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

WP2 – Deliverable 4

Methodological Report

Date: 20 February 2019

Orietta Luzi (Istat)
Claudia De Vitiis (Istat)
Francesca Inglese (Istat)
Roberta Varriale (Istat)
Alessio Guandalini (Istat)
Marco Dionisio Terribili (Istat)
Barry Schouten (CBS)

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

Contents

Preface

In mixed-mode data collection, particular attention has to be devoted to the survey design in order to both prevent as much as possible potential biasing effects due to the combined use of different modes, and properly treat such effects when they occur in final outputs. This is needed to ensure high quality statistics at affordable costs and low response burden.

Actually, well-designed mixed-mode surveys may reduce costs and non-sampling errors (coverage, nonresponse, and measurement errors). However, possible *mode selection effects* (resulting from errors of non-observation), and *mode measurement effects* (resulting from observation errors) can affect the survey results due to the combined use of different data collection modes. Mode effects need to be properly assessed and adjusted for in order to ensure accurate estimates. A wide range of methodological solutions and strategies have been proposed in literature to deal with these problems and improve the quality of the produced estimates.

Within the MIMOD Project, Work Package 2 (WP2 hereafter) addresses mode effects in mixed-mode survey designs with the purpose of investigating ways to deal with this issue (e.g. weighting, imputation, other data processing) and analyze differences in the final sample composition based on different modes across time, countries and survey types, providing practical evidence-based guidelines for the National Statistical Institutes (NSIs hereafter) in the European Statistical System (ESS).

To this purpose, within WP2 three main objectives have been pursued:

1. to provide an updated overview on methodologies for mode effect assessment and adjustment in mixed-mode designs, particularly those currently used in the ESS, with a discussion of assumptions, advantages and disadvantages of the various approaches. This review has been complemented with information on methods and strategies currently adopted in the ESS countries based on the MIMOD query which has been carried out in 2018;

2. to evaluate the suitability of selected statistical approaches and methods to deal with selection and measurement effects in mixed-mode data collection surveys based on practical applications and statistical analyses. In particular, the following methods have been applied on some current mixed-mode social surveys: 1) re-interview designs for mode effect estimation and adjustment, with a cost-benefit analysis for decomposing mode effects into selection and measurement components; 2) methods for treating mode bias/mode effects at the estimation stage;

3. to provide general guidance and assistance about methodological approaches which can be adopted to deal with mode effects in mixed-mode designs. Based on the results achieved through the analysis of recent literature, the MIMOD query outcomes and the practical application of selected methods, general operational and evidence-based advices and suggestions for the use of methodologies to deal with mode effects in mixed-mode surveys have been elaborated.

The results of WP2 are expected to provide all ESS countries not only with an updated overview about methodological solutions to improve the quality of estimates produced in mixed-mode surveys, but also with a tool - represented by a set of guidelines - that could support them in properly design methodological strategies to properly deal with mode effects.

Three technical deliverables have been produced by the partner countries involved in this work package (the Italian National Statistical Institute – Istat – and Statistics Netherlands – CBS): an updated literature review and current status of detection and adjustment methodology (Deliverable 1), a cost-benefit analysis of re-interviews based on two CBS case studies (Deliverable 2), and an application of a subset of these methods to an ISTAT case study (Deliverable 3). This report contains a summary of the results achieved within WP2 and described in detail in these deliverables, as well as general guidance and advises derived from the analyses carried out and the results obtained during the project.

The reader of this report is expected to be familiar with basic concepts of survey sampling in general and mixed-mode designs in particular.

The report is structured as follows.

Section 1 contains a summary of the update of the literature review on methodologies for mode effect assessment and adjustment, including the main results from the query conducted in the MIMOD project.

Section 2 reports summaries of the applications of selected methods and approaches on current mixed-mode social surveys which have been carried out within WP2.

Section 3 contains general guidelines based on a schematization of the approaches and methods that can be adopted to assess mode effects and/or to adjust for mode effects in mixed-mode surveys, consisting in a general discussion about how to choose and use these methods and approaches in practical survey contexts.

Concluding remarks close the report (Section 4).

1. Mode bias/mode effects and adjustment for mode-effects: the state-of-the art

1.1 Introduction: mixed-mode surveys and mode effect

The increasing use of the web for data collection has driven NSIs to move their surveys from single to mixed-mode designs in which web is combined with traditional survey modes. Given the low cost and the relatively short return times of web surveys, and despite their low response rates, mixed-mode designs involving the web are now becoming rule rather than exception, especially in social surveys.

Mixed-mode designs can employ multiple data collection modes in different ways. A first classification of mixed-mode designs can be made regarding the choice of modes: does the agency conducting the survey assigns sample units to mode groups, or can the sample units choose the mode through which they respond to the survey? A second classification is based on a distinction between designs in which each respondent can only respond through a single mode (assigned or chosen), and designs in which different modes are offered to the same respondents. Mixed-mode designs in which multiple modes are used simultaneously are known as *concurrent designs*. In contrast, *sequential designs* use one mode first and then re-approach nonrespondents using a different mode; combinations with more than two modes are also possible.

All mixed-mode surveys, regardless of their precise design, result in a bipartition of the sample into respondents and nonrespondents. The respondents have provided answers to the survey questions, and not all of them did so through the same data collection mode. This phenomenon can give rise to *mode effects*. The term *mode effect* is used differently in different contexts, and in its most general form refers to effects that are due to the use of one mode compared to another, or a combination of modes to a single mode, or to a different combination of the same or other modes. Effects of this kind manifest themselves in the survey outcomes, typically estimates of population means and totals. Mode effects are related to bias and variance of the estimators of the survey variables.

In WP2 two kinds of mode effects are distinguished. First, *selection effects* are caused by the selection mechanism of a mixed-mode survey design which results in partitioning the sample into respondents and nonrespondents. Selection effects are a combination of coverage and nonresponse effects. Second, *measurement effects* are caused by specifics of the modes employed in the survey and affect the recorded responses to the survey questions. They arise from the same respondent potentially giving different answers to the same questions in different modes. Sometimes measurement effects are referred to as *measurement bias*, or as *pure mode effects*. Often, only a joint mode effect can be observed, which is the combined effect of selection and measurement effects. Apart from experimental designs, selection and measurement effects are generally *confounded* and are difficult to separate.

The WP2 concentrates on methods for the detection, estimation and adjustment of mode-specific biases in survey statistics.
The first activities carried out within WP2 focused on the review of recent literature on methodologies to assess and/or to adjust for mode effects (Buelens *et al*., 2018a), and on the analysis of the outcomes of the MIMOD survey held among statistical agencies in ESS countries about their mixed-mode experiences and activities. The aim of the next two sections is to provide a summary of these activities.

