**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 analyses performed on re-interview designs

## *A cost-benefit analysis of re-interview designs for mode-specific measurement bias*

WP2 – Deliverable 2

Date: December 20, 2018

Barry Schouten (CBS)
Thomas Klausch (CBS/VUmc)
Bart Buelens (CBS)
Jan van den Brakel (CBS)

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

*Summary: WP2 of MIMOD concentrates on the estimation, detection and adjustment of mode-specific biases in survey statistics. Three intermediate deliverables are produced: a literature review and current status of detection and adjustment methodology, the results of the analyses performed on re-interview designs based on two CBS case studies and an application of a subset of these methods to an ISTAT case study. This deliverable 2 addresses the cost-benefit analysis of re-interviews. A final Methodological Report is also foreseen.*

*Re-interview designs are a potential tool to estimate and adjust for mode-specific measurement bias. In 2011, a re-interview design was successfully applied to the Dutch Safety Monitor, which led to a redesign of the survey. Re-interview designs may, however, be very costly, especially when face-to-face is included as a survey mode. The crucial question is whether benefits outweigh costs, i.e. whether the potential increase in accuracy of survey statistics is worth the investment. The answer to this question depends heavily on the purpose of the re-interview, i.e. assessment versus adjustment, the size of the mode-specific measurement biases, and the relative costs of the modes. Re-interview designs also make a number of assumptions that will not hold for every setting.*

*In this deliverable, we perform a cost-benefit analysis for two surveys, the Dutch Health survey and the Dutch Labour Force survey, and discuss the utility and validity of re-interviews. We conclude that for the Labour Force survey a re-interview may not be useful due to relatively small measurement differences, while for the Health survey it may be useful.*

## 1 - Introduction

With the emergence of the online survey mode, many national survey institutes transitioned their surveys to mixed-mode designs in which Web is combined with traditional survey modes. Mixed-mode designs are not new and have been explored for decades. However, given the low cost, relatively short return times but low response rates of the online survey mode, mixed-mode designs are now becoming rule rather than exception.

The online survey mode is mostly implemented as a self-administered mode and, as a consequence, is relatively disparate to interviewer-administered telephone and face-to-face survey modes, e.g. Dillman et al (2014). This disparity implies an increased risk of incomparability in time or between relevant population subgroups due to mode-specific measurement bias. While such a risk may have been reason not to combine modes in a single design in the past, the low cost of Web often simply overrules such considerations; the risk of incomparable statistics may be ignored or taken for granted. This risk is further alleviated by the growing range of devices on which the Web can be accessed and the gradual future change in the shares of the modes to the total response that is likely to come with it.

Survey methodology offers three options to overcome the risk of method effects when 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, see deliverable 1 of WP2. In this paper, we focus on the last two options, although estimates of measurement bias may inform questionnaire redesigns. We believe that advanced mixed-mode questionnaire design is the most important and effective option to reduce method effects, but not capable of removing all mode-specific measurement biases. We assume that questionnaires have been tested and evaluated extensively and consider settings where additional efforts are needed to decrease the risk of incomparability of survey statistics.

The estimation and evaluation of mode-specific measurement bias is inherently hard due to the confounding with mode-specific selection bias. Deliverable 1 of WP2 provides an overview of methods to detect mode-specific biases and an overview of methods to adjust such biases. Deliverable 3 of WP2 applies various methods to the Italian Aspects of Daily Life Survey.

Even with sophisticated designs, it may be hard to separate the two types of bias (Klausch 2014). However, without such designs, it is, in general, impossible to assess to what extent measurement biases arise due to method effects. In this paper, we consider mixed-mode re-interview designs, see Biemer (2001) and Schouten et al (2013), in which a sample of respondents to the regular survey is invited to participate in one of the other modes that is employed. More specifically, we restrict attention to sequential mixed-mode designs, where some of the modes are offered only to nonrespondents in the other modes. A re-interview has been successfully applied to the Dutch Crime Victimization Survey (CVS), see Schouten et al (2013), and estimation methodology has been evaluated and optimized by Klausch et al (2017). However, re-interview designs can be very costly and unbiased estimation of mode-specific measurement biases can only be made under a number of assumptions that may be implausible in some survey settings. In this paper, we, therefore, investigate the conditions under which re-interview designs may be sensible tools.

The utility of re-interview designs depends on the purpose. Re-interviews may be used to explain mode differences in measurement and to inform data collection design through the choice of survey modes. Re-interviews may also be used to adjust survey statistics. The adjustment setting is much more demanding as imprecision and bias in estimates for the mode-specific measurement bias directly translate to resulting survey statistics. Under the design setting, estimates merely guide decisions. For this reason, we evaluate the two settings separately in the following.

Our main research question is: When do the benefits of a re-interview outweigh the costs of the re-interview? We answer this question by considering two realistic case studies. Both concern household surveys in which the online survey mode has been introduced in a sequential design next to telephone and face-to-face interviewing: the Dutch Health Survey and the Dutch Labor Force Survey. We perform a cost-benefit analysis in which we compare bias and precision with and without re-interview. We consider both the Web mode and the interviewer mode as benchmark for measurement. We conclude that for the Health Survey re-interviews may be useful, while for the Labor Force Survey they are not.

The outline of the deliverable is as follows: In section 2, we motivate and explain the use of re-interview designs. In section 3, we present the methodology to optimize re-interview designs and to estimate mode-specific measurement biases. In section 4, we apply the methodology to the two surveys. Finally, in section 5, we discuss results and future study.

## 2 - The use and utility of re-interview designs

Re-interview designs may only be useful in certain circumstances and under certain conditions. Before we explain how cost – benefit analyses may be done, we give a motivating example linked to the Dutch Health survey.

### 2.1 – A motivating example

We consider the Dutch Health survey which has a sequential design with Web and face-to-face (F2F) as survey modes; an invitation to respond online is sent to a general population random sample and after a month nonrespondents are assigned to a F2F follow up. Roughly half of the respondents come from Web and half from F2F.

We consider 12 population strata based on age, health problems and medicine use {young, middle age, elderly} × {health problems, no health problems} × {medicine use, no medicine use }. Age of the sample unit is known from the sampling frame but health problems and medication are not. Medication is considered less important and is asked towards the end of the questionnaire.

*Relative selection and measurement biases*: Suppose that younger and older persons, persons with health problems and persons who use certain medication exhibit lower response rates in Web. The F2F follow-up adjusts this in part and stratum response rates are more similar after F2F respondents are added. Hence, the F2F follow-up has a beneficial selection bias and we would prefer the response to the combined modes over Web only. Suppose that the selection bias on health problems for Web only relative to the Web → F2F response is 5%. If we employ a weighting adjustment based on age, then this selection bias decreases to 3% due to the collinearity between the two variables. Suppose for simplicity, that for medication the unadjusted and adjusted biases are the same.
Suppose, next, that Web respondents are more honest than F2F respondents, because health is a sensitive topic, but that F2F interviewers are able to keep respondents more concentrated to the end of the questionnaire, especially for younger persons. Due to social desirable answering, the percentage health problems in F2F is lower. Due to insufficient interpretation and recall effort, the percentage medication in Web is lower. Suppose that for both variables there is a net 8% relative measurement bias between the two modes, but downwards for health problems and upwards for medication. When the modes are combined in the Web → F2F design, then the relative measurement bias reduces to 4%, due to the 50%-50% distribution of response over the two modes. The relative measurement bias is not noticeably affected by a weighting adjustment on age.

