are equal between different data collection modes.

The variables selected for the analyses are described in Table 15.

Table 15. Description of the variables used for the MCFA

Item	Original scale	Transformed scale
How many times did you go to the theatre in the last 12 months?	1='Never' 2='1-3 times' 3 ='4-6 times' 4 ='7-12 times' 5 ='More than 12 times'	1='Never' 2='1-3 times' 3 ='More than 3 times'
How many times did you go to the cinema in the last 12 months?		
How many times did you go to classical music concerts and opera in the last 12 months?		
How many times did you go to museum and exhibitions in the last 12 months?		
How many times did you go to other music concerts in the last 12 months?		
How many times did you go to sport performances in the last 12 months?		
How many times did you go disco, night clubs and other places to dance in the last 12 months?		
How many times did you go to archaeological sites and monuments in the last 12 months?		

Different factor model specifications are possible with ordinal responses, the more used being the cumulative link model and the adjacent-category logit model (Agresti, 2002). More specifically, with an ordered polytomous y_{ih} variable with categories c = 1, 2, ..., C-1 for individual i and item h, the standard cumulative link model, expressing the probability of being in one of the highest categories, is:

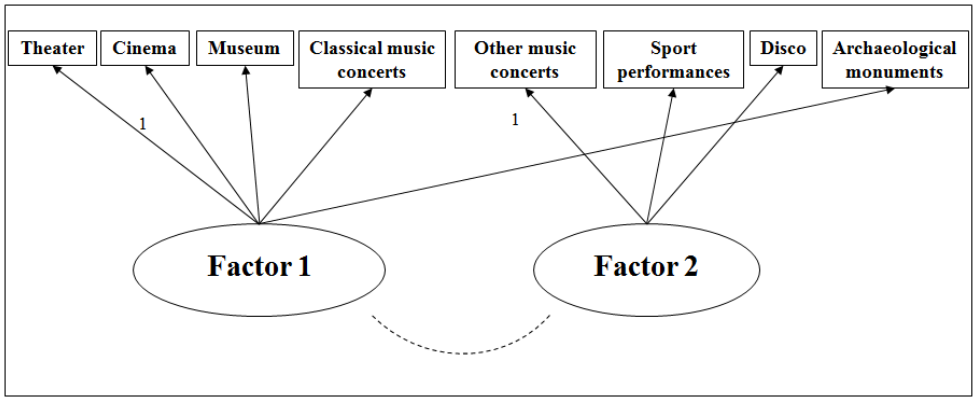
$$F^{-1}[P(y_{ih} \geq c \mid \boldsymbol{\eta_i})] = \alpha_{hc} + \Sigma_j \lambda_{hj} \eta_{ij}$$

where F is commonly chosen as a normal or logistic distribution function (Muthén and Asparouhov, 2002), $\boldsymbol{\eta}_i$ is a vector containing the J latent variables for individual i, η_{ij} is the value of the j-th (j = 1, 2, ..., J) latent variable for the i-th individual and λ_{hj} are the factor loadings relating the latent factor η_j with the observed variable y_{ih}; α_{hc} are the model intercepts, called thresholds.

The main feature of this model is that the effect of ηij are the same for each cumulative probability invariant to the choice of categories for y. Furthermore, each cumulative link has its own intercept. In this application there are 3 categories per each item and two thresholds αhc are estimated; we chose F as a standard normal cumulative distribution function.

The confirmatory factor analysis (CFA) indicates the presence of two underlying correlated factors (Fig. 2).

Figure2. Confirmatory factor analysis, two correlated latent factors



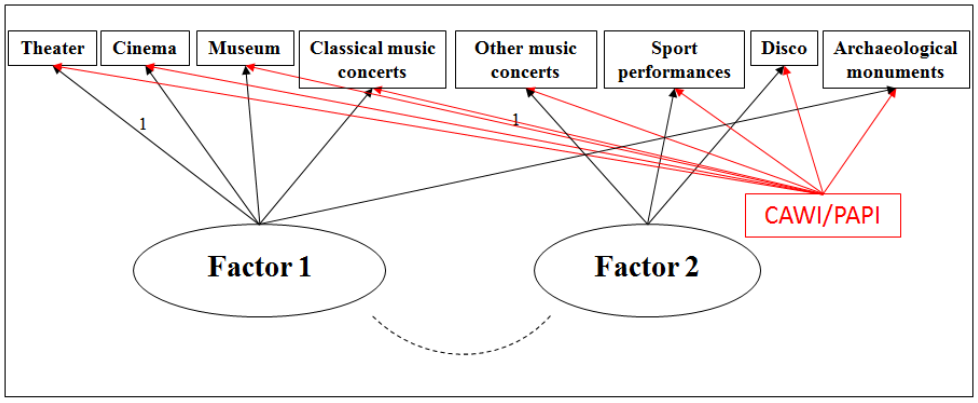
As shown in Table 16, all factor loadings are positive: the higher the factor score is, the higher is $P(y_{ih} \geq c \mid \eta_i)$, that is the higher the frequency of going to the cinema, theatre, etc. The most important aspect related to the first factor is the frequency of going to the theatre, and the highest coefficient measuring the second factor is that one related to the frequency of going to other (not classical) music concerts.

Table 16. Confirmatory factor analysis, two correlated latent factors. Parameters.

Item	Factor loadings		Thresholds	
	Latent factor 1	Latent factor 2	α1	α2
Theatre	1		-1,35	-2,59
Cinema	0,82		-0,07	-1,23
Museum	1,32		-1,01	-2,85
Classical music concerts	1,16		-217	-3,03
Other music concerts		1	-1,69	-3,28
Sport performances		0,59	-0,89	-1,78
Disco		0,60	-1,19	-1,82
Monuments	1,15		-1,10	-2,50
Variance, latent factor	1,33	3,21		
Covariance, latent factors		1,73		

After analysing the factor structure underlying the phenomenon of behavioural model regarding the frequency of going to the cinema, theatre, etc., a MCFA has been applied in order to evaluate the presence of measurement invariance between respondents with PAPI and web (CAWI) techniques. To control for the selection effect weights have been used (propensity score method) (Hox *et al.*, 2015).

Figure 3. Confirmatory factor analysis, two correlated latent factors and data collection technique



After estimating two models with the same form (same dimensions, same pattern of fixed, free and constrained parameters) for both web and PAPI respondents with a good fit to the data, the presence of measurement invariance has been tested (see Figure 3 and Table 17), in particular, configural and scalar invariance.