1.2 Update of literature review

Comparisons between surveys conducted using different data collection modes are available in the literature almost from the time when sample surveying became common practice. It seems to be the emergence of web technology that has instigated renewed interest in research into the effects of using different modes of data collection. The year 2005 appears to mark the onset of this latest wave of interest, with particular attention to the combined use of multiple modes in the same survey. In that year, some often-cited articles were published. De Leeuw (2005) lists advantages and pitfalls of mixing modes. Voogt and Saris (2005) discuss the trade-off between improved selection and possibly hampered measurements in multi-mode surveys. Dillman and Christian (2005) recognize the issue of differential measurement effects between modes and suggest preventing this issue through the design of questionnaires that prevent this phenomenon from occurring. Fricker *et al*. (2005) conducted an experiment to compare web and telephone surveys.

Mode assessment studies are sometimes limited to quantifying the total mode effect, but are more insightful when they separate the total effect into selection and measurement components. In Deliverable 1 (Buelens *et al*., 2018a) the approaches that have been published in the literature are listed and commented. Experimental designs specifically aimed at studying mode effects are preferable, but costly, and hence less common. These include parallel, independent surveys, or re-interview studies. Some authors report methods to separate mode effects in observational studies, usually relying on socio-demographic covariates that explain the selection mechanism. Using such covariates, approaches like reweighting or sample matching have been reported.

Less common than mode assessments are mode adjustments. Adjustment techniques are aimed at correcting survey estimates for bias induced by one or several modes, or by the specific combination of several modes. Adjustments for bias require the presence of a definition – or choice – of reference mode or design that serves as a benchmark, since bias of some design is only meaningful with respect to some other design. Adjustment techniques that have appeared in the literature include reweighting and calibration approaches, imputation, and prediction approaches. Faithful adjustment methods require mode effects to be separable into selection and measurement effects, which is most successful in experimental designs. References to specific articles are provided in WP1 deliverable 1 Buelens *et al*. (2018a).

1.3 Key results from the MIMOD survey

Within the MIMOD project, a survey was held among statistical offices in ESS countries. The survey contained questions on data collection strategies, questionnaire design, use of smartphones and tablets, methods to deal with mode effects, and case management systems. Here, we report on answers received to the questions in the section on methods to deal with mode effects. For a comprehensive analysis of the survey see Deliverable 1 of WP1 (Murgia *et al.*, 2018).

Responses to the MIMOD survey were received from all 31 countries in the survey: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, The Netherlands, United Kingdom.

Table 1 summarizes the key results of the query in terms of methods to deal with mode effects in mixed-mode social survey adopted by the responding ESS Agencies.

Concerning activities undertaken to assess mode effects, one third of the countries did not conduct any assessments of mode effects in their social surveys, as can be seen from the last row of the table. Out of the activities undertaken by the Agencies who did conduct some assessments, pre-testing or experiments with questionnaire designs are most common. Other often conducted assessments include pre-testing or experiments with sensitive or core questions, conducting pilot surveys, comparing distributions in socio-

demographic or target variables, comparing various quality indicators, and parallel runs of different data collection strategies.

It cannot be seen how many activities were undertaken in a combined fashion by individual Agencies. Analysis of the results learns that most agencies who report at least some activity did actually undertake several activities. Countries reporting only a single activity are exceptions.

As for the measures taken to adjust for mode effects, it results that two thirds of the Agencies have not taken any measures so far. The minority of countries that have taken measures did so predominantly by weighting corrections. Only a few countries applied calibration or correction adjustments. Finally, 14 of the 31 countries report to have future plans for research into mode effect assessment and/or adjustment methods. Most of these plans focus on assessment and to a lesser extent on adjustment. The plans for assessments are often quite rigorous in that they involve pilot studies, experimental designs or parallel execution of different strategies. Some Agencies anticipate the need for mode effect adjustments, but none report to have plans for research into adjustment strategies specifically. The plans involve mostly empirical and applied research.

Table 1. Activities undertaken by 31 responding ESS countries to deal with mode effects in mixed-mode designs

Objective	**Activity undertaken**	**Percentage of countries**
Assess mode effects	Pre-tests, experiments on questionnaire design	48 %
	Pilot surveys	42 %
	Differences in distributions of socio-demographic or target variables	39 %
	Differences in quality indicators (e.g. total or item non response rates, break-off rates, reliability indicator, failure rates of consistency rules, …)	35 %
	Pre-tests, experiments on sensitive or core questions	35 %
	Previous and new data collection strategies running simultaneously (independent sampling)	32 %
	Separating selection, nonresponse and measurement effects	26 %
	Calculation of representativeness indicators of various designs	23 %
	Pre-tests, experiments on split sample approach	19 %
	Subsampling of groups receiving different data collection strategies (e.g. control group)	19 %
	Pre-tests, experiments on the use of different devices (smartphones, tablets, …)	19 %
	Re-interview studies	6 %
	Other types of pre-tests and/or experiments	3 %
	Other activities	6 %
	No activity conducted in recent years	*32 %*
Adjust for mode effects	Weight adjustments	26 %
	Calibration to fixed mode distributions	13 %
	Estimate measurement errors and correct responses to a benchmark mode	10 %
	Other	13 %
	No measure taken	*61 %*

The ESS country experiences reported in the MIMOD survey reflect findings in the literature reviews on methods for mode effect assessment and adjustment. Both reported activities and published literature on mode effect assessments are more widespread than on mode effect adjustment techniques. Sometimes assessment of mode effects may be sufficient, but when detected, some effects may need to be corrected for, in particular measurement effects.

While in our view a distinction between selection and measurement effects is essential to make, this is not always done in the literature on mode effect assessments. An important reason is that it is difficult to separate selection from measurement effects, but easy to assess their combined effect. The main difficulty is the confounding of selection and measurement effects in observational studies. The two effects can be separated in experimental studies, but these are rather rare because of costs.

Since separating selection from measurement effects are a prerequisite for successful mode effect assessments and adjustments in mixed-mode designs, a promising line of future research is the development of mixed-mode designs that allow for this, for example through embedded experiments. An example of such a design consists of conducting re-interviews through a second mode for a subset of respondents who already responded through a first mode (Klausch *et al*., 2018). Alternative designs that allow for separating measurement and selection effects, and for which suitable mode adjustment estimators can be defined, are expected to appear and would deliver a very valuable contribution to the practical usability and theoretical validity of mixed-mode sample surveys.

Some approaches for mode effect assessment and adjustment have been actually explored in the context of WP2, as described in section 2.

2. The applications

2.1. Introduction

Survey methodology offers three options to deal with mode effects when data collection modes are combined. They can be prevented through questionnaire design, e.g. Dillman *et al*. (2014), avoided through data collection design, e.g. Schouten, Peytchev & Wagner (2017), and adjusted through estimation design, e.g. Klausch *et al*. (2018). In WP2 the focus is on the last two options, although estimates of measurement bias may inform questionnaire redesigns. Actually, even if accurate mixed-mode questionnaire design is the most important and effective option to reduce mode effects, it is not capable of removing all mode-specific measurement biases.