*Benchmark designs*: We can distinguish two benchmark designs (Klausch, Schouten, & Hox, 2017): A) a Web → F2F design where answers are given as in Web, and B) a Web → F2F design where answers are given as in F2F. Both are virtual designs with unobserved potential outcomes; the real designs are Web and Web → F2F with answers in the mode in which a person responds. The Web only design has a selection bias of 3% against both benchmarks and, additionally, a measurement bias of 4% against benchmark B. The Web → F2F has no selection bias but has a measurement bias that depends on the choice of benchmark.
When social desirability is deemed to be the major concern, then Web → F2F gives a relative measurement bias. Alternatively, when satisficing behaviour is deemed most risky, then Web has a relative measurement bias. The choice what behaviour is more influential amounts to a choice of measurement benchmark. That choice may be different for each survey variable. We suppose it is Web for health problems and F2F for medication.
Consider first the variable health problems where Web is the measurement benchmark (benchmark A). The measurement bias in F2F is 8% and negative; fewer health problems are reported. The gain in selection bias of 3% when F2F is added is offset by a loss of 4% in measurement bias, which leads to a net total bias increase of 1%. The Web only design has a bias of 3%, whereas the Web → F2F has a bias of 4%. The F2F seems to have little added value as the percentage health problems changes by only 1%.
Consider next the variable medicine use where F2F is the measurement benchmark (benchmark B). The measurement bias in Web is 8% and again negative, because respondents fail to report medicine use. When F2F is added, the gain in selection bias is 3% and the Web measurement bias of 8% gets attenuated to 4%. This leads to a net total bias change of 7%, and, now, F2F, seems to have a clear added value. The Web only design has a bias of 11%, whereas the Web → F2F has a bias of 4%.

*Re-interview*: A re-interview implies that (a subsample of) Web respondents are assigned to F2F, next to the F2F follow-up of Web nonrespondents, see Klausch (2014). They get the same questionnaire with some modifications, an alternative invitation letter is sent and interviewers are informed and receive additional training. The modifications in the questionnaire may consist of removing and/or replacing some less important modules or questions by new modules or question. These new modules/questions help justify a new interview on the one hand and may assist in explaining measurement differences on the other hand. In section 3, we will explain the estimation strategy based on the re-interview response.
The relative measurement and selection biases can be estimated by a re-interview under two main assumptions: 1) The F2F re-interview answers are not affected by the preceding Web participation, 2) re-

interview measurement behaviour is not a cause of missing data in the re-interview, and 3) the true value of the survey outcome variable shows negligible real change between the two measurements. The assumptions can be made more plausible by careful design of the timing and invitation. However, for some settings and surveys, the assumptions are unlikely to hold, even with careful design. Health problems and medication are relatively stable statistics, so that answers to a timely re-interview, say after a month, should not have changed much. The first assumption is, however, harder to verify and to safe-guard through design. Re-interview respondents may indeed still show social desirable answering as long as no reference is made to the answers from the Web interview and the impression is avoided that re-interview answers are evaluated against the Web responses. In other words, the respondents perceive the re-interview as a new request for data. When it comes to satisficing behaviour, there is an apparent risk that respondents are less motivated because they now have already done a similar survey. For this reason, the F2F re-interview needs to be framed and announced slightly differently as the original survey and interviewers need to take extra care to keep respondents motivated. Hence, it is clear the re-interview data collection demands a subtle change of design.

*Design and/or adjustment*: Let us suppose that in this example a re-interview is effective in separating selection from measurement bias on both survey variables. Based on a pilot, it is found that the percentage of health problems hardly changes when adding F2F response, whereas there is a strong increase in medicine use. Designers of the Health survey wonder if and how selection and measurement biases are confounded. There are two possible purposes for the re-interview: 1) To decide whether F2F is applied at all in future design, and 2) to determine relative measurement bias for health problems and medication in order to adjust future waves. Since the interest lies in changes in health and in associations between health survey variables, and not in absolute values, comparability in time and between age groups is deemed more important than accuracy. For this reason, the assessment purpose holds. Since age group comparability is important, the biases need to be disentangled per age group
The precision of the estimated size of the measurement bias depends on the size of the re-interview (sub)sample and the size of the bias itself. The measurement bias of 8% is obviously unknown but is set at a conservative estimate of 10%. The Health survey is a repeated cross-sectional survey with approximately 800 respondents per month, i.e. 400 respondents in both modes. If all Web respondents receive a re-interview, then the standard error of the measurement bias is 1.5% for one month of re-interview and 0.9% for three months of re-interview. It is decided to perform three months of re-interview and to decide per age group whether F2F follow-up is applied.

### 2.2 - Cost – benefit analysis

The example in the previous subsection points at a number of parameters and decisions that determine the utility and purpose of a re-interview. In general, a re-interview may be beneficial when the relative measurement bias is large, when either accuracy of survey statistics or comparability of survey statistics between population subgroups is important, when survey statistics are relatively stable in time, and when the impact of the regular interview on the re-interview is small. Furthermore, the choice of measurement benchmark moderates the utility of the re-interview.

The re-interview design follows the mixed-mode survey design. For two modes, say $m_1$ and $m_2$, there are only five options: single mode $m_1$, single mode $m_2$, a concurrent design $m_1 + m_2$, and two sequential designs $m_1 \rightarrow m_2$ and $m_2 \rightarrow m_1$. In this paper, we assume that a sequential design $m_1 \rightarrow m_2$, with the cheapest mode $m_1$ first, is considered against a single mode $m_1$. The re-interview then implies that $m_1$ respondents are again invited to participate in $m_2$. Re-interview designs for survey designs with three modes are sketched in Klausch et al (2017).

The costs of a re-interview are a function of the re-interview sample size and the fixed and variable costs per re-interview unit. In the Health survey example, the re-interview is conducted F2F and is, therefore, relatively costly. In the Crime Victimization Survey application, described in Schouten et al (2013), it was found that re-interview response rates in F2F were higher than nonresponse follow-up response rates in F2F. This implies that the re-interview unit costs were slightly higher than for a nonresponse follow-up. As a consequence, optimization of the re-interview design is attractive. The re-interview sample size depends on the required statistical power, which in turn depends on the purpose of the re-interview, i.e. estimates of bias to inform design decisions or an adjustment.

We speculate that, in the majority of settings, design decisions require less precise estimates of relative measurement bias than adjustments. This is because the precision of estimated measurement bias is inherited by the adjusted survey statistics, whereas design decisions impact mostly the bias of survey statistics. This becomes even more apparent when it is assumed that relative measurement bias may change over time, so that re-interviews may have to be repeated.

The design and adjustment scenarios are natural under different circumstances. The design scenario is more natural when a survey focusses on time change and on associations between survey and auxiliary variables, rather than on absolute levels of survey statistics. In such a setting, comparability is favoured to accuracy. The adjustment scenario comes from a focus on accurate statistics, i.e. the levels are of direct importance. In the cost-benefit analysis, we will consider both the assessment and adjustment scenarios.

## 3 - Analysis strategy

In this section, we optimize the design of re-interview samples and present a strategy for the cost-benefit analysis. We, first, discuss four scenarios for the analysis. Next, we move to the estimators in the analysis. In the final two subsections, we discuss the bias and variance of the estimators.

### 3.1 – Scenarios in the cost-benefit analysis

We evaluate the benefit of a re-interview design in terms of relative bias to the benchmark design, precision of the estimators and the overarching mean square error (MSE) relative to the benchmark design. The costs are evaluated in terms of the budget needed to conduct the survey for a specified period of time. In the cost assessment, we consider only variable costs, and assume they are scale-independent; the costs for one sample unit are, thus, independent of the sample size.

We select the optimal multi-mode design, possibly with a re-interview, under four scenarios:

1) minimize MSE with respect to the benchmark design, while assuming that the relative measurement bias is stable in time, under the constraint that the budget is equal to the sequential mixed-mode design without re-interview;
2) minimize MSE with respect to the benchmark design, while assuming that the relative measurement bias may change in time, under the constraint that the budget is equal to the sequential mixed-mode design without re-interview;
3) minimize bias with respect to the benchmark design while assuming that the measurement bias is stable in time, under the constraint that the precision equals that of the regular sequential mixed-mode design;
4) minimize bias with respect to the benchmark design while assuming that the measurement bias may change in time, under the constraint that the precision equals that of the regular sequential mixed-mode design;

The time-independence of the relative measurement bias is an influential assumption; when the bias does not change (or only very gradually) in time, then an estimate in a particular period can be re-used and forwarded to future data collection periods. That means that a re-interview becomes an investment that may be funded from future savings. Under time-dependence, we assume that the re-interview needs to be repeated in each new wave of the survey.

Scenarios 3 and 4 are different from scenarios 1 and 2 in the requirement that the precision is not affected. This is a constraint that may demand a larger budget, especially under time-dependence of the measurement bias. It means that the volatility of time series of the survey statistics must remain the same. This is a strong requirement implying that relative measurement biases need to be estimated with high precision and, consequently, demand for more budget.