Indeed, in the context of factor models with ordinal variables, the relevant parameter set for studies of invariance are the thresholds and the factor loadings (Muthén and Asparouhov, 2002): changes in loadings imply changes in intercepts as well. If scalar invariance holds, we test if there are differences in factor means or variances. To evaluate the significance of the restrictions both the Likelihood Ratio Test for nested models (the model to be tested is obtained by constraining some parameters of the previous model) and comparison of Bayesian Information Criteria index (BIC) for not nested models was used.

Configural invariance assumes that the same observed variables are associated with the same latent factors. Scalar invariance assumes that factor loadings and item thresholds are equal across data collected with PAPI and web techniques. The Likelihood Ratio Test between the first two models suggests that only configural equivalence holds: the two groups measure the two latent constructs in a different way.

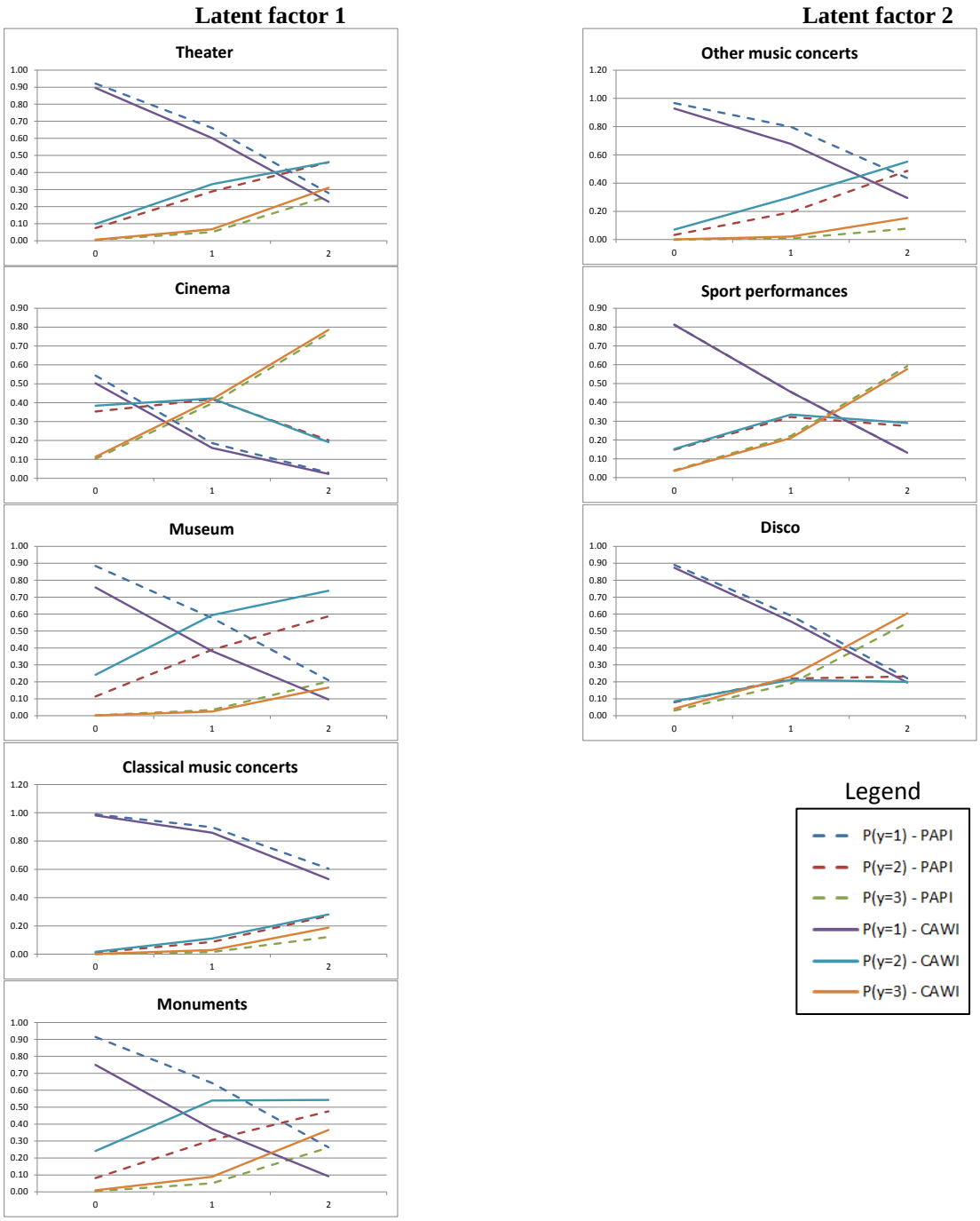
Table 17. Test of invariance for the factor model for web and PAPI

MODEL	Log Likelihood	BIC	# parameters
Configural invariance: unrestricted model	-115593	231692,8	50
Scalar invariance with different covariance	-116020	232323,8	28
Scalar invariance with different covariance and factor means	-115931	232167,1	30

Figure 4 shows the posterior probabilities of the observed items conditional on some values of the latent factor (0,1,2) and the group membership (web and PAPI), for an "typical respondent".

While the items "Cinema" and "Sport performances" show a very similar response to the shift of the latent factor, the posterior conditional probabilities of the items "Museum" and "Other music concerts" show a shift in the intercepts, and those of the item "Monuments" show a completely different behaviour in the two groups.

Figure 4. Posterior probabilities of the observed items conditional on the latent factor value and the group membership (web and PAPI), for a "typical respondent"



In the next application, the analysis of the households habits of collecting waste separately, namely through appropriate waste containers in the street (bins) and/or the door to door service is described.

The questionnaire asks questions for the following king of waste: paper, plastic drugs, batteries, glass, containers of aluminium and other metals, organic/wet waste, textile waste. The original nominal scale with 5 categories was transformed in order to obtain dichotomous indicators (yes/no), as shown in Table 18. The difference with the previous application is in the nature of the indicators and in the type of respondent, all family components in the former application and only the householder in the latter.