The estimation and evaluation of mode-specific measurement bias is fundamentally hard due to the confounding with mode-specific selection bias. Among the methods to detect mode-specific biases and to adjust such biases, which are illustrated in Deliverable 1 of WP2 (Buelens *et al*., 2018), some specific methodologies have been selected and applied to real social surveys within WP2:

- a cost-benefit analysis to re-interview designs to optimize re-interview designs and to estimate mode-specific measurement biases in two Dutch surveys: the Dutch Health Survey and the Dutch Labor Force Survey (see WP2 Deliverable 2 - Buelens *et al*., 2018);
- some methods to assess mode effects and adjust for measurement effects to the Italian Aspects of Daily Life Survey (see WP2 Deliverable 3 - De Vitiis *et al*. (2018).

In the following sub-sections, the summaries of the applications which have been carried out and the main results achieved are reported. A discussion of the main evidences emerged from the applications made is reported as well.

2.2 Cost-benefit analysis of re-interview designs for mode-specific measurement bias

One option to estimate, and potentially also adjust, mode-specific measurement biases is through so-called re-interview designs. Re-interview designs re-approach respondents to one or modes by another mode. As a result, two measurements are available for part of the respondents in different modes. The two measurements are used to estimate biases. This can be done in two ways, a direct and an indirect option. The direct option estimates mode-specific measurement bias. The indirect option estimates mode-specific selection bias first and then deduces the mode-specific measurement bias by subtracting the selection bias from the total mode bias. The two options will be explained briefly.

Re-interview designs are typically used in sequential mode designs where the more expensive interviewer-assisted modes follow self-administered modes. They are less suited for detecting measurement biases in concurrent mode designs where the mode choice is up to the respondent. This is because the respondents would be forced to also answer parts of the survey in mode they did not choose. They can, however, be applied to designs where modes are assigned concurrently but without a respondent choice, such as telephone for those sample persons that have a registered phone number and face-to-face for those sample persons that do not. In the latter case, telephone respondents are also allocated to face-to-face. To fix thoughts, we give three examples:

- in a sequential web – telephone design, the web respondents are re-approached by telephone at the same time as the telephone follow-up to the web nonrespondents;
- in a concurrent telephone – face-to-face design with allocation based on phone number registration, telephone respondents are also approached face-to-face;

- in a sequential design with three modes web – telephone – face-to-face, both web respondents and telephone respondents are re-approached by face-to-face;

We must stress that a re-interview does not guarantee that mode-specific measurement biases can be estimated accurately; they require assumptions. The designs assume that re-interview respondents are unaffected by the first mode contact and interview. Furthermore, the re-interview itself leads to nonresponse and it must be assumed that this is not related to the difference in measurements between the two modes. These assumptions may not always hold, even with careful timing of the re-interview and with careful introduction of the purpose of the re-interview. A natural presentation of the re-interview survey combined with a mix of repeated and new questions are crucial. In order to avoid context effects as much as possible, the first part of the survey contains the repeated survey questions without and changes.

If the assumptions hold, then the biases can be estimated in two ways. In the direct option, the two answers to a repeated question are compared, the measurement differences are modelled and the estimated measurement bias model is applied to predict answers of those not in the re-interview, i.e. the nonrespondents to the first mode and to the re-interview. In the indirect option, the response to the first mode is calibrated to the combined response to the re-interview mode and differences between the unadjusted and adjusted estimates are attributed to mode-specific selection bias. The estimated selection bias is then subtracted from the total bias to arrive at an estimate for the mode-specific measurement bias.

To proceed, both the direct and indirect options suppose that there is a benchmark design relative to which biases arte estimated and possibly adjusted. Such a benchmark design consists of a benchmark for selection and a benchmark for measurement. The common practice, e.g. Schouten *et al* (2013), is to assume that the mixed-mode design is the selection benchmark but one of the modes is the measurement benchmark. We again look at the examples:

- in the web – telephone sequential design, the web – telephone selection may be treated as selection benchmark and web may be treated as the measurement benchmark;
- in the telephone – face-to-face concurrent design, the telephone - face-to-face selection may be benchmark and face-to-face may be treated as measurement benchmark;
- in the web – telephone – face-to-face design, the full mixed-mode selection may be considered the benchmark and face-to-face may be treated as measurement benchmark;

Obviously, a re-interview introduces an extra element to the overall design and, consequently, implies an investment. WP2 Deliverable 2 explores whether the investment is worthwhile and performs a cost-benefit analysis, using the Dutch Health Survey and the Dutch Labour Force Survey as case studies. It concludes that the investment may be worthwhile for the Health Survey, but not for the Labour Force Survey, for reasons we will explain later.

In the cost-benefit analysis, four scenarios are considered to compare estimates unadjusted for measurement biases and estimates adjusted for such biases. The scenarios follow from crossing two conditions. The first condition is that measurement biases are assumed constant for a specified time period versus time-varying. Under the time-varying setting, the biases need to be re-estimated for each wave, whereas for the time-independence setting they are estimated only once at the starting wave. The investment for the time-varying option is, obviously, much larger. The second condition is the quality criterion adopted by the main stakeholders of the survey. This condition is motivated by the loss of precision that comes from the measurement bias adjustment as effort is partially redirected to estimate additional parameters in the measurement models. The condition has also two settings. The first setting is that stakeholders view the mean square error of the final estimate as the criterion to judge quality, i.e. they are willing to weigh a reduced precision against a reduced bias. The second setting is that stakeholders constrain the precision to be

the same after adjustment for measurement biases. In general, the variance constraint setting implies that the re-interview will demand extra budget. In Deliverable 2, these four scenarios are explored.

The accuracy of the measurement bias adjusted estimate of a survey variable depends on a number of parameters: the actual size of the measurement bias for the survey variable, the correlation between repeated measurements of the survey variable in time, the sample size, the response rates to the modes of interest, the re-interview subsampling probability and the follow-up subsampling probability. The two subsampling probabilities determine the proportion of re-interview respondents and nonrespondents that are allocated to the re-interview and follow-up, respectively. The two probabilities are under the control of the re-interview designer. They may stratified based on known population characteristics beforehand. However, typically, they are chosen equal for all units as (usually) no prior knowledge is available about variation in mode-specific measurement biases between population subgroups/strata. Obviously, part of the parameters, notably the mode-specific measurement bias, is unknown beforehand. The strategy that is adopted in Deliverable 2, is to set a range of plausible values for the mode-specific measurement bias and to assess under what values the adjusted estimate is superior to the unadjusted estimate. In constructing the range of values, it is assumed that historic estimates for the total mode bias, i.e. the compound of mode-specific selection bias and mode-specific measurement bias, are available. We refer to Deliverable 2 for details. Deliverable 2 assumes that mode response rates can also be estimated from historic survey data, and it supposes that correlations can be guessed by experts. Because of these guesstimates, the study does have a subjective nature which must be accounted for in the evaluation.