As discussed in section 2.2, the re-interview may serve two purposes: inform survey design by estimating the size of measurement bias and adjust estimates for relative measurement bias. Under the design-option, only scenario 1 applies, as the re-interview is conducted once and future survey statistics are based on future data collection only. Since the design-option follows from a focus on comparability, it is not sensible to assume scenario 2; under this scenario, the design may be subject to constant change. All four scenarios may apply to the adjustment-option; the exact scenario that is evaluated depends on the importance of statistical power when evaluating temporal change of survey statistics, and a hypothesized change of relative measurement bias in time.

### 3.2 – Estimators

Consider a sequential mixed-mode design with two modes, $m_1$ and $m_2$, in which a sample of $m_1$ respondents is re-interviewed with $m_2$ and a sample of $m_1$ nonrespondents receives a follow-up in $m_2$. Assume that both samples are simple random samples without replacement from $m_1$ respondents and $m_1$ nonrespondents respectively, but with different subsampling probabilities. Let $\pi_1$ denote the (constant) inclusion probability for the re-interview and $\pi_2$ for the follow-up. The subsampling designs may, of course, be extended to stratified simple random sampling, depending on differences in the natural variation in population strata on key survey variables. Such an extension would be relatively straightforward but cumbersome in notation and optimization.

We consider three estimators: 1) the unadjusted response mean of the single mode $m_1$ design, 2) the unadjusted response mean of the sequential design $m_1 \rightarrow m_2$, and 3) an adjustment estimator using re-interview data in the sequential design $m_1 \rightarrow m_2$.

Klausch, Schouten, Buelens and Van den Brakel (2017) compare the statistical properties of a range of estimators that adjust for relative measurement bias with respect to the two possible measurement benchmarks ($BM = m_1$ and $BM = m_2$). The estimators include a fixed-effect estimator, a regression estimator, an inverse regression estimator, an inverse propensity weighting estimator and multiple imputation. Their comparison shows overlap with earlier analyses by West and Little (2013) in a slightly different context. In the Klausch et al (2017) comparison no attempt is made to optimize efficiency; 60% of the $m_1$ nonrespondents received a follow-up and 50% of the $m_1$ respondents are assigned a re-interview. The estimators based on a full follow-up ad re-interview are compared to the response means of the single mode $m_1$ design and the sequential $m_1 \rightarrow m_2$ design that are not adjusted for measurement bias. Based on simulations for various choices of parameters in nonresponse and measurement error models, they conclude that overall the inverse regression estimator[1] is the most accurate, i.e. has the smallest mean square error[2].

[1] The inverse regression estimator is maximum likelihood under normality for a pattern mixture model and that it is referred to as classical calibration in the measurement error calibration literature

Only when it is known that mode-specific measurement error implies merely a shift of the location of survey variables and not a rescaling, it is that the inverse regression estimator is outperformed by the fixed-effect estimator. Since we assume general measurement error models, we apply the inverse regression estimator.

Figure 1 shows the missing data pattern of a two mode sequential mixed-mode design with re-interview. For measurement benchmark $BM = m_2$, the inverse regression estimator regresses the answers to $m_1$ in area A on the answers to $m_2$ in area C, predicts the answers to $m_2$ in area D by inverse regression and combines the predicted answers with the answers in areas C and E. For measurement benchmark $BM = m_1$, the inverse regression estimator regresses the answers to $m_2$ in area C on the answers to $m_1$ in area A, predicts the answers to $m_1$ in area B by inverse regression on area E and combines the predicted answers with the answers in area A. Nonresponse to the follow-up is considered in a separate overall adjustment and ignored in this paper.

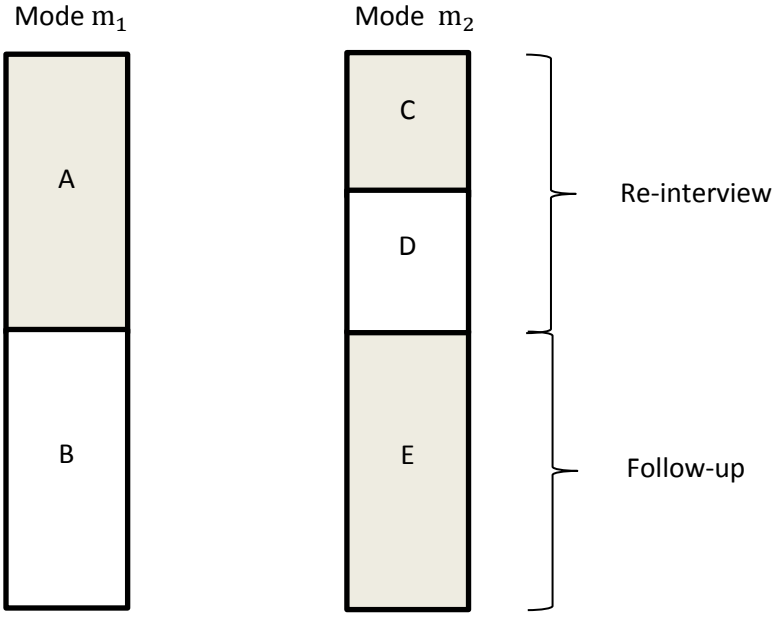


*Figure 1: Re-interview design for a $m_1 \rightarrow m_2$ sequential survey design. Grey areas represent $m_1$ response (A), re-interview $m_2$ response (C) and follow-up $m_2$ response (E). White areas represent $m_1$ nonresponse (B) and re-interview $m_2$ nonresponse (D). Nonresponse to both modes is omitted.*

We label the three estimators as $\hat{Y}_{m_1}$, $\hat{Y}_{m_1 \rightarrow m_2}$ and $\hat{Y}_{INV}$. In appendix A, we present an extension of the estimators in Klausch et al (2017) to mixed-mode designs with three modes, as they are applied at Statistics Netherlands, for example.

### 3.3 - Intervals for mode-specific selection and measurement biases

In this section, we set up intervals for selection and measurement biases. The bias and variance of the unadjusted single mode $m_1$ design, of the unadjusted sequential $m_1 \rightarrow m_2$ design and of the sequential $m_1 \rightarrow m_2$ design adjusted (using the inverse regression estimator) depend on these two biases. In real survey settings, these biases are unknown in advance. Klausch et al (2017) show that the bias of the inverse regression estimator is robust to the sizes of the relative measurement and relative selection bias; it does not change when the two biases vary. This is not true for the two unadjusted response means $\hat{Y}_{m_1}$ and $\hat{Y}_{m_1 \rightarrow m_2}$. We, therefore, have to make informed pre-assessments of theses biases in order to analyze the potential utility of a re-interview. Such pre-assessments could be made by methods presented in deliverable 1 of WP2.

[2] The results do not hold for regression coefficient estimates, only for means

We assume that an estimate for the total relative bias between the single mode $m_1$ and the sequential design $m_1 \rightarrow m_2$ is available for variable $Y$, say $\Delta_y$. This bias is the sum of a relative selection bias and a relative measurement bias when adding the follow-up response in $m_2$. The two bias terms may have the same sign but may also have opposite signs.

We need some further notation first, following the areas in figure 1. Let, $p_A$ be the probability of a $m_1$ response, $p_C$ be the probability of a $m_2$ re-interview response, and $p_E$ be the probability of a $m_2$ follow-up response. The share of mode $m_1$ to the total response is

$$P_1 = \frac{p_A}{p_A + (1-p_A)p_E}. \quad (1)$$

Now, we have $\Delta_y = (1 - P_1)(SB_y + MB_y)$, where $SB_y$ is the relative selection bias and $MB_y$ the relative measurement bias.