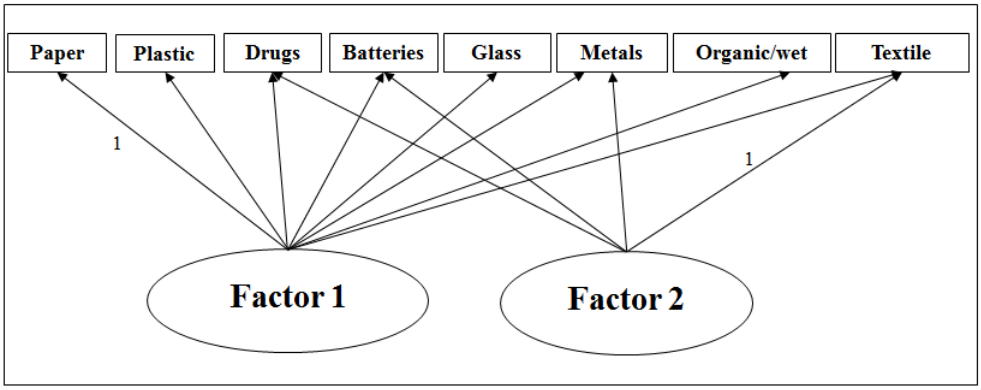
Table 18. Description of the variables used for the MCFA

Item	Original scale	Transformed scale
Do you collect the following waste separately through the appropriate waste containers in the street (bins) and / or the door to door service?		
Paper Plastic Drugs Batteries Glass	1='Through bins, always' 2='Through bins, sometimes' 3='Through door to door service, always' 4='Through door to door service, sometimes'	1='Yes'
Containers of aluminum and other metals	5 ='Never	2='No (never)'
Organic/wet waste		
T textile waste		

For the analysis, the strategy described in the previous paragraphis followed, using the logit model for all (binary) indicators instead of the cumulative link model.

The factor model describing the data is represented in Figure 5. The first factor represents the general sensibility of the family to the waste problem, while the second factor is linked to the waste that usually are less common to be produced by a family, and for which there are usually fewer collection services.

Figure 5. Confirmatory factor analysis, two correlated latent factors



After estimating two models with the same form for both web and PAPI respondents (configural invariance), the presence of measurement invariance, namely scalar invariance through the application of a MCFA with weights for controlling the selection effect was tested.

Table 19. Test of invariance for the factor model for CAWI and PAPI

MODEL	Log Likelihood	BIC	# parameters
Configural invariance: unrestricted model	-26972.5	54320.17	40
Scalar invariance with different covariance	-27054.4	54315.18	22
Scalar invariance with equal covariance	-27056.1	54299.76	20

Also in this application, the results of the LRT suggests that the two groups measure the two latent constructs in a different way. Differently from the previous example, the values of the BIC for the different models are not that dissimilar and suggest to use the scalar invariance hypothesis.

This is probably due to the fact that the results of the LRT are affected by the sample size (Friston, 2013). Furthermore, since the parameters measuring the difference between the two data collection techniques in the variance-covariance matrix of the latent factors are not significant, we can conclude that there is measurement invariance and that factor loadings and item intercepts are equal across data collected with PAPI and CAWI techniques.

Table 20 reports parameters estimates of the factor model describing the families habits of collecting waste.

Table 20. Confirmatory factor analysis, two correlated latent factors. Parameters

Item	Factor loadings		Intercepts
	Latent factor 1	Latent factor 2	
Paper	1		-16,15
Plastic	0,80		-12,98
Drugs	0,59	7,50	-5,29
Batteries	0,48	5,69	-3,43
Glass	0,66		-11,27
Metals	0,29	0,67	-3,06
Organic/wet	0,21		-3,18
Textile	0,19	1	-1,46
Variance, latent factor	90,23		1,06

6. The adjustment of mode effect in the MM sample

6.1. Weighting methods and multiple imputation

In this section some methods for adjusting mode effect are applied: the weighting methods, as propensity score and calibration,are used to correct the selection effect; the multiple imputation method based on MAR assumption is used to correct measurement errors.

The weighting methods assume that the selection effect is ignorable and the measurement error determined by the mix of techniques is negligible or remains constant over time, so as not to affect the estimates of variation. The assumption of the invariance over time of measurement error in repeated sequential mixed-mode surveys is not very sustainable, because the composition of the respondents by mode can change in the survey occasions, leading to variations in the total measurement error.

To avoid the misinterpretation of variations in the composition of respondent samples as variations in the estimates, a calibration procedure that takes into account fixed levels of mode proportions is used. This method is proposed by Buelens and Van den Brakel (2011) and aims at keeping the measurement error constant over the survey occasions. The calibration procedure simultaneously performs with respect to both auxiliary variables, that correct the selection effect, and to fixed levels of proportions of response by mode, that stabilize the total measurement error.

In the methodological approach based on imputation models, since the measurement error is conceptualized as a problem of missing data, the adjustment of measurement effects is performed by converting the response provided with a specific mode into counterfactual responses. This approach, starting from the construction of a complete data set of respondents with the same technique, for example that assumed as a reference, leads to obtain mode-specific survey estimates. Standard methods for the treatment of partial non-response and implemented in the SAS Proc MI procedure were used.

6.2. Comparison of the results

Tables 21 and 22 show the comparison of the estimates for "Reading books" and "Continuity in sport activity" deriving from the application of different methods.

These methods are based on calibration procedures with respect to distributions of the same socio-demographic totals (age class, sex, educational level) at geographical area level, but differ for other aspects of the procedure: 1) calibration on only socio-demographics; 2) calibration on socio-demographics and observed fixed levels of mode proportions by six municipal typologies; 3) calibration on socio-demographics and hypothesized fixed levels of mode proportions by six municipal typologies; 4) calibration on socio-demographics with sampling weights corrected for the web selection effect through correction factors w_k; 5) multiple imputation (counter factual PAPI response for web respondents) and calibration on socio-demographics. In the first column, as a reference, are displayed the estimates obtained from the SM sample with nonresponse adjustment and calibration on socio-demographics.

Table 21. Estimate of "Reading books in the last 12 months" variable with different methods

Variable	Item	Estimate (%)					
		SM estimate	Meth. 1	Meth. 2	Meth. 3	Meth. 4	Meth. 5
Reading books (last 12 months)	No	57,81	59,92	59,00	58,66	59,92	68,94
	Yes	39,68	36,51	37,43	37,73	36,33	28,23
	NR	2,49	3,58	3,56	3,61	3,75	2,48

Table 22. Estimate of "Continuity in sport activity" variable with different methods

Variable	Item	Estimate (%)					
		SM estimate	Meth. 1	Meth. 2	Meth. 3	Meth. 4	Meth. 5
Continuity in sport activity	No	73,22	75,79	74,97	74,81	75,47	82,66
	Yes	25,28	22,57	23,35	23,45	22,79	16,01
	NR	1,51	1,64	1,68	1,73	1,74	0,94

What emerges from the tables is that the two calibrations including the constraints with respect to fixed level of mode proportions (methods 2 and 3) determine a difference in the estimate of about one percentage point. Important differences in the estimates of the two parameters of interest are highlighted when the measurement error correction is used for the web responses. The results of method 5, multiple imputation, suggest that without the adjustment of the selection effect the estimates are far from the others and in particular from the SM estimates.