We summarize the results of the two case studies. Both the Health Survey (HS) and the Labour Force Survey (LFS) are repeated monthly surveys. The HS has a sequential web - face-to-face design and the LFS a hybrid design where web is followed by telephone and face-to-face. LFS web nonrespondents with registered phone number are sent to telephone and otherwise allocated to face-to-face. In the evaluation, for the sake of simplicity, telephone and face-to-face are treated as a single mode in the LFS. Table 2 contains the estimates for the LFS key variable, unemployment, and for four key HS variables. Also given is the anticipated correlation between interview and re-interview, assuming a time lag of six weeks.

Table 2: Selected survey outcome variables with estimates per mode. Also provided is the estimated/anticipated reliability (correlation between repeated measurements).

	Survey	*Estimate* m_1	*Estimate* m_2	*Correlation*
Unemployment rate	LFS 2014-2015	5.6 %	6.7%	0.5
% good health	HS 2014	78.0%	75.6%	0.9
% smoker	HS 2014	19.9%	29.8%	0.9
% obese	HS 2014	12.1%	13.9%	0.9
% visit to dentist	HS 2014	82.3%	74.5%	0.7

The budget is frozen to be the same as the total budget of *T+1* waves. *T=0* means that biases are time-varying and are re-estimated each wave. For the HS and LFS, in addition, *T=3, T=7* and *T=19* are considered where a wave consists of a quarter, i.e. time periods of one year, two years and five years. Here, we show only the results for *T=3, T=7* and *T=19.*

Table 3 contains the results for the MSE criterion for the two measurement benchmarks and, per variable, three measurement bias levels. See Deliverable 2 for the values of the levels. It was concluded that for the LFS the unadjusted often outperform the adjusted estimates, or are close. For the HS, the picture is the other way around.

Table 3: RMSE values (in %) for the HS and LFS survey variables per time period and relative measurement bias level. Highlighted values in blue have the lowest RMSE. Highlighted values point at the preferred survey design under the scenario 1 adjustment perspective.

a) benchmark $BM = m_1$.

	T	LFS			HS											
		Unemployment ME bias level			Health ME bias level			Smoking ME bias level			Obesitas ME bias level			Dentist ME bias level		
		Left	Mid	right	left	mid	right	left	mid	right	Left	mid	right	left	mid	Right
adjusted	3	0.7	0.7	0.7	1.1	1.1	1.1	1.1	1.1	1.1	0.8	0.8	0.8	1.1	1.1	1.1
	7	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.8	0.8	1.0	1.0	1.0
	19	0.6	0.6	0.6	0.9	1.0	0.9	0.9	0.9	0.9	0.7	0.7	0.7	1.0	1.0	1.0
not adjusted	-	0.5	0.2	0.5	1.7	1.3	1.5	1.6	4.0	3.1	1.3	1.0	1.2	3.9	3.1	2.3

b) benchmark $BM = m_2$.

	T	LFS			HS											
		Unemployment ME bias level			Health ME bias level			Smoking ME bias level			Obesitas ME bias level			Dentist ME bias level		
		Left	Mid	right	left	mid	right	left	mid	right	Left	mid	right	left	mid	Right
adjusted	3	0.7	0.7	0.7	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.7	0.7	1.0	1.0	1.0
	7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.9	0.9	0.9
	19	0.6	0.6	0.6	0.8	0.8	0.8	0.9	0.8	0.8	0.6	0.6	0.6	0.8	0.8	0.8
not adjusted	-	0.7	0.2	0.6	1.8	1.3	1.6	5.3	4.3	3.3	1.4	1.0	1.3	4.3	3.3	2.4

Under the alternative setting, the precision needs to be the same after adjustment. Table 4 presents the required increase in budget to guarantee the same precision after adjustment for mode-specific measurement bias. The smallest increase in budget, 41%, is for the HS under the face-to-face benchmark and a five year time period (*T=19*). This still means a sizeable increase in budget. The largest increase if for the LFS with a year time period and web as measurement benchmark, almost 200%. From the evaluation, it was concluded that the precision constraint leads to unrealistic increases in budget.

Table 4:Relative increase in required budget per benchmark, time period and survey.

	Health survey						Labor Force Survey					
	$BM = m_1$			$BM = m_2$			$BM = m_1$			$BM = m_2$		
	T=3	T=7	T=19	T=3	T=7	T=19	T=3	T=7	T=19	T=3	T=7	T=19
ΔB	148%	106%	67%	80%	61%	41%	199%	139%	85%	146%	105%	66%

Deliverable 2 advocates to replicate findings for other case studies, but argues there is a positive business case under certain setting and for certain surveys.

2.3 Methods to asses and adjust mode effect on a social survey

Deliverable 3 of WP2 (De Vitiis *et al*., 2018) illustrates 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 WP2 first deliverable (Buelens *et al*., 2018a).

It is worth mentioning that 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. Therefore, the surveys based on mixed mode must be designed and carried on 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 modes, suspected to be responsible of mode effect.

As described in Deliverable 1 (Buelens *et al*., 2018a), 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. The 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 approach 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 the considered 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 analyses is the evaluation of the impact on final estimates of the switching from single to mixed mode 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, focusing only on the mixed mode sample (web-PAPI), a comparison is made between estimates obtained using different methods for adjusting mode effect (weighting/calibration and multiple imputation).

The set of 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, as outlined in Table 5.

Table 5. Synthesis of the experimental context

General Survey Contest	Experimental : Parallel independent samples (single mode SM, mixed mode MM)
Mixed mode specific context	Sequential web-PAPI; PAPI is the single mode
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	Register demo-social covariates
Steps of the analyses	– Comparison between the SM and MM samples – Assessment of the mode effect in the MM design

	– Adjusting for mode effect in the MM design

The analyses in the specific survey context

The analyses carried out on the survey data aimed to evaluate firstly the impact on the survey estimates of the introduction of mixed mode design with respect to the previous single mode design and, subsequently, to analyze in depth the reasons that determine significant differences in the estimates obtained with the two designs. For this purpose, the study was developed on several levels of analysis, corresponding to different operational steps:

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 using web and PAPI in the MM design;
3. the third level consists of some experiments to adjust for mode effect using the MM data.

In the first step 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.