In the bias pre-assessment, we make three steps: 1) Determine the likely direction of the selection bias, 2) construct an interval for the selection bias, and 3) derive the interval for the measurement bias. The first step is based on literature and on experience with nonresponse monitoring and analysis of auxiliary variables, i.e. variables for which the values are known for nonrespondents. For the percentage smokers, we anticipate that the selection bias is positive; smokers tend to have lower response rates to Web. We base this conjecture on biases in income, educational level and other variables that are related to smoking behaviour. In the second step, we derive an interval by bounding the absolute value of the selection bias from above by a constant times the standard deviation, $S(y)$, of the survey outcome variable

$$|SB_y| \leq SB_y^{Max}(\alpha) = \alpha S(y), \quad (2)$$

where $\alpha$ is some constant larger than zero. For dichotomous variables, (2) amounts to $Sb_y^{Max}(\alpha) = \alpha\sqrt{\mu_y(1-\mu_y)}$, where $\mu_y$ is the response mean of the sequential design $m_1$.

It is our experience, e.g. Bethlehem, Cobben and Schouten (2011), that selection biases of dichotomous variables seldom exceed 5%. We choose to set $\alpha = 0.1$, so that when $\mu_y = 50\%$, then $SB_y^{Max}(\alpha) = 5\%$. For $\alpha = 0.1$, $SB_y^{Max}(\alpha) = 2.2\%$, when $\mu_y = 5\%$ or $\mu_y = 95\%$. For the percentage smokers, we have $\mu_y = 20\%$, and, consequently, $SB_y^{Max}(\alpha) = 4\%$. The resulting interval for the relative selection bias is $SB_y \in [0\%, 4\%]$. The third step is simple and we take the complement of the total relative bias and the relative selection bias interval. For percentage smokers, the relative measurement bias interval becomes $MB_y \in [6\%, 10\%]$, i.e. we expect that in F2F respondents will more often say they smoke than in Web.

We are aware that the functional form and scaling of (2) is arbitrary and should have a more empirical basis. As a result, the construction of bias intervals also becomes somewhat arbitrary. We note, however, that we make the pre-assessments only to explore the potential benefit of a re-interview; the decision to perform a re-interview should be robust to small changes in anticipated selection biases. The latter is the subject of our exploration.

### 3.4 - Optimization of the precision of re-interview estimators

We discuss optimization under the four scenarios of section 3.1. In the optimization, we either minimize the Mean Square Error (MSE) of estimators under budget constraints or minimize bias of estimators under constraints on variance. The three estimators that we consider are the unadjusted single mode $m_1$ response mean, the unadjusted sequential mixed-mode response mean, and the adjusted sequential mixed-mode response mean based on a re-interview.

In deriving optimal re-interview designs, we make three simplifications. First, we ignore the differences in variance of survey variables between modes. Klausch et al (2017) show that the variance of the inverse regression estimator depends on the variance of the random measurement error in the mode that is not the measurement benchmark; the larger this variance, the lower the reliability of the mode and the higher the variance of the inverse regression estimator. We include this additional variance in the simulation study in section 4. However, in the optimization, we assume that the survey outcome variable variance, $S^2(y)$, is comparable for $m_1$ respondents, $m_2$ re-interview respondents and $m_2$ follow-up respondents, which may be a strong assumption. However, in practice we may not know the difference in reliability in advance and we expect it does not have a strong impact. Under the simplification of equal variation, we can focus completely on the number of respondents in $m_1$, re-interview and follow-up strata. Second, although in practice we employ design-weighted estimators, we restrict ourselves here to simple random samples with replacement. Third, we assume that mode $m_2$ dominates costs of the survey design, which is true when $m_1 = \text{Web, Mail}$ or Telephone, and $m_2 = \text{F2F}$. Let contact costs in $m_2$ be $c_1$ and interview costs in $m_2$ be $c_2$. So we ignore any variation between individual sample units. We do acknowledge the difference in costs between a re-interview and follow-up. The costs for a re-interview per unit are $c_{RE} = c_1 + p_C c_2$ and for a follow-up they are $c_{FU} = c_1 + p_E c_2$. Let $B$ be the total available budget and let $n$ be the sample size. Then it must hold that

$$np_A\pi_1 c_{RE} + n(1 - p_A)\pi_2 c_{FU} \leq B. \quad (3)$$

In (3), it is the ratio $B_U = \frac{B}{n}$ that matters in the cost constraint, i.e. the available budget per sample unit.

As we will show in the following, eight population parameters turn up in the MSE and variance expressions of the three estimators that we consider: the response rates $p_A$, $p_C$, $p_E$, the unit costs $c_1$, $c_2$, the relative selection bias $SB_y$, the relative measurement bias $MB_y$, and the correlation between repeated measurements $\rho_{1,2}$. Next to these parameters, there is the budget per sample unit $B_U$, which is a constraint. We treat the response rates and the unit costs parameters as fixed and given, i.e. they are not subject to sampling variation The relative biases are varied as described in the previous section, but they are estimated and do affect the sampling variation.

The decision variables in the optimization problems are the subsampling probabilities $\pi_1$ and $\pi_2$, for re-interview and follow-up, respectively, and the overall sample size, $n$. The sample size plays a role in scenarios 3 and 4 because of the variance constraint.

In order to optimize, we need expressions for the biases relative to the benchmark design and the variances. From these, we can compute estimates of the MSE. Klausch et al (2017) derive bias expressions for the three estimators, which we do not repeat here and will employ as well. In appendix B, we derive variance approximations for the three estimators.

In the following subsections, we discuss each of the four optimization scenarios introduced in section 3.1. We focus on the choice of subsampling probabilities which determine the variances, and, thus, also the MSE's, but not the biases.

*3.4.1 – Scenario 1: Time-independence and trade-off between MSE and budget*

Scenario 1 assumes time-independence and the estimated measurement bias between $m_1$ respondents and $m_2$ re-interview respondents is used in future waves.

Obviously, the future is not indefinitely long. Say $T$ future waves are anticipated to use the same design. The total budget for $m_2$ is $B = (T + 1)n(1 - p_A)c_{FU}$. The re-interview needs to be done in the first upcoming

wave. The sample size of this wave may be taken larger, say $n_0 = \beta_0 n$, and the sample sizes of the future waves are all equal, say $n_F = \beta_1 n$.

Given that the estimated measurement bias is re-used, any subsampling for the re-interview would be inefficient, as it is the only time it is conducted, i.e. $\pi_1 = 1$. Furthermore, it is optimal that the follow-up sample is the same in size over all waves and the precision coming of the $m_1$ response is also the same over all waves. These can be translated as

$$n_0(1 - p_A)\pi_2 = n_F(1 - p_A), \tag{4a}$$

$$n_0 p_A p_C = n_F p_A. \tag{4b}$$

The conditions (4a-b) lead to $\pi_2 = \frac{\beta_1}{\beta_0}$, and $\beta_0 = \frac{\beta_1}{p_C}$. Finally, $\beta_1$ can be derived from the total budget constraint

$$(T + 1)n(1 - p_A)c_{FU} = n_0 p_A c_{RE} + (T + 1)n_F(1 - p_A)c_{FU}. \tag{5}$$

Some manipulation gives the optimal solution

$$\beta_1 = \frac{(T+1)(1-p_A)c_{FU}p_C}{p_A c_{RE}+(T+1)(1-p_A)c_{FU}p_C}, \quad \beta_0 = \frac{(T+1)(1-p_A)c_{FU}}{p_A c_{RE}+(T+1)(1-p_A)c_{FU}p_C}, \pi_2 = p_C. \tag{6}$$

The optimal solution (6) can now be used to estimate the bias, variance and MSE of the estimators.

*3.4.2 – Scenario 2: Time-dependence and trade-off between MSE and budget*

Under scenario 2, mode-specific measurement biases are time-dependent and the re-interview is repeated for each wave. We, therefore, have to consider a single data collection wave and cannot exceed the budget of one wave. It requires that the optimal subsampling probabilities are chosen such that $B = n(1 - p_A)c_{FU}$.

With (3), we then get the constraint

$$np_A\pi_1 c_{RE} + n(1 - p_A)\pi_2 c_{FU} \leq n(1 - p_A)c_{FU}, \tag{7}$$

which can be rewritten as

$$p_A\pi_1 c_{RE} \leq (1 - p_A)(1 - \pi_2)c_{FU}. \tag{8}$$

In the optimization, we perform a brute force optimization and derive the MSE under all pairs of subsampling probabilities in $\{0,0.01,0.02, \dots ,1.00\} \times \{0,0.01,0.02, \dots ,1.00\}$ that satisfy (8).