7. Discussion and conclusions

7.1. Summary and discussion of the results

The analyses presented brought out several issues deriving from the introduction of the mixed mode in a social survey. The survey context was peculiar for the presence of a control sample which allowed to carry out a deeper assessment of the impact of mixed mode.

The analysis shows that in the mixed mode survey the bias due to the total nonresponse is reduced, confirming what stated in Deliverable 1 of WP2 (Buelens, Van den Brakel and Schouten, 2018) . It remains difficult to get an overall evaluation of the total measurement error determined by different conflicting factors, such as the response process and the mode choice. Moreover in a multipurpose survey the complexity of the assessment is even more affected by the need to control many target variables on which multiple effects have different impacts.

If the objectives of cost reduction and of better population coverage are achieved, the impact on the quality of the estimates seems negative and moreover difficult to assess. In fact it is a complex task to interpret the results because it is not easy to understand if the different effects are correctly disentangled and estimated.

The analyses presented highlight, moreover, the complexity of the survey context, deriving from the variety of indicators and from the sequential nature of this mixed mode. In fact, the mixed mode design catches better the overall population resulting more "representative" than the single mode design. Anyway, the positive impact of mixed mode in terms of obtaining a less selectivity response, does not necessarily become an improvement of the estimates of the target variables.

The outcome presented in this report, anyway, would need a significance assessment, based on tests or replication methods, which are in progress.

The results are conditioned by the limits of the auxiliary variables currently available, although ISTAT is confident that in the future the system of register will improve also the potentialities of mode effect assessment. So for the Italian NSI this experiment represents an useful exercise rather than the search of a solution for the present case.

7.2. Concluding remarks

The set of the analyses presented and applied in a specific survey context can be considered as a possible checklist, a sequence of steps usable by researchers of other NSIs to carry out an assessment of mode effect in similar situations. They try to cover all the different approaches applicable in this specific survey context, even if without claiming to be exhaustive.

The experience presented in this report was very useful for ISTAT because it was the occasion for experimenting several methods for assessing and adjusting mode effect in an experimental context, not very frequent for this NSI. From this experience it can be underlined that the introduction of mixed mode has an important impact both on the composition of the sample (and its representativeness) and on several indicators, the quality of which seems to be affected by measurement effect not always easily assessable. A similar research path can be followed when an experimental design is set up to evaluate the impact of the switching from single to mixed mode.

Naturally the application of all the presented methods is subject to the validity of the hypotheses that all these methods assume and the researcher has to verify them as far as possible. Besides, the results of the methods depends on the extent to which the specified models support the analyses, taking into consideration also the availability and the quality of the auxiliary information, which should be mode insensitive and well explaining the selection effect.

In conclusion, this report shows the analysis process carried out and the obtained results and what can be highlighted is that the underlying effort is hardly compatible with the usual resources and the timing of a statistical process: only in some cases such a deepening is feasible, in general situations an accurate planning of the data collection phase is advisable, in order to limit as far as possible ex-ante the measurement effect, which is the main drawback of the mixed mode.

References

Agresti, A. (2002). Categorical Data Analysis 2nd edition. Wiley.

Ballabio S., Carra A., Casacci S., Ferrazza D., Verrecchia F., Vitalini A., Viviano L. C. (2018) Local decisions and new guidelines of the Official Statistics, Q2018 European Conference on Quality in Official Statistics, Cracovia June 2018.

Buelens B. and Van den Brakel J. A., (2015) Measurement error calibration in mixed-mode, Sociological methods & Research, 4483, pp 391-426.

Buelens B., Van den Brakel J. A. and Schouten B.,(2018) Current methodologies to deal with mode effects and mode bias in multi-mode designs, MIMOD Deliverable 1 – WP2

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21(2), pp. 233–255

Friston (2013), Sample size and the fallacies of classical inference, Neuro Image, 81, 503–504.

Garofalo, G. (2014). Il Progetto ARCHIMEDE obiettivi e risultati sperimentali. Istat Working Paper (2014).

Gordoni, G., Schmidt, P., and Gordoni, Y. (2012), Measurement invariance across face-to-face and telephone modes: the case of minority-status collectivistic-oriented groups, International Journal of Public Opinion Research, 24, 185—207.

Hox, J., de Leeuw, E. D., e T. Klausch. (2015). "Mixed Mode Research: Issues in Design andAnalysis." Invited paper presented at the International Conference on Total survey error: improving quality in the era of big data. Baltimore, 19-22 September.

Hox, J. J., de Leeuw E. D., e E. A. O. Zijlmans. (2015). "Measurement equivalence in mixed mode surveys." Frontiers in psicologhy 6, 1-11.

Krosnick, J. A. (1999). Survey research. Annual Review of Psychology, 50, 537-367. Annual Reviews

Link, M. W. and Mokdad, A. H. (2005). Effects of Survey Mode on Self-Reports of Adult Alcohol Consumption: A Comparison of Mail, Web and Telephone Approaches, Journal of Studies on Alcohol, 66(2): 239–45

Martin P. and P. Lynn, (2011). The effects of mixed mode survey designs on simple and complex analyses. Centre for Comparative Social Surveys. Working Paper Series. Paper n.04.

Rosenbaum P. R. and D. B. Rubin, (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, 70, pp. 41-55.

Schouten B., N. Shlomo, and C. Skinner, Indicators for Monitoring and Improving Representativity of Response. Journal of Official Statistics 27 (2011), pp. 231–253.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. Psychological. Bulletin, 133, 859–883.

Vandenplas, C., Loosveldt, G., and Vannieuwenhuyze, J. T. A. (2016). Assessing the use of mode preference as a covariate for the estimation of measurement effects between modes. A sequential mixed mode experiment. Method, data, Analyses. Vol. 10(2), 2016, pp. 119-142.

Vannieuwenhuyze, J. T. A., Loosveldt, G. and Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. Public Opinion Quarterly, Volume 74, Issue 5, 1 January 2010, Pages 1027–1045,https://doi.org/10.1093/poq/nfq059

APPENDIX A

For each considered variable listed in Figure 1 (section 4.1.), a plot with the percentage of answers, with the two data collection mode presented, was estimated. Furthermore, the p-value for the Chi-squared test and the t-test are reported.