Subsequent analyses were conducted to study the bias caused by the total nonresponse in the two samples. To this end, auxiliary variables were acquired from administrative archives on the individuals included in the samples and redefined at the household level as the household is involved in the response process and in the "choice" of the mode. The response processes were analyzed and some indicators of representativeness were evaluated 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 nonresponse 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 at 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 nonresponse with respect to a benchmark estimate. Moreover, in order to estimate the measurement and selection effects in the MM sample, a method that takes the SM survey as a benchmark is experimented.

In 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.

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

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 *et al.*, 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 adjustment methods.

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. In Table 6 the steps and the methods considered in the study are listed.

Table 6. Operational steps of the analysis

	Method	**Objective**	**Assumptions/Conditions**
First Step	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 Step	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 Step	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

The analyses carried out brought out several evidences deriving from the introduction of the mixed mode in a social survey. The results show that in the mixed mode survey, the bias due to the total nonresponse is reduced, confirming what stated in Deliverable 1 of WP2 (Buelens *et al*., 2018a). 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.

If the objectives of cost reduction and of better population coverage are achieved, the quality of some of the produced estimates seems affected by a measurement effect, 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 design. 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.

When the assessment of mode effect is carried out for specific target 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 a 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.

Regarding the adjustment methods experimented, what can be underlined is that the application of weighting correction (both based on propensity score and calibration) brings to close estimates. When, on the contrary, one tries to utilize imputation (standard multiple imputation in the analyses carried out) the outcome can be strongly different and difficult to be interpreted, also considering that the PAPI technique has been considered as a reference mode (benchmark), as not affected by measurement error, which in general is a strong assumption.

So in order to better define an adjustment in the cases where a double measure is not available (re-interview or register data), more sophisticated and resource consuming method should be applied, such as the method proposed in Suzer-Gurtekin *et al*. (2012).

Regarding the evaluation of measurement effects, the diagnostic method *multi-group confirmatory factor analysis* has been applied. The analysis 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, and highlighted the presence of measurement invariance for the analyzed variables. The results provide useful advices for the planning of future editions of the survey.

The outcomes presented in Deliverable 3, finally, would need a significance assessment, based on tests or replication methods, which have not been carried out but that in general are needed to complete the assessment.

From the experience made, 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, whose quality seems to be affected by measurement effect which cannot be always easily assessed. 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.

It is obvious that the application of all the presented methods is subject to the validity of the hypotheses underlying them, and that are to be properly verified by the researcher. Besides, the results obtained by the applied methods depend 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, the analysis process carried out and the results illustrated in Deliverable 3 highlight that the effort required to carry out such studies highly overcomes 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 more advisable, in order to prevent as much as possible the measurement effect, which is the main drawback of mixed mode surveys.

3. General discussion

3.1 Introduction

This section provides some general guidance related to the design of strategies to control for potential mode bias/mode effect in mixed-mode surveys.

In general, in deciding if and how to estimate mode effects and/or to adjust for their biasing effects on survey results, there are three key decisions to be made:

- the quality criterion (e.g. the MSE) against a cost limit: how to assess whether mode effect adjustment is beneficial?;
- the multi-dimensionality of a survey: what key estimates and population parameters of interest need to be evaluated?;
- the time perspective: is the survey repeated and can effects be assumed constant?

Without a consensus on how quality and costs are quantified, it is, generally, hard or impossible to make an objective choice between unadjusted and adjusted estimates. Since true values are often unknown, one inevitable sub-question is what mode is chosen as benchmark for measurement. In other words, to what benchmark is the adjustment made. The answer to this question may ideally be different for different survey variables. However, in practice, a single choice has to be made.

Surveys obviously contain many questions, so that it is imperative that stakeholders select the most crucial variables in order to limit complexity of decisions.

As alluded to in the re-interview case study in section 2.2, it makes a big difference when surveys are repeated and decisions to adjust may stretch over a longer time period.

When defining a methodological strategy to deal with mode effect estimation and adjustment, there are essentially three main requirements to be defined:

- a design,
- auxiliary data (from administrative data/frame data/paradata) referred to as *covariates*,
- a set of assumptions.

Concerning the design, Table 7 below summarizes the type of design, experimental and non-experimental, within which it is possible to carry out analyses aimed at either assessing or adjusting selection and/or measurement effects.

Experimental designs allow controlling for selection effects, and hence the unbiased assessment of measurement differences between modes. An *experimental design* is of course optional and not standard practice.

Observational studies require covariates that explain the selection mechanisms. If available, differences between mode groups are attributed to measurement differences, conditional on the *covariates*. Validating this assumption can be achieved when variables that are observed without error are available, potentially obtainable from external data sources.

There are two types of *auxiliary data*: data informative about selection into modes and data informative about measurement within a mode. In causal inference literature, the two have been referred to as backdoor and frontdoor variables and may be employed to improve external and internal validity. Data informative of selection typically consist of linked frame data and administrative data and paradata from the contact and participation processes. Data informative of measurement consist of record check or validation data, repeated measurements and paradata from the answering process.

Table 7. General scheme of survey settings and objective of the analyses

Design type	Objective
Experimental	
Parallel independent surveys (single mode and mixed mode)	Mode assessment Mode adjustment
Re-interview study - repeated measurement designs	
Other (Embedded experiments, Split sample designs)	
Non-experimental	
Observational studies (Mixed-mode design only)	Control for selection effects through weighting or regression-based inference methods Adjusting for measurement effect

Assumptions may be divided into three types: assumptions about the explanation of the missing data mechanism due to mode selection, assumptions about the explanation of measurement differences due to modes, and assumptions about the absence of experimental influence on (non)respondents in experimental designs.

It is straightforward to mention that when the available covariates do not fully explain the selection mechanism, the decomposition of the total mode effect into selection and measurement effects is incorrect.

Concerning the *assumptions*, they depend on the type of auxiliary data and type of design. It is straightforward to mention that, with the same auxiliary data and design, different estimation strategies/methods should/must not be too influential.

In mode effect estimation, the following steps may be followed:

1. Identify the main quality and cost criteria
    - What benchmark is chosen for measurement?
    - Is it sufficient to consider accuracy (i.e. MSE) or also comparability in time and/or between subgroups?
    - What is the time horizon for which the mode design and budget are fixed and mode effects are estimated?
    - What are the key variables/population parameters of interest?
2. Decide whether mode effect estimation serves explanation only, design choice or adjustment
3. Identify available auxiliary data that is informative about
    - Mode selection
    - Mode measurement
4. Evaluate anticipated validity of assumptions for mode selection, mode measurement and absence of experimental influences
5. Decide whether an experimental design (such as re-interview or parallel run) is required and feasible to serve the purposes of the mode effect estimation;
6. Conduct experimental designs if deemed feasible and necessary

Based on these assumptions, in the following of this section general guidance when selecting methods to deal with the mode effects are provided.