*3.4.3 – Scenario 3: Time-independence and trade-off between bias and variance*

Scenario 3 replaces the MSE objective function by the bias objective function and the budget constraint by a variance constraint.

This scenario is handled by adding a benchmark-dependent precision constraint. In analogy to scenario 1, let the survey design be constant for $T$ waves, let $n_0 = \beta_0 n$, and the sample sizes of the future waves all be equal to $n_F = \beta_1 n$. Now, both $\beta_0$ and $\beta_1$ must larger than 1 as the re-interview requires budget and simultaneously decreases precision. The total required budget over all waves becomes

$$\beta_0 n(p_A\pi_1 c_{RE} + (1 - p_A)\pi_2 c_{FU}) + T\beta_1 n(1 - p_A)c_{FU}, \tag{9}$$

whereas the regular $m_1 \rightarrow m_2$ without re-interview would cost

$$(T+1)n(1-p_A)c_{FU}. \quad (10)$$

The $\pi_1$, $\pi_2$, $\beta_0$ and $\beta_1$ are chosen as follows: Depending on the benchmark, we, first, choose subsampling probabilities by minimizing the variance of the inverse regression estimator, $S^{BM}_{IREG}$, as given in the appendix. Next, fixing the optimal subsampling probabilities, we minimize (9) under two precision constraints, with $S_{m_1 \to m_2}$ being the variance of the unadjusted sequential design response means:

1) $S^{BM}_{IREG}$ based on a sample of size $n_0$ must be smaller than or equal to $S_{m_1 \to m_2}$
2) the variance of future response means of the $m_1 \to m_2$ design adjusted for the mode-specific measurement bias and based on a sample of size $n_F$ must be smaller than or equal to $S_{m_1 \to m_2}$.

The first constraint implies that the re-interview wave is sufficiently precise, whereas the second constraint implies that future waves are sufficiently precise.

The precision of the adjusted response mean of the $m_1 \to m_2$ design is approximately equal to

$$S^{ADJ,m_1}_{m_1 \to m_2} = \frac{1}{\beta_1} S_{m_1 \to m_2} + \frac{1}{\beta_0}(S^{BM}_{IREG} - S_{m_1}), \quad (11)$$

under $BM = m_1$, and under $BM = m_2$ it is

$$S^{ADJ,m_2}_{m_1 \to m_2} = \frac{1}{\beta_1} S_{m_1 \to m_2} + \frac{1}{\beta_0}(S^{BM}_{IREG} - S_{m_1 \to m_2}). \quad (12)$$

where the second terms in (11) and (12) come from the adjustment. The rationale behind the approximations in (11) and (12) is that the added variance is the same as the added variance by the inverse regression estimator.

The bias of the inverse regression estimator is compared to the single mode and sequential design response means in order to evaluate whether a re-interview is beneficial.

*3.4.4 - Scenario 4: Time-dependence and trade-off between bias and variance*

Scenario 4 is the most demanding as a re-interview needs to be repeated in each wave and precision needs to be kept constant. Without increasing budget, this requirement is not feasible. The precision constraint demands a larger sample size for each wave, say $\tilde{n} = \beta n$.

First, $S^{BM}_{IREG}$ is minimized by choosing optimal subsampling probabilities, but without a budget constraint. This is possible as the optimal subsampling probabilities are independent of the sample size. Next, $\beta$ is derived by constraining the precision of $S^{BM}_{IREG}$ based on a sample of size $\tilde{n}$. Unlike scenario 3, we have to consider only the precision of the current wave as re-interviews are repeated. It, therefore, suffices to let $S^{BM}_{IREG}$ based on a sample of size $\tilde{n}$ be equal to $S_{m_1 \to m_2}$. It follows easily that

$$\beta = \frac{S^{BM}_{IREG}}{S_{m_1 \to m_2}}. \quad (13)$$

The budget per wave can be computed for the increased sample and is equal to

$$\beta n(p_A \pi_1 c_{RE} + (1-p_A)\pi_2 c_{FU}). \quad (14)$$

As a last step, again, the bias of the inverse regression estimator is compared to those of the unadjusted response means.

**4 - Application to Dutch Health survey and Labour Force survey**

We optimize the re-interview design for two Dutch surveys that are conducted by Statistics Netherlands, the Health Survey (HS) and the Labour Force Survey (LFS). We start with some background to the surveys, including derivations of bias intervals for the relative selection and measurement bias, and then discuss optimization under the design and adjustment options.

### 4.1 - Data and survey designs

The HS and LFS are repeated surveys that employ monthly samples. The HS is purely cross-sectional, while the LFS is a rotating panel with five waves that have three month time lags. Table 1 contains some details of the surveys. For the HS, only annual statistics are made, while for the LFS monthly statistics are made. For this reason, the LFS is much larger than the HS. The LFS has three modes, Web, telephone (CATI) and F2F (CAPI). However, for the sake of illustration, we combine the response to the two interviewer modes. The estimation strategy in appendix A may be followed to separate the biases to all three modes.

*Table 1: Sample size, modes, share of each mode to total response, target population, target population size and publication frequency.*

| | *Sample size* | $m_1$ | $m_2$ | $P_1$ | *Population* | *Population size* | *Publication frequency* |
|---|---|---|---|---|---|---|---|
| LFS | 8.000 | Web | CATI+CAPI | 59% | 16-64 years | 11 000 000 | Month |
| HS | 850 | Web | CAPI | 52% | 12+ years | 14 000 000 | Year |

For both surveys, we consider a re-interview spread over three consecutive months. Table 2 contains the survey outcome variables for which a decomposition of relative selection and measurement bias is evaluated. The LFS has one key variable, the unemployment rate, while for the HS four statistics are chosen from a range of key statistics. Table 2 also shows the unadjusted response means for Web ($m_1$) and the interviewer modes ($m_2$). Most statistics show relatively small differences, except for HS statistics percentage smoker and percentage visit to dentist in last year.

Apart from the means for the survey variables, we need an assessment of the reliability, operationalized as the correlation between original measurements and re-interview measurements. This correlation determines the performance of the inverse regression estimator, see Klausch et al (2017); the larger the reliability, the more powerful the re-interview and the smaller the MSE of the inverse regression estimator. Table 2 contains the anticipated reliabilities of the survey variables, i.e. correlation between repeated measurements. The reliability depends on the time lag between the two measurement and the intrinsic volatility of the characteristic itself. We assume a re-interview time lag between one and two months. We deem unemployment to be relatively volatile, while we view health, smoking and obese as relatively stable. We set the reliability of dentist visits in between.

*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* $m_2$ | *Reliability* |
|---|---|---|---|---|
| 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 |

Table 3 displays the three steps of section 3.3 for the pre-assessment of selection and measurement biases per variable. The differences $\Delta_y$ are computed from table 2. The anticipated signs of the selection bias are based

on known biases in income, registered employment, age and different forms of government allowance. The selection bias interval is constructed by taking $\alpha = 0.1$ in (2). The measurement bias interval then follows directly. Only for percentage smoker and percentage visit to dentist, it follows that the anticipated measurement bias is large. The other three variables have intervals that contain zero.

*Table 3: Total relative bias, anticipated sign of relative selection bias, intervals for relative selection and measurement bias and share of relative measurement bias to total relative bias.*

| | | | *Interval* | |
|---|---|---|---|---|
| | $\Delta_y$ | *Sign* | $SB_y$ | $MB_y$ |
| Unemployment rate | 1.2% | + | (0.0%, 2.3%) | (-1.1%, 1.2%) |
| % good health | -2.4% | - | (-4.1%, 0.0%) | (-2.4%,1.7%) |
| % smoker | 9.9% | + | (0.0%, 4.0%) | (5.9%, 9.9%) |
| % obese | 1.8% | + | (0.0%, 3.3%) | (-1.5%, 1.8%) |
| % contact dentist | -7.8% | - | (-3.8%, 0.0%) | (-7.8%, -4.0%) |

*Table 4: Measurement bias levels in percentage of total relative bias.*

| | Left | Mid | Right |
|---|---|---|---|
| Unemployment rate | 1.2% | 0.05% | -1.1% |
| % good health | 1.7% | -0.35%% | -2.4% |
| % smoker | 9.9% | 7.9% | 5.9% |
| % obese | 1.8% | 1.65% | -1.5% |
| % contact dentist | -4.0% | -5.9% | -7.8% |

In the results, we will consider three relative measurement bias values: the two extreme values following from the selection bias interval and the midpoint of the interval. These are labelled the left, mid and right measurement bias levels, see table 4, taking the relative selection bias as viewpoint. The left extreme value is where the selection bias is smallest and the right extreme value is where the selection bias is largest. For example, the left point of the selection bias interval of 0.0% for the unemployment rate implies a 1.2% measurement bias and the right point of 2.3% means a measurement bias of –1.1%.