The p-value can be defined as the smallest value of α, significance level[3], for which the sample estimates will lead to rejection of H_o. In practice:

$$\begin{array}{lll} \text{accept } H_o & \text{if} & \text{p-value} > a \\ reject\ \ H_o & if & \text{p-value} < a \end{array}$$

The analysis were conducted considering a significance level to 0.05. It is important to point out that for rejecting the null hypothesis using the t-test – that is a bilateral test – the p-value must be lower than $a/2$ (0.025). On the contrary, for the unilateral Chi-squared test it must be lower than 0.05. The case in which the difference between the estimates is significant (t-test) and the case in which the two distribution are different (Chi-squared test) with respect to the data collection mode, are in bold font.

The case in which the difference between the estimates is significant (t-test) and the case in which the two distribution are different (Chi-squared test) with respect to the data collection mode, are in bold font.

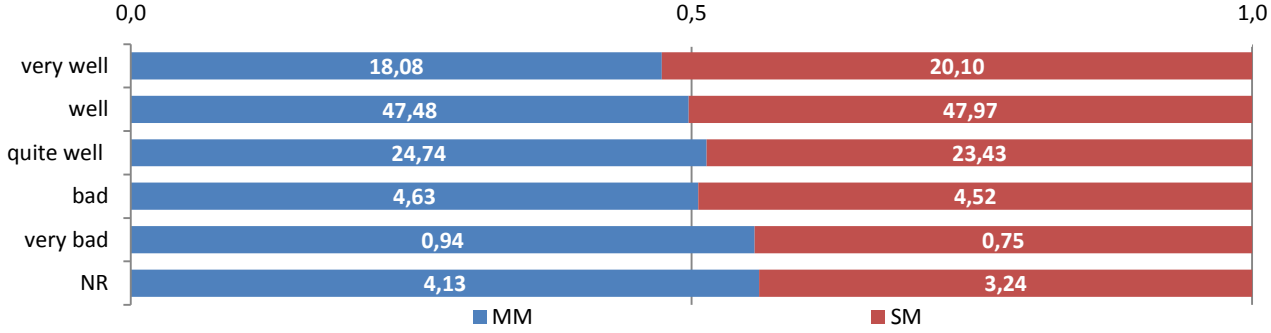
In the following figure the distributions of the responses for MM and SM and tests on significance of the difference in proportions for several survey target parameters are shown.

Figure A1. Distributions of the responses for MM and SM and tests on significance of the difference in proportions for several survey target parameters

Number of MM respondents	30746
Number of SM respondents	12175

	p-value											
	0	1	2	3	4	5	6	7	8	9	10	NR
t-test	0.0944	0.8902	**<.0001**	**<.0001**	**0.0447**	**<.0001**	0.6749	0.7956	**<.0001**	0.2378	**0.0338**	**<.0001**
χ^2-test		**<.0001**										

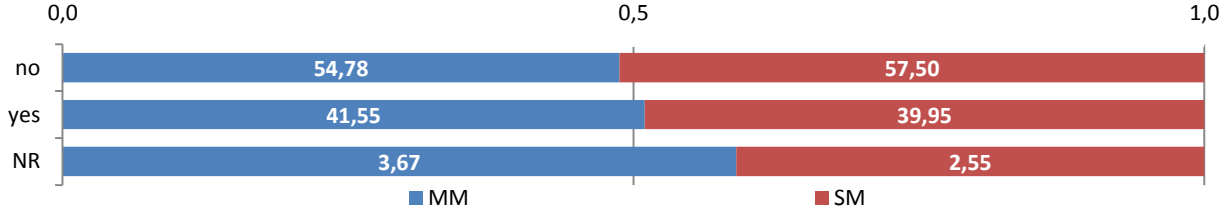
Health conditions



Number of MM respondents	35005
Number of SM respondents	13870

	p-value					
	very well	well	quite well	bad	very bad	NR
t-test	**<.0001**	0.2844	**0.0007**	0.5539	0.0250	**<.0001**
χ^2-test	**<.0001**					