We make two side remarks. First, we note that mode effects do not refer to biases only, but may also affect precision. Modes may affect, for example, motivation and concentration. Less motivated or concentrated respondents may give less reliable, i.e. more noisy, answers, leading to a loss of precision. Furthermore, interviewer effects have been a widely studied source of potential variation in survey statistics. In this report, we focus on bias adjustment. Second, we note that mode-specific measurement differences must be prevented above all through careful questionnaire design and testing. We, thus, assume that mode effect estimation is conducted to estimate remaining differences that are hard to detect and prevent in questionnaire design.

3.2. Assessing mode effects

In practice, biases and mode effects can be estimated according to two main approaches:

1) record check approach when the true scores are available from an external source;

2) measurement benchmark mode approach that requires the choice of a reference mode to produce best answers for a question.

The first approach is rarely feasible in practice but it allows estimating all biases and effects. The second approach assumes the benchmark measurement equals the true scores. The methods referred to here mainly follow this last approach.

The following schemes outline the methods which can be applied for different objectives of the study in different survey/experimental contexts, given the following types of analyses to be carried out:

- *Analysis of total mode effect* (Table 8), and
- *Analysis to disentangle measurement and selection effects* (Table 9).

The *objectives of the study* considered are:

1. assessing differences between estimates obtained based on data collected through different survey designs (single-mode and mixed-mode), in order to evaluate the total mode effect and the measurement equivalence;
2. analyzing the response processes and evaluation of the bias caused by the total nonresponse (selection errors);
3. assessing mode effect - disentangling measurement and selection effects.

Table 8. Analyses of total mode effect

Objective of study: Assessing differences between estimates obtained based on data collected through different survey designs (single-mode and mixed-mode), in order to evaluate the total mode effect and the measurement equivalence		
Method	**Analysis**	***Context* / Conditions**
Regression modelling approach to test whether design has a significant effect on the mean or distribution of the item (Martin and Lynn, 2011)	Univariate analysis of items to evaluate the impact on marginal distributions of mixed-mode design	– *Parallel independent surveys* Appropriate statistical models and tests
Tests on differences in the estimates (Martin and Lynn, 2011)	Univariate analysis to highlight significant differences in the estimates calculated on the two sample designs	– *Parallel independent surveys* Appropriate statistic tests for independent samples
Tests on indicators of *completeness* (item nonresponse) Tests on indicators of *accuracy* (comparisons with external data) (Jackle *et al*., 2010)	Analysis on differences in the quality indicators	– *Parallel independent surveys* Appropriate statistic tests
Multi-group confirmatory factor analysis (Martin and Lynn, 2011; Hox *et al*., 2015)	Analysis of the measurement equivalence when concepts are measured through more than one variable	– *Parallel independent surveys* – *Mixed mode survey design* Identification of the latent structure of the phenomenon, Control of selection effect
The proportional odds modelling technique (or parallel regression model, grouped continuous model) (Jackle *et al*., 2010)	Analysis to assess measurement equivalence of ordinal data on comparable samples	– *Parallel independent surveys* – *Mixed mode survey designs* Control of selection effect Validity of model assumption about covariates (covariates "shift" the distribution of responses proportionately across all categories)
Regression modelling approach whit one or more predictor variables and a binary indicator of single-mode and mixed-mode respondents (Martin and Lynn, 2011)	Multivariate analysis on estimates of the association between variables	– *Parallel independent surveys* Appropriate statistical models and tests on significant interaction effects
Objective of study: Analysing the response processes and evaluation of the bias caused by the total nonresponse (selection errors)		
Method	**Analysis**	**Context / Conditions**
Tests on the response rates respect to some characteristics of sample units (Jackle *et al*., 2010)	Analysis on the response rates	– *Parallel independent surveys* – *Single and mixed mode designs* Appropriate statistic tests for independent samples
Summary statistic tests	Analysis of deviations from mode independence (absolute and relative selection error per benchmark variable)	– *Parallel independent surveys* – *Comparison between single mode and mixed mode designs* Appropriate statistic tests
R-indicator, Conditional and Unconditional partial *R*-indicator (Klausch *et al*., 2015; Schouten *et al*., 2011; Shlomo and Schouten, 2013; Schouten, *et al*., 2017)	Analysis of the representative response (absolute selection error for sets of benchmark variables)	– *Parallel independent surveys* – *Single and mixed mode designs* MAR assumption for Response model
Tests on the differences between benchmark variables (true value) and estimates (Roberts and Vandenplas 2017)	Analysis on benchmark variables known for selected sample units	– *Parallel independent surveys* – *Single and mixed mode designs* Appropriate statistic tests

Table 9. Analyses to disentangle measurement and selection effects

Objective of study: Assessing mode effect - disentangling measurement and selection effects			
Method	**Analysis**	**Conditions**	***Context***
Weighting • Propensity score (PS) • Calibration • Post-stratification (Vandenplas *et al*., 2016; Rosenbaum and Rubin, 1983; Vannieuwenhuyze, *et al*., 2014)	Analysis based on response model to control for respondent characteristics (comparable samples in MM)	MAR assumption Mode-insensitive auxiliary variables Balancing assumption in PS	– *Mixed mode survey designs (observational studies)*
Regression model (Kolenikov and Kennedy, 2014)	Model analysis to estimate measurement and selection errors	Mode-insensitive auxiliary variables in the model to control selection effect	– *Mixed mode survey designs (observational studies)*
Other methods *-double robust estimation* that combines an outcome regression with a propensity score model - *matching*	Model to estimate causal effect	Appropriate statistical models	– *Mixed mode survey designs (observational studies)*
Instrumental variable approach (Vannieuwenhuyze *et al*., 2010)	Analysis based on benchmark single-mode design	Validity of comparability and representativity assumptions	– *Parallel independent surveys*
Re-interview (Biemer, 2001)	Analysis based on re-interview data, administrative data and paradata. The response of each mode is calibrated to the combined response of the re-interview and follow-up. Measurement effect (ME) is estimated as remaining difference between modes. Selection effect (SE) is estimated using mix of re-interview data, administrative data and paradata.	Re-interview does not affect measurement behavior of respondent. Nonresponse to re-interview is unrelated to survey variables of interest given administrative data and paradata.	– *Re-interview of subset of mixed-mode respondents (experimental design with sequential mixed mode survey)*

3.3. Adjusting for mode effects

The following Table 10 outlines the methods which are applicable for adjusting for mode effect in experimental contexts (re-interview, parallel single mode), or when auxiliary data from either administrative data or paradata are available, or in the case of longitudinal or repeated over time surveys.

Table 11 presents, for the standard covariate-based adjustment approach, a set of methods that can be used to correct selection and/or measurement effects.