In order to facilitate interpretation, we estimate Root Mean Square Error (RMSE) values instead of MSE; these are on the same measurement level as the values of the survey target variable that we investigate.

**4.2 - A cost – benefit analysis from the design perspective**

Under the design perspective, no adjustment is performed but the re-interview is merely done to facilitate a decision between the single mode $m_1$ design and the sequential design. In the case studies, $m_1 =$ web and $m_2 =$ F2F/phone for the LFS and $m_2 =$ F2F for the HS. Under the design perspective, only scenario 1 is meaningful to explore: We minimize the MSE under the assumption that the relative measurement bias changes only very gradually in time and while fixing the total budget over a period of $T$ waves. Since we implement the re-interview for three months, $T$ implies multiples of three months. We consider three values, $T = 3, 7, 19$, i.e. we fix budget for a year, for two years and for five years. Since the re-interview is not repeated nor are future waves adjusted, we have to choose between a design without or with $m_2$ follow-up.

Table 5 contains the optimal subsampling probabilities and the two sample size scaling parameters $\beta_0$ and $\beta_1$ for the three choices of $T$ per survey, following from (6). The optimal values depend on the survey, but are independent of the survey outcome variable.

Table 6 has the RMSE values (in %) of the estimators, $\hat{Y}_{m_1}$ and $\hat{Y}_{m_1 \to m_2}$ for each of the survey variables for the two benchmark modes $BM = m_1$ and $BM = m_2$. The highlighted values give the preferred choice between the two options, i.e. single mode or sequential.

The results for the HS are not surprising: When the benchmark mode is web, then in the majority of cases, 10 out of 12, the single mode web design is preferred. When the benchmark mode is F2F, then it is the other way around; for 8 out of 12 the sequential design web → F2F has a smaller RMSE. In other words, the optimal design comes down to the choice of benchmark. However, there are a few exceptions, where the re-interview could help. These are the cases where the relative measurement bias has an opposite sign as the relative total bias, i.e. there is large measurement bias but an even larger selection bias of opposite sign. The re-interview design may be used to conclude that these settings do not hold for the HS and then the choice of design amounts to a choice of benchmark.

*Table 5: Optimal subsampling probabilities and sample size scale parameters for scenario 1 per time period and survey.*

| | *Health survey* | | | *Labor Force Survey* | | |
|---|---|---|---|---|---|---|
| | *T=3* | *T=7* | *T=19* | *T=3* | *T=7* | *T=19* |
| $\pi_1$ | 1 | 1 | 1 | 1 | 1 | 1 |
| $\pi_2$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |
| $\beta_0$ | 1.09 | 1.23 | 1.34 | 1.12 | 1.26 | 1.35 |
| $\beta_1$ | 0.76 | 0.86 | 0.94 | 0.79 | 0.88 | 0.95 |

*Table 6: 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 design 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* |
| $\hat{Y}_{m_1}$ | - | <0.1 | 0.4 | 0.9 | 0.1 | 1.0 | 2.0 | 0.2 | 0.1 | 1.9 | <0.1 | 0.8 | 1.6 | 0.1 | 1.0 | 1.8 |
| $\hat{Y}_{m_1 \to m_2}$ | - | 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* |
| $\hat{Y}_{m_1}$ | - | 1.1 | 0.5 | 0.2 | 2.4 | 1.3 | 0.2 | 9.9 | 8.9 | 7.8 | 1.8 | 1.0 | 0.1 | 7.8 | 6.8 | 5.8 |
| $\hat{Y}_{m_1 \to m_2}$ | - | 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 |

The results for the LFS are less obvious; under both benchmarks it is possible that each of the designs is to be preferred. Hence, for the LFS a re-interview is beneficial so that the right decision can be made. It must, however, be remarked that differences in RMSE between the designs are relatively small, so that one may accept a suboptimal choice.

### 4.3 - A cost – benefit analysis from the adjustment perspective

Under the adjustment option, we consider all scenarios. This option implies that we remove the relative measurement bias to adjust towards the benchmark design. Scenario 1 is explored in the preceding section and revisited under the adjustment perspective. We consider again $T = 3, 7, 19$ in scenarios 1 and 3. In section 4.2, we noted that from the design perspective, we can only choose between the single mode $m_1$ design and the sequential design $m_1 \rightarrow m_2$. From the adjustment perspective, there is a third option, namely the adjusted sequential design $m_1 \rightarrow m_2$, where the estimated relative measurement bias is removed using the re-interview.

*4.3.1 Scenario 1: Minimize MSE under stable measurement bias and budget constraints*

The adjustment of the relative measurement bias implies a gain in bias but also comes at a price: the variance will increase.

*Table 7: 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* |
| $\hat{Y}^{\text{adj}}_{m_1 \rightarrow m_2}$ | 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 |
| $\hat{Y}_{m_1}$ | - | <0.1 | 0.4 | 0.9 | 0.1 | 1.0 | 2.0 | 0.2 | 1.0 | 1.9 | <0.1 | 0.8 | 1.6 | 0.1 | 1.0 | 1.8 |
| $\hat{Y}_{m_1 \rightarrow m_2}$ | - | 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* |
| $\hat{Y}^{\text{adj}}_{m_1 \rightarrow m_2}$ | 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 |
| $\hat{Y}_{m_1}$ | - | 1.1 | 0.5 | 0.2 | 2.4 | 1.3 | 0.2 | 9.9 | 8.9 | 7.8 | 1.8 | 1.0 | 0.1 | 7.8 | 6.8 | 5.8 |
| $\hat{Y}_{m_1 \rightarrow m_2}$ | - | 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 |

Table 7 presents the RMSE values of the three estimators. The values for $\hat{Y}_{m_1}$ and $\hat{Y}_{m_1 \rightarrow m_2}$ are the same as in table 6. The highlighted values in table 7 point at the preferred design.

For the LFS, adjustment is not favourable in all but one case, and even for this case the gain is very small. Hence, for the LFS it is not sensible to use a re-interview to adjust from an RMSE point of view.

For the HS, the picture is quite different and in the majority of cases the RMSE values for the adjusted design are smaller, although the gain is sometimes very modest. Only when the benchmark mode is web and the relative selection biases are small (the levels labelled as "left"), it is not sensible to adjust.

Remarkably, the length of the time period in which the design is kept stable plays a relatively minor role; the RMSE values do get much smaller for longer time periods, but only very gradually.

*4.3.2 Scenario 2: Minimize MSE under time-dependent measurement bias and budget constraints*

Under scenario 2, each wave has a re-interview and the inverse regression estimator is used directly.

The optimal subsampling probabilities are given in table 8 for the two benchmarks and the two surveys. For the HS, there is subsampling for the re-interview and the follow-up. For the LFS, there is no subsampling of the re-interview.