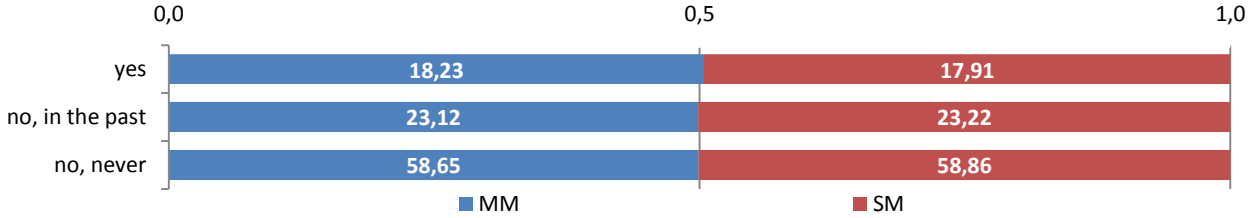
Reading books read in the last year



Number of MM respondents	33365
Number of SM respondents	13197

	yes	no	NR
t-test	**<.0001**	**0.0006**	**<.0001**
χ^2-test	**<.0001**		

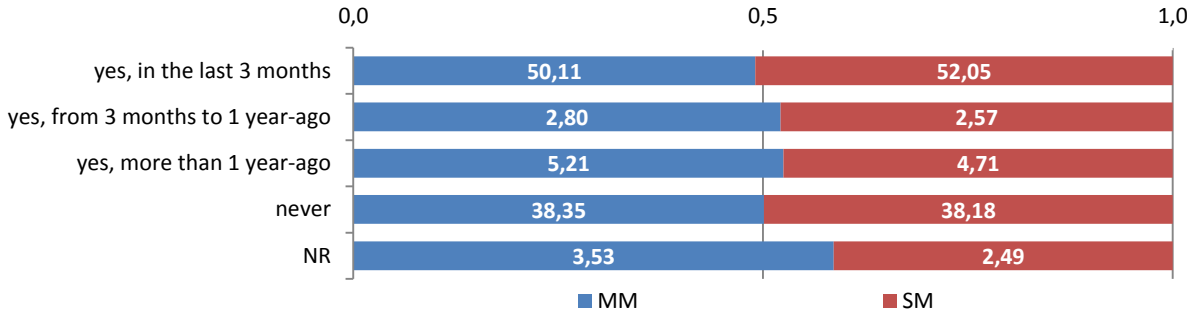
Habitual smoker



Number of MM respondents	30658
Number of SM respondents	12201

	yes	no	NR
t-test	0.6539	0.4766	0.1736
χ^2-test	0.5085		

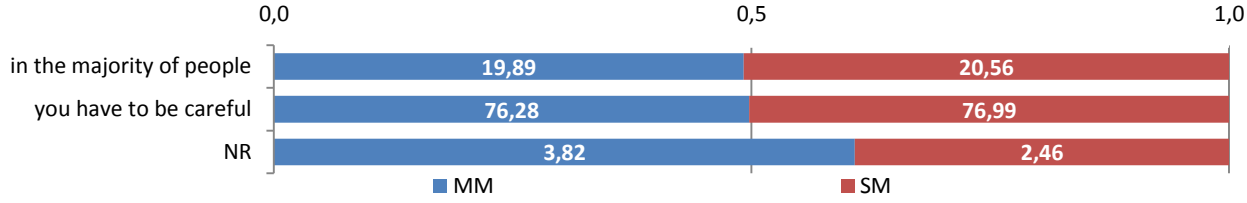
Use of internet



Number of MM respondents	34226
Number of SM respondents	13568

	yes, in the last 3 months	yes, from 3 months to 1 year-ago	yes, more than 1 year-ago	never	NR
t-test	**<.0001**	0.1205	**0.0141**	0.8186	**<.0001**
χ^2-test	**<.0001**				

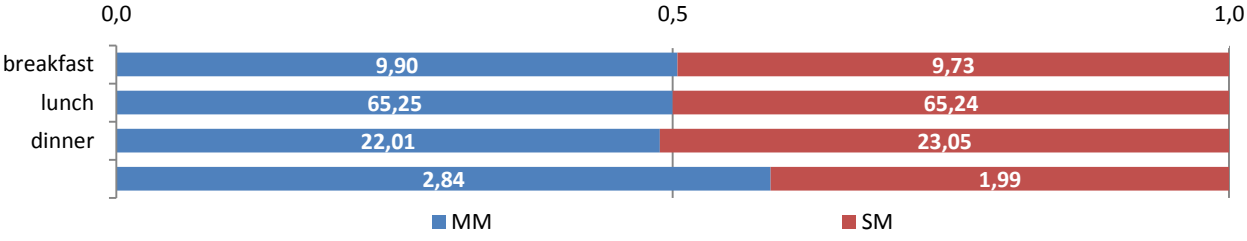
Trust in others



Number of MM respondents	30746
Number of SM respondents	12175

	in the majority of people	you have to be careful	NR
t-test	0.0923	0.1504	**<.0001**
χ^2-test	**<.0001**		

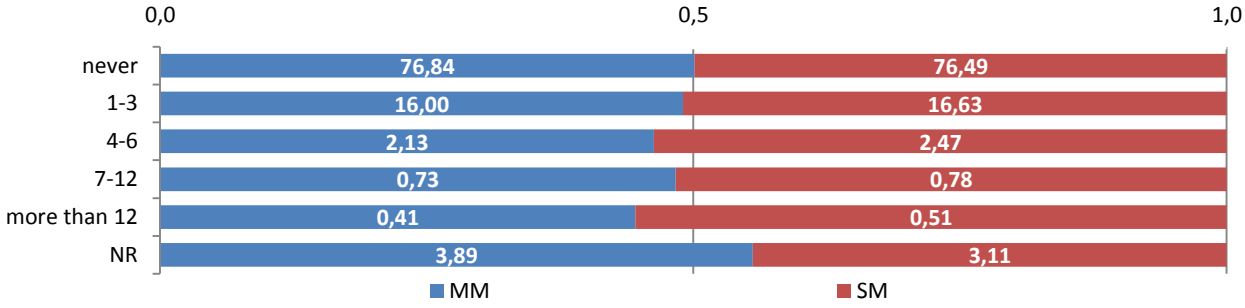
Main meal



Number of MM respondents	34226
Number of SM respondents	13568

	breakfast	lunch	dinner	NR
t-test	0.5716	0.8171	**0.0052**	**<.0001**
χ^2-test	**0.0009**			

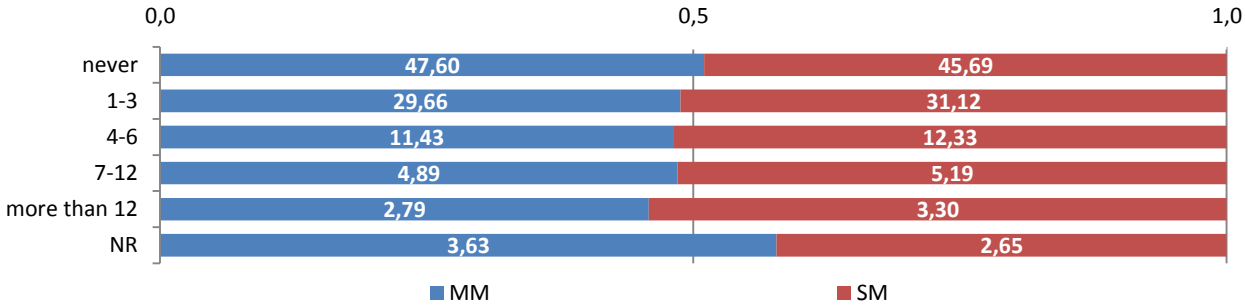
Number of times in the last year you went to theatre



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	0.4181	0.0685	**0.0149**	0.5043	0.1152	**<.0001**
χ^2-test	**0.0002**					

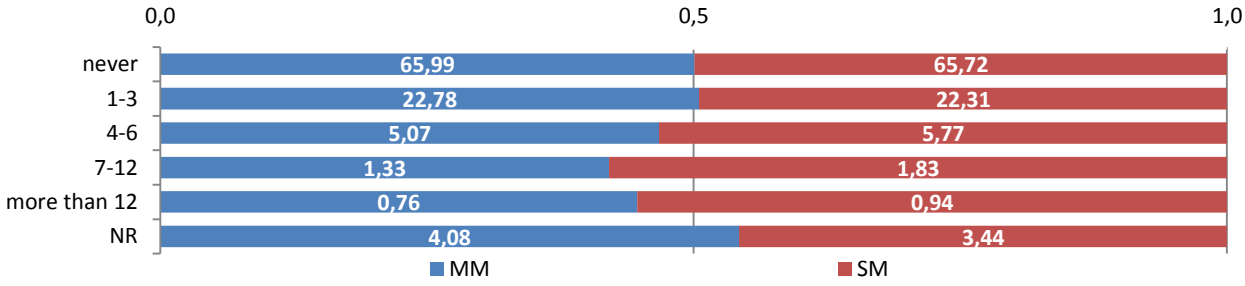
Number of times in the last year you went to cinema