Table 10. Approaches to adjust for mode effects

Objective of study: Adjustment methods			
Method	**Data requirements**	**Assumptions**	**Advantages/Disadvantages**
Standard Covariate-based adjustment	• Sampling frame data • Paradata • Survey responses	Missing at random potential outcomes (MAR) Exogeneity of auxiliary data	Too strong assumptions in many settings (-) Adjustment on individual level possible (+)
Time-series stabilization/ mode calibration (Buelens and van den Brakel, 2015, 2017)	Repeated cross-sectional / longitudinal survey	Independence of measurement and selection error Time-stability of measurement error (ME)	Does not decompose (-) Avoids ME estimation problem (+) Strong assumption on mode contributions (not fluctuate) (-)
Instrumental variable method (Vannieuwenhuyze *et al*., 2010)	Single-mode reference survey parallel to mixed-mode	Single-mode and mixed-mode survey have same selection bias (SB)	Avoids MAR and exogeneity assumption (+) Representativeness assumption usually implausible (-) Not available for >2 modes
Re-interview method (Klausch *et al*., 2017)	Re-interview of subset of mixed-mode respondents	Measurement equivalence	More plausible MAR assumption (+) MNAR estimators available (+) Measurement equivalence traded off against true score time-stability (-)

Table 11 . Methods to adjust for mode effect

Objective of study: Adjusting selection/measurement effects in MM (observational studies)		
Method	**Aim**	**Conditions**
Weighting - Propensity score - Calibration - Post-stratification (Vandenplas *et al*., 2016; Rosenbaum and Rubin, 1983; Austin, 2011; Vannieuwenhuyze, *et al*., 2014)	To equate samples To correct selection effect	Ignorability of selection mechanism (MAR) Mode-insensitive auxiliary variables Measurement error negligible
Regression (Kolenikov and Kennedy, 2014)	To estimate measurement and selection effects To correct measurement error	Appropriate statistical models
Other methods -*double robust estimation* that combines an outcome regression with a propensity score model - *matching*	To estimate causal effect To correct measurement error	Appropriate statistical models
Multiple imputation		
1.Multiple (standard) imputation	To predict counterfactual data (potential outcomes) To correct measurement error	Choice of benchmark mode MAR assumption
2.Multiple imputation with response and selection models proposed by Suzer-Gurtekin *et al*. (2012)		Choice of benchmark mode Sequential design and two modes (Possibility – non-ignorability of selection mechanism)
3.Fractional multiple imputation proposed by Park *et al*. (2016)		Sequential design and more than two modes Possibility – non-ignorability of selection mechanism

3.4. Final Discussion

It has to be reminded that mode effects are not necessarily bad. Mode effects, when present, can either improve or worsen the quality of survey estimates. An obvious improvement that results from mode-specific selection is a less selective sample of respondents in a mixed-mode survey compared to a single-mode survey. In this case a selection effect may be present, which manifests itself as a difference in survey estimates. Researchers can study the representativity and may come to the conclusion that the mixed-mode survey is to be preferred, and that the mode effect introduced is an improvement compared to the former survey design, the single-mode survey. Generally, mode dependent selection effects indicate a difference in representativity of the response collected through a mixed-mode design and a benchmark design. If the difference is such that the mixed-mode response is less selective, the selection effect corresponds to an improvement in survey estimates.

Measurement effects in mixed-mode designs are generally not desirable. Such effects typically arise when different modes have different associated biasing effects: they do not measure the target quantity at the same level, or with the same precision. Since mixed-mode designs produce responses using a combination of modes, the individual responses may become incomparable, as they are not all measured using the same measurement instrument (data collection mode in this setting).

Assessing mode effects

Both the ESS country experiences reported in the MIMOD survey and the literature review on methods for mode effect assessment and adjustment show that **activities and published literature on mode effect assessment are more widespread than on mode effect adjustment techniques**.

Mode assessment analyses are sometimes **limited to quantifying the total mode effect**. An important reason is that it **is difficult to separate selection from measurement effects**, but easy to assess their combined effect.

The main difficulty is the **confounding of selection and measurement effects**.

Sometimes assessment of mode effects may be sufficient, but **when detected, some effects may need to be corrected for, in particular measurement effects**. Methods to disentangle measurement and selection mode effects are needed.

Assessments (as well as adjustments) **are most sensibly conducted in a comparative manner**, by comparing a mixed-mode design with a single mode design, or with another multimode design.

In assessment studies, the **representativity of the response**, the **response rate**, and **distributional socio-demographic characteristics of the respondents** can be studied to gain insight into the selection mechanism of a mixed-mode design.

Generally it is of course desirable that the response collected through a mixed-mode design is better in some way: less selective and/or higher than for example through a single-mode design. **In this sense, selection effects are desirable and could reduce selection bias of survey estimates**. Adjustments for selection effects in mixed-mode designs are no different from adjustments in single-mode designs, and are generally needed because of selective coverage and nonresponse.

Adjusting for mode effects

Appropriate adjustment methods require the separation of selection and measurement effects in order to correct each, potentially by different types of approaches.

Adjustment methods in the context of mixed-mode designs are aimed at correcting survey estimates for **undesired mode effects**, **typically bias resulting from measurement effects**. Measurement effects arise

when respondents give different answers to the same questions in different modes. As a result, comparability of population subgroups who responded through different data collection modes may be compromised.

Assessment of measurement effects may show that there are systematic differences between measurements obtained through one mode compared to a different mode. **When applying adjustments, the researcher must choose a reference design as the benchmark**, since true measurement errors with reference to some unknown underlying construct are impossible to recover. The benchmark design can consist of a single data collection mode, or of a mix of several modes where the proportion of each mode in the mix is fixed at a specific level.

Measurements that deviate from the benchmark design are said to suffer from measurement effects and are in need of adjustments to remove the bias with respect to the benchmark.

Adjustments can be applied by using **different approaches**:

- **Weighting approaches** seek to correct through applying adjustments to the usual survey weights.
- In some situations one could use an **imputation approach** where counterfactuals are imputed: it consists in the application of prediction methods that attempt to predict at the item level measurements that would have been obtained had the data been collected through a different mode.
- Alternatively, systematic measurement differences between two modes could be **estimated at aggregated levels**, and subsequently used in an additional correction, for example.

Experimental vs observational studies

Measurement and selection mode effects are confounded in mixed-mode designs. **The two effects can be separated in experimental studies.**

Experimental studies are rather rare because of the **associated costs**.

However, **assessment and adjustment strategies are most reliable and hinge less on assumptions when conducted in experimental settings**. In such cases selection and measurement effects can be separated, which is important specifically in adjustment approaches. Separation of selection from measurement effects generally proceeds by explaining the selection **using some covariates** (which are **assumed to be mode-insensitive**), and attributing remaining differences to measurement. Hence, when separating the effects is not completely successful, selection effects are not fully explained, and as a result estimated measurement effects are biased.