*Table 8: Optimal subsampling probabilities for scenario 2 per benchmark and survey.*

| | *Health survey* | | *Labor Force Survey* | |
|---|---|---|---|---|
| | $BM = m_1$ | $BM = m_2$ | $BM = m_1$ | $BM = m_2$ |
| $\pi_1$ | 0.84 | 0.53 | 1.00 | 1.00 |
| $\pi_2$ | 0.51 | 0.69 | 0.53 | 0.53 |

*Table 9: RMSE values (in %) for the HS and LFS survey variables per relative measurement bias level. Highlighted values have lowest RMSE.*

*a) benchmark $BM = m_1$.*

| | *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* |
| $\hat{Y}_{INV}$ | 0.3 | 0.3 | 0.3 | 1.1 | 1.1 | 1.1 | 1.2 | 1.2 | 1.2 | 0.9 | 0.9 | 0.9 | 1.6 | 1.6 | 1.6 |
| $\hat{Y}_{m_1}$ | <0.1 | 0.4 | 0.9 | 0.1 | 1.0 | 2.0 | 0.1 | 1.0 | 1.9 | <0.1 | 0.8 | 1.6 | <0.1 | 0.9 | 1.8 |
| $\hat{Y}_{m_1 \to m_2}$ | 0.5 | 0.2 | 0.5 | 1.7 | 1.3 | 1.5 | 5.0 | 4.0 | 3.1 | 1.3 | 1.0 | 1.2 | 3.9 | 3.1 | 2.2 |

*b) benchmark $BM = m_2$.*

| | *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* |
| $\hat{Y}_{INV}$ | 0.3 | 0.3 | 0.3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.8 | 0.8 | 0.8 | 1.3 | 1.3 | 1.3 |
| $\hat{Y}_{m_1}$ | 1.1 | 0.5 | 0.1 | 2.4 | 1.3 | 0.3 | 9.9 | 8.9 | 7.8 | 1.8 | 1.0 | 0.1 | 7.8 | 6.8 | 5.8 |
| $\hat{Y}_{m_1 \to m_2}$ | 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 |

The RMSE' values of $\hat{Y}_{m_1}$, $\hat{Y}_{m_1 \to m_2}$ and $\hat{Y}_{INV}$ for each of the survey outcome variables are shown in table 9. The highlighted values in table 9 give the preferred design.

For $BM = m_1$, adjustment using a re-interview is only sensible for cases where both the relative measurement and selection bias are large. These are the cases labelled as "right". For all other settings, the single mode or the sequential design is superior in terms of RMSE. In other words, we end up in the design perspective again; the re-interview is conducted to make a design decision.

For $BM = m_2$, the picture is different: Adjustment is sensible for the HS in almost all cases. Here, the re-interview may, thus, be a way to decrease the RMSE. For the LFS, it is only sensible when the relative selection bias is small ("left") and, hence, the relative measurement bias is large.

*4.3.3 Scenario 3: Minimize bias under stable measurement bias and constraints on precision*

Scenario 3 seeks to minimize bias while maintaining precision and assuming a stable relative measurement bias.

*Table 10: Optimal subsampling probabilities, sample size scale parameters and relative increase in required budget for scenario 3 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* |
| $\pi_1$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\pi_2$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\beta_0$ | 2.7 | 3.7 | 5.2 | 1.5 | 2.2 | 3.2 | 4.1 | 5.1 | 7.2 | 3.1 | 3.9 | 5.5 |
| $\beta_1$ | 1.9 | 1.5 | 1.3 | 1.6 | 1.4 | 1.2 | 2.0 | 1.7 | 1.4 | 1.8 | 1.5 | 1.3 |
| $\Delta B$ | 148% | 106% | 67% | 80% | 61% | 41% | 199% | 139% | 85% | 146% | 105% | 66% |

*Table 11: Bias values (in %) for the HS and LFS survey variables per relative measurement bias level. Highlighted $\hat{Y}_{INV}$ values lead to a gain bias of more than 0.5%..*

*a) benchmark $BM = m_1$.*

| | *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* |
| $\hat{Y}_{INV}$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -0.1 | -0.1 | -0.1 | 0.1 | 0.1 | 0.1 |
| $\hat{Y}_{m_1}$ | 0.0 | 0.4 | 0.9 | 0.0 | 1.0 | 2.0 | 0.0 | 1.0 | 1.9 | 0.0 | 0.8 | 1.6 | 0.0 | 0.9 | 1.8 |
| $\hat{Y}_{m_1 \to m_2}$ | 0.5 | 0.0 | 0.4 | 1.1 | 1.6 | 0.8 | 4.8 | 3.8 | 2.8 | 0.9 | 0.1 | 0.7 | 3.7 | 2.8 | 1.9 |

*b) benchmark $BM = m_2$.*

| | *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* |
| $\hat{Y}_{INV}$ | 0.0 | 0.0 | 0.0 | -0.1 | -0.1 | -0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -0.1 | -0.1 | -0.1 |
| $\hat{Y}_{m_1}$ | 1.1 | 0.5 | 0.1 | 2.4 | 1.3 | 0.2 | 9.9 | 8.9 | 7.8 | 1.8 | 1.0 | 0.1 | 7.8 | 6.8 | 5.8 |
| $\hat{Y}_{m_1 \to m_2}$ | 0.7 | 0.0 | 0.6 | 1.2 | 0.2 | 0.9 | 5.1 | 4.1 | 3.1 | 0.9 | 0.1 | 0.8 | 4.1 | 3.1 | 2.1 |

The optimal subsampling probabilities are given in table 10 for the two benchmarks, the two surveys and three time periods. Table 10 also contains the sample size scale parameters $\beta_0$ and $\beta_1$ and the required increase in survey budget to do a re-interview at the fixed precision level. The increase in required budget to keep variances at regular levels is considerable. The smallest increase, 41%, is for the HS under the web benchmark mode for a five year period. The largest is 199% for the LFS under a time horizon of a year. In all cases, no subsampling is performed. The wave 1 sample is increased by a factor between 1.5 and 7.2. The future waves' sample sizes are increased by a factor 1.2 to 2.0. As expected, the longer the time horizon the larger is wave 1 and the smaller are future waves.

The bias of $\hat{Y}_{m_1}$, $\hat{Y}_{m_1 \to m_2}$ and $\hat{Y}_{INV}$ for each of the survey outcome variables is shown in table 11. Since the re-interview design removes the relative measurement bias, it is always preferable to do the re-interview, except in trivial cases where there is no measurement bias to adjust. Table 11 shows the remaining relative

bias towards the benchmark design, which must be weighed against the increase in budget as estimated in table 10.

Under $BM = m_1$, the inverse regression estimator outperforms the other estimators, but yield can be small. In many cases the gain in bias may not be worth the extra budget. There are a few exceptions in the settings where the relative selection bias is largest (labelled as "right").

Under $BM = m_2$, again the inverse regression estimator is best, but now the gains are much larger, especially for the HS. For this benchmark, the investment may be worthwhile for longer time horizons for the HS.

*4.3.4 Scenario 4: Minimize bias under time-dependent measurement bias and constraints on precision*

Finally, scenario 4 minimizes bias like scenario 3, but assumes time varying relative measurement bias. The re-interview is done in every wave of the survey. Table 12 displays the optimal subsampling probabilities for all settings, the sample size scaling parameter $\beta$ and the increase in budget in order to maintain the same precision level for the re-interview.

*Table 12: Optimal subsampling probabilities, sample size scale parameter and corresponding increase in costs for scenario 4 per benchmark and survey.*

| | *Health survey* | | *Labor Force Survey* | |
|---|---|---|---|---|
| | $BM = m_1$ | $BM = m_2$ | $BM = m_1$ | $BM = m_2$ |
| $\pi_1$ | 1 | 1 | 1 | 1 |
| $\pi_2$ | 1 | 1 | 1 | 1 |
| $\beta$ | 1.8 | 1.1 | 2.6 | 1.8 |
| $\Delta B$ | 181% | 66% | 284% | 157% |

From table 12, we can see that again, as expected, the increase in required budget is large to very large; it ranges from 66% up to 284%.

The relative biases in table 11 also hold for scenario 4 as they are independent of sample size. Hence, the most favourable setting is $BM = m_2$ for the HS. Bias reductions can be considerable but the budget must be 66% higher in order to maintain precision. This is a complicated trade-off that might only be positive when both accuracy and comparability are deemed very important.

*4.3.5 Summary of results*

All in all, we can conclude that a re-interview design can be attractive under certain objectives and benchmark design choices. When the measurement benchmark mode is the more expensive second mode, then re-interviews are often useful for the HS. This is because relative measurement biases are fairly large and survey outcome variables are expected to be time stable and have relatively high correlations. For the LFS, biases and correlations are smaller, making it less profitable. When the measurement benchmark mode is the cheaper first mode, then re-interviews are only profitable when relative selection biases are large. This is because the gain in selection bias from the expensive second mode is no longer offset by the relative measurement bias; this bias exists in the unadjusted sequential design but is identified by the re-interview. Scenarios 3 and 4, in which also the precision is constrained, implies large increases in required budget and for the HS and LFS will often not have a positive business case. When the objective is design, i.e. a choice between a single mode and a sequential design, then a re-interview may be worth the effort, especially for the HS.