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	**<.0001**	**0.0007**	**0.0028**	0.1488	0.1284	**<.0001**
χ^2-test	**<.0001**					

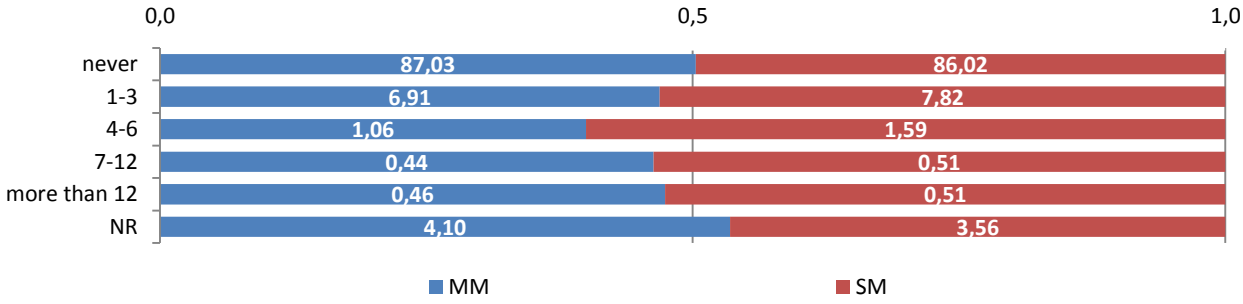
Number of times in the last year you went to museum



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	0.5628	0.2243	**0.0008**	**<.0001**	0.0340	**0.0003**
χ^2-test	**<.0001**					

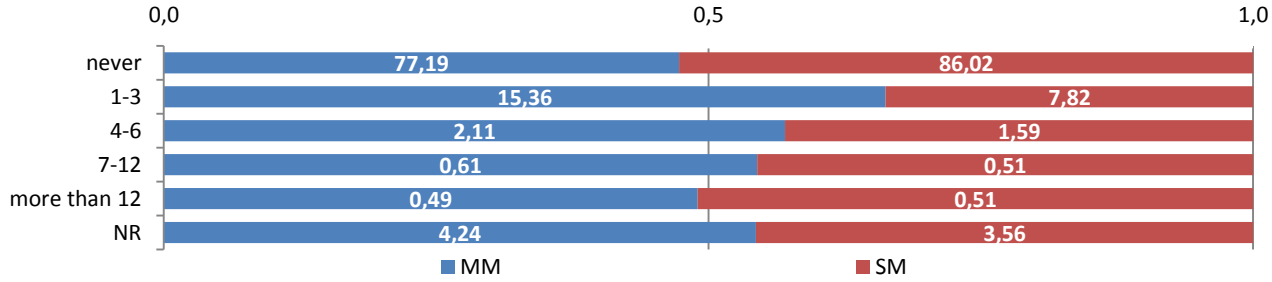
Number of times in the last year you went to concerts of classical music, opera



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	**0.0057**	**0.0002**	**<.0001**	**0.2957**	0.4669	**0.0013**
χ^2-test	**<.0001**					

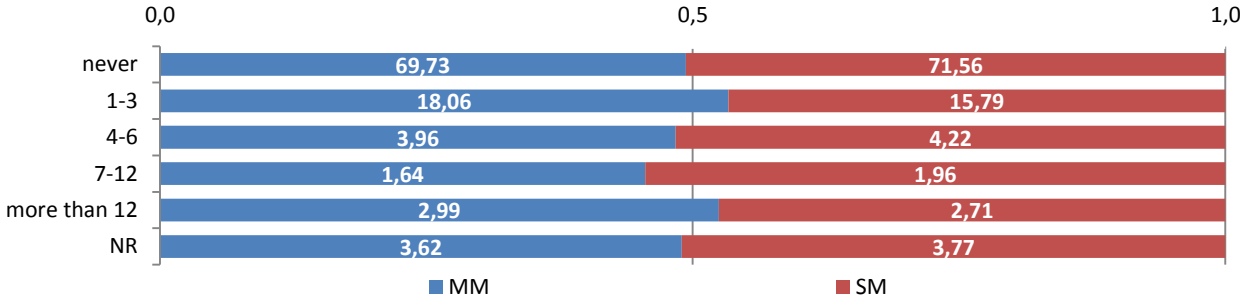
Number of times in the last year you went to other music concerts



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	**0.0057**	**0.0002**	**<.0001**	0.2957	0.4669	**0.0013**
χ^2-test	**<.0001**					

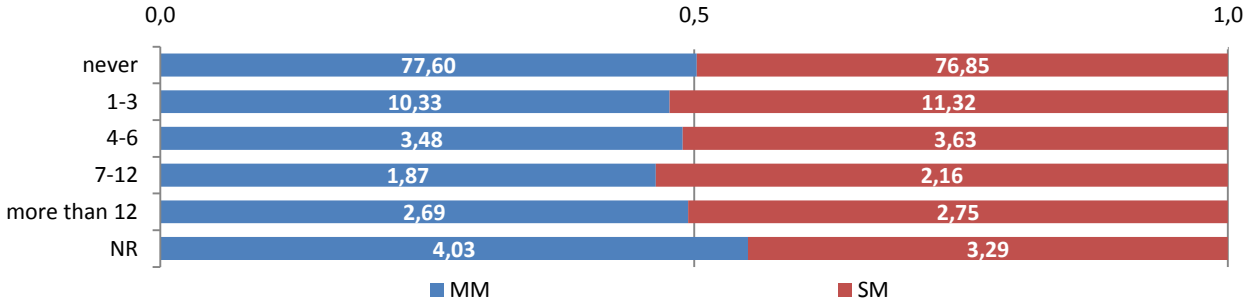
Number of times in the last year you went to went to sport events



Number of MM respondents	33365
Number of SM respondents	13197

	never	1-3	4-6	7-12	more than 12	NR
t-test	**0.0230**	0.1588	**0.0037**	**0.0009**	0.2073	**0.0004**
χ^2-test	**0.0007**					