Since separating selection from measurement effects are a prerequisite for successful mode effect assessments and adjustments in mixed-mode designs, a promising line of future research is the development of mixed-mode designs that allow for this, for example through **embedded experiments**. An example of such a design consists of conducting re-interviews through a second mode for a subset of respondents who already responded through a first mode.

4. Concluding remarks

In last decades, the use of combined data collection modes in social surveys has undergone a considerable increase due to the need of increase the quality of the data while limiting costs. Actually, the use of mixed-mode strategies produces advantages in terms of increasing response rates and coverage of target populations, and reduction of surveys costs and respondents burden.

However, mixed-mode data collection originates the so-called *mode effects* (selection and measurement effects), which may highly affect estimates accuracy. Furthermore, selection effect and measurement differences across modes are usually confounded. Mode effects need to be properly taken in to account at both the survey design and the estimation phases in order to reduce their biasing effects on target parameters and to ensure accurate estimates.

Within WP2 of the MIMOD project, methods and approaches to deal with mode effects in mixed-mode surveys are investigated. The activities have been firstly focused on an update of recent literature on the topic and on the results of the query carried out in the ESS during the project, in order to collect information about the recent methodological development and research in this field.

Hence, experimental applications of selected methods for mode effects assessment and adjustments have been carried out, using data from current social surveys at the Italian National Statistical Institute and at Statistics Netherlands.

Based on evidences and outcomes of the above activities, a general discussion and high-level guidelines and advices about possible methodological approaches and solutions which can be adopted and to deal with mode-effects in mixed-mode designs have been provided in this report.

However, given that only two countries have been involved in WP2, the results of the performed analyses allowed to delineate quite general guidelines on possible risks and advantages of combining modes of data collection in this context.

For the same reason, the WP2 results could not cover all the specific situations and application context of each country in the ESS.

However, the results of WP2 provide all countries in the ESS with an updated overview about methodological solutions to improve the quality of estimates produced in mixed-mode social surveys. The general guidance and discussions reported in this deliverable represent a good starting point for all countries who plan to design their own methodological strategies to assess and possibly adjust for mode effects in surveys using mixed-mode data collection.

Further research and analyses are necessary in this area at National and European level.

At European level, it is recommended that suitable modes of collaboration could be identified in the future to proceed with developments in this area, e.g. through a network of countries interested in continuing the discussion on methodological issues.

In particular, even if standardization is difficult in this context due to the complexity of the methodological elements involved in the design of strategies to deal with mode bias/mode effects, the general guidelines and advises provided in this report can be considered a first step to proceed towards the development of generalized tools supporting ESS countries in the methodological design of their own strategies.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46:399-424, 2011. Taylor & Francis Group, LLC.

Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. Journal of Official Statistics 17:295-320.

Buelens B., van den Brakel J. A., Schouten B. (2018a). Current methodologies to deal with mode effects and mode bias in multi-mode designs. MIMOD – MIxed MOde Designs for social surveys. Work Package 2: Mode bias/mode effects and adjustment for mode-effects. Deliverable 1: A report containing an overview on current methodologies adopted at the ESS NSIs to deal with mode bias/mode effects in multi-mode designs.

Buelens B., Klausch T., van den Brakel J. A., Schouten, B. (2018b). A cost-benefit analysis of re-interview designs for mode-specific measurement bias. MIMOD – MIxed MOde Designs for social surveys. Work Package 2: Mode bias/mode effects and adjustment for mode-effects. Deliverable 2: Report containing the results of the analyses performed on re-interview designs.

Buelens B. and van den Brakel J. A. (2015). Measurement error calibration in mixed-mode, Sociological methods & Research, 44(3), pp. 391-426.

Buelens, B. and van den Brakel J. A. (2017). "Comparing two inferential approaches to handling measurement error in mixed-mode surveys," Journal of Official Statistics, 33(2), pp. 513-531.

De Vitiis C., Inglese F., Guandalini A., Terribili M. D., Varriale R. (2018). Experimenting methods to asses and adjust mode effect on a social survey: a case study on the Italian "Aspects of daily life" survey. MIMOD – MIxed MOde Designs for social surveys. Work Package 2: Mode bias/mode effects and adjustment for mode effects. Deliverable 3: Report containing the results of the applications of selected methods on mixed-mode social surveys.

de Leeuw, E. (2005), To mix or not to Mix Data Collection Modes in Surveys, Journal of Official Statistics, 21(2), pp. 233-55.

Dillman D. A. and Christian L. M. (2005). Survey mode as a source of instability in responses across surveys. Field methods, 17(1), pp. 30-52.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. Public Opinion Quarterly, 69(3), pp.370-392.

Murgia. M., Gravem D.F., Lo Conte M. (2018). MIxed MOde Designs for social surveys. Work Package 1: Investigation of mode organisation (concurrent vs consecutive multi-mode data collection). Deliverable 1: Report on MIMOD survey on the state of the art of mixed mode for EU social surveys.

Jackle A., Roberts C. and P. Lynn. 2010. Assessing the Effect of Data Collection Mode on Measurement. International Statistical Review (2010), 78, 1, pp. 3–20.

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

Klausch T., J. Hox and B. Schouten. 2015. Selection error in single and mixed mode surveys of the Dutch general population. Journal of the Royal Statistical Society Series A 178: 945–961.

Klausch, T., Schouten, B., Buelens, B., and Van Den Brakel, J. (2018). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. Journal of Survey Statistics and Methodology, 5(4), pp. 409-432.

Kolenicov, S. and C. Kennedy. (2014). Evaluating three approaches to statistically adjust for mode effects. Journal of Survey Statistics and Methodology 2, pp. 126–158.

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, November 2011.

Park S., Kim J. K. and Park S. (2016), An imputation approach for handling mixed-mode surveys, The annals of Applied Statistics. Vol. 10, No. 2, pp. 1063-1085.

Roberts C. and C. Vandenplas. 2017. Estimating Components of Mean Squared Error to Evaluate the Benefits of Mixing Data Collection Modes.

Rosenbaum P. R. and D. B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, 70 (1), pp. 41-55.

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

Schouten, B., Peytchev, A., Wagner, J. (2017). Adaptive survey design. Chapman and Hall/CRC.

Shlomo N. and B. Schouten. 2013. Theoretical Properties of Partial Indicators for Representative Response.

Suzer-Gurtekin, Z. T., Heeringa, S., & Vaillant, R. (2012). Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys. Proceedings of the JSM, Section on Survey Research Methods, 4711-2.

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

Vannieuwenhuyze, J. T.A., G. Loosveldt and G. Molenberghs. 2014. Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models. *Journal of Official Statistics*, Vol. 30, No. 1, 2014, pp. 1–21, http://dx.doi.org/10.2478/jos-2014-0001.

Voogt, R. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. Journal of official statistics, 21(3), 367.