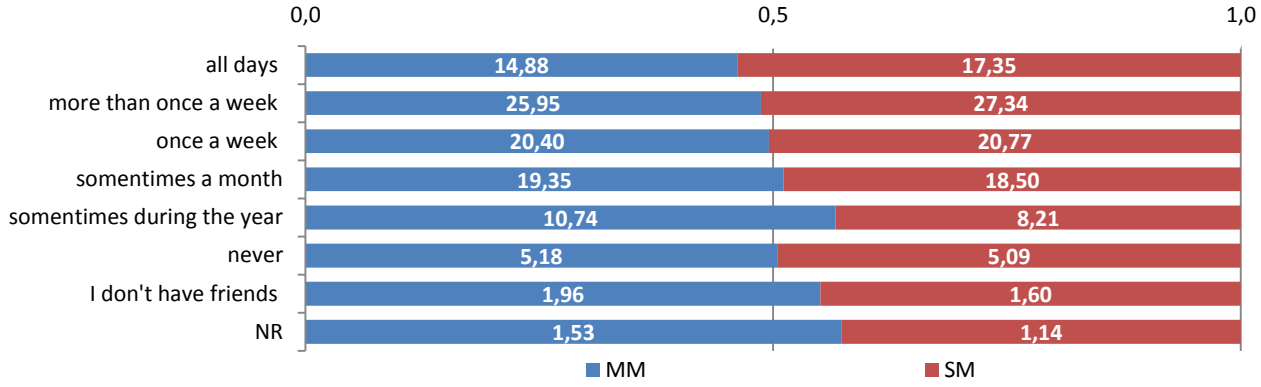
Number of times in the last year you went to went to disco or clubs



Number of MM respondents	33365
Number of SM respondents	13197

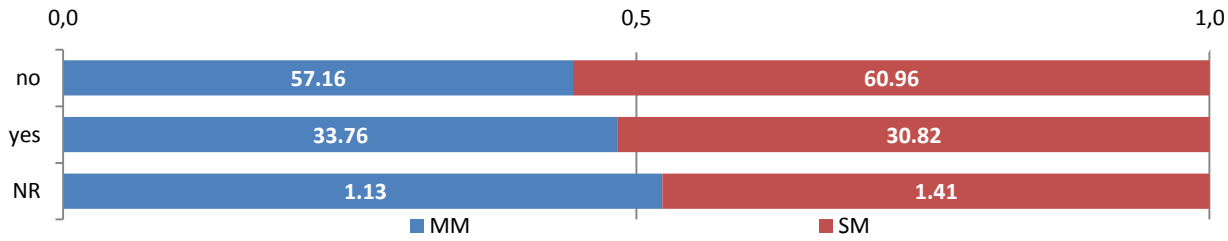
	never	1-3	4-6	7-12	more than 12	NR
t-test	0.0767	**0.0007**	0.3551	**0.0235**	0.7265	**<.0001**
χ^2-test	**0.0341**					

Frequencies of seeing friends



	p-value							
	all days	more than once a week	once a week	sometimes a month (less than 4)	sometimes during the year	never	I don't have friends	NR
t-test	**<.0001**	**0.0008**	0.3397	**0.0202**	**<.0001**	0.6559	**0.0032**	**0.0003**
χ^2-test	**<.0001**							

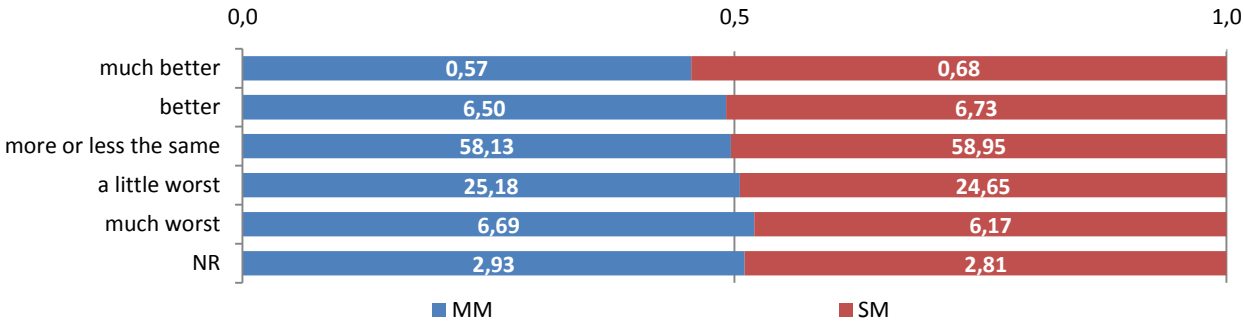
Continuity in sport activity



Number of MM respondents	34226
Number of SM respondents	13568

	no	yes	NR
t-test	0.9864	0.3169	0.0402
χ^2-test	0.4956		

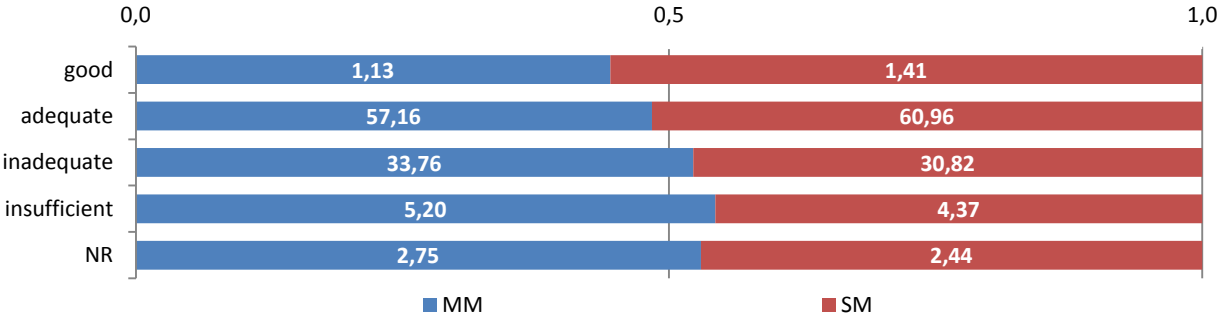
Economic situation with respect to the previous year (household level)



Number of MM respondents	15007
Number of SM respondents	5947

	much better	better	more or less the same	a little worst	much worst	NR
t-test	0.1396	**0.0121**	0.1685	0.1337	**0.0123**	0.2971
χ^2-test	0.0785					

Household economic resources level



Number of MM respondents	15007
Number of SM respondents	5947

	good	adequate	inadequate	insufficient	NR
t-test	0.7827	**<.0001**	**<.0001**	**0.0007**	**0.0109**
χ^2-test	**<.0001**				