## 5 - Discussion

Our main research question is: When do the benefits of a re-interview outweigh the costs of the re-interview? We researched this question from two perspectives, that of design choices and that of adjustment/estimation choices. Under the design perspective, a re-interview merely serves a choice between designs, in this paper, the design that has the highest accuracy, i.e. lowest mean square error, under a budget constraint. Under the adjustment perspective, a re-interview is conducted in order to remove relative measurement biases in the current survey wave and, possibly, also future waves.

The two case studies in the paper show that a re-interview can be profitable under both the design perspective and the adjustment perspective. This is especially true when relative measurement biases between modes are fairly large, when the more expensive mode is the measurement benchmark and when correlations between measurements are relatively large. In our examples, we conclude that a re-interview may be favourable for the Dutch Health Survey, but not for the Dutch Labour Force Survey.

The utility of a re-interview exists only under three main assumptions. One assumption is that the re-interview does not affect measurement behaviour in the alternative mode; here, the more expensive mode. A second assumption is that mode-specific measurement behaviour itself is not a cause of missing data in the re-interview. In other words, there is no association between mode selection and mode measurement. The last assumption is that the time change in answers is absent between the two measurements. The last two assumptions can be combined in stating that the re-interview measurement behaviour itself is not a cause of missing data in the re-interview. The three assumptions can to some extent be ascertained by a careful choice of timing of the re-interview and introduction and motivation of the re-interview. The time lag should be short enough to avoid time change but long enough to neutralize any experimental conditioning. Even so, the assumptions can be quite strong or even unrealistic for some surveys or specific survey questions. It is important to keep this in mind. We expect that for a Health Survey a repeated measurement can be conducted without strong experimental influence.

There are a number of simplifications and limitations in our study. The most important, perhaps, is that we translated the methodology of Klausch et al (2017), designed for continuous survey variables, to categorical variables. Although this may be acceptable for ordinal variables or binary variables, the methodology would need to be revised for nominal variables. For this reason, we restricted ourselves to binary survey variables. A second limitation is that we considered simple random sampling without replacement. In practice, sampling designs are usually more complex, which may impact some of the conclusions of this paper. A third limitation is that we simplified the variance approximations of the inverse regression estimator and ignored sampling variation in the regression coefficients. In general, we will underestimate the variance of the estimator and, consequently, overestimate its accuracy. In the analyses, we found, however, that biases dominate the mean square error. A final simplification is that we ignored costs for the first and cheaper mode. In web – F2F designs, such a simplification may be acceptable, but in paper – telephone designs cost differences are small and costs of both modes need to be included. For all these simplifications, however, one may state that our approach merely facilitates a decisions and does not pretend or require exact approximations.

Obviously, there are a number of directions to further explore and elaborate the methodology and findings of this paper. We already mentioned the extension of the methodology to all measurement levels, in particular categorical variables. Next, as we have shown, the utility of a re-interview depends on the choice of measurement benchmark. In this paper, we left this decision open, but, in practice, it needs to be made. Such a uniform choice may not be easy for the survey as a whole. Natural approaches are questionnaire profiles summarizing the characteristics of individual survey items and survey blocks.

It would be very useful if findings of this paper can be replicated for other surveys and/or countries/settings.

**References**


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.

Dillman, D. A., Smyth, J. D., Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, Hoboken, NJ: John Wiley.

Klausch, L.T. (2014), *Informed Design of Mixed-Mode Surveys. Evaluating mode effects on measurement and selection error*, Ph.D. dissertation, Utrecht University: Utrecht, The Netherlands.

Klausch, L.T., Schouten, B., Buelens, B., Van den Brakel, J. (2017), Adjusting measurement bias in sequential mixed-mode surveys using re-interview data, Journal of Survey Statistics and Methodology, 5 (4), 409 – 432.

Klausch, L.T. Schouten, B., Hox, J. (2017), Evaluating bias of sequential mixed-mode designs afgainst benchmark surveys, Sociological Methods and Research, 46 (3), 456 – 489.

Schouten, B., Brakel, J. van den, Buelens, B., Laan, J. van der, Klausch, L.T. (2013), Disentangling mode-specific selection and measurement bias in social surveys, Social Science Research, 42, 1555 – 1570.

Schouten, B., Peytchev, A., Wagner, J. (2017), *Adaptive Survey Design*, Statistics in the Social and Behavioral Sciences Series, Chapman & Hall/CRC.

West, B.T., Little, R.J.A. (2013), Non-response adjustment of survey estimates based on auxiliary variables subject to error, Journal of the Royal Statistical Society: Series C, 62 (2), 213 – 231.

**Appendix A – Adjustment for measurement bias in sequential designs with three modes**

Statistics Netherlands has a design with Web, telephone and face-to-face (F2F), Web → telephone + F2F as the default survey design. The whole sample is first invited to participate in a Web survey through an advance letter with login and possibly a QR-code. Nonrespondents with a registered phone number receive a follow-up by telephone and all other nonrespondents by F2F. The default design is essentially a mix of a sequential and concurrent design. The population is divided into five subpopulations: 1 = households with a registered phone responding to web, 2 = households without a registered phone responding to web, 3 = households with a registered phone responding to telephone after not responding to web, 4 = households without a registered phone responding to F2F after not responding to web, and 5 = households not responding to the MM design. Subpopulation 5 is out of scope and no outcomes are estimated for this subpopulation. We label the remaining subpopulations as $l = 1,2,3,4$.

We introduce some additional notation: Let $(l_1, l_2)$ be the union of subpopulations $l_1$ and $l_2$. Furthermore, let $\widehat{N}_l$ be the estimated size of subpopulation $l$ based on the sample, and $\widehat{N}$ be the estimated population size. Let $\hat{Y}^{s}_{BM,t}$ be the type $t \in \{\text{unadj}, \text{ratio}, \text{reg}, \text{ireg}, \text{fixed} - \text{effect}, \ldots\}$ estimator applied to section $s \in \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$ of the population assuming $BM \in \{\text{web}, \text{tel}, \text{F2F}\}$ as the benchmark mode for measurement. Let $\bar{Y}_l$ be the design-weighted mean of the outcomes for subpopulation $l$.

Under BM=web, estimators have the form

$$\hat{Y}_{web} = \frac{\widehat{N}_1+\widehat{N}_3}{\widehat{N}} \hat{Y}^{(1,3)}_{web,t_1} + \frac{\widehat{N}_2+\widehat{N}_4}{\widehat{N}} \hat{Y}^{(2,4)}_{web,t_2}, \qquad \text{(A.1)}$$

where the estimators are allowed to be different per section given conjectures about the form of the measurement model.

The estimators for BM=tel and BM=F2F are conceptually the same; we give only the estimator for BM=F2F. Under BM=F2F, estimators have the form

$$\hat{Y}_{F2F} = \frac{\widehat{N}_1+\widehat{N}_4}{\widehat{N}} \hat{Y}^{(1,4)}_{F2F,t_1} + \frac{\widehat{N}_2+\widehat{N}_4}{\widehat{N}} \hat{Y}^{(2,4)}_{F2F,t_2} + \frac{\widehat{N}_3+\widehat{N}_4}{\widehat{N}} \hat{Y}^{(3,4)}_{F2F,t_3} - 2\frac{\widehat{N}_4}{\widehat{N}} \bar{Y}_{F2F}, \qquad \text{(A.2)}$$

where the term $\frac{\widehat{N}_4}{\widehat{N}} \bar{Y}_{F2F}$ is included in each of the estimators and needs to be subtracted twice. Again, the estimators may be different per section.

The estimators (A.1) and (A.2) are written in such a form that they can be computed easily using any existing implemented code for a two-mode design.

**Appendix B – Variance approximations**

We provide variance approximations to the three selected estimators under a simple random sample without replacement design. More general designs can be dealt with in a relatively similar way, but, obviously, lead to different solutions. Following Klausch et al (2017), a fixed response model is assumed, all strata are taken as fixed. We, then, have three probability mechanisms: the sampling from the population, the re-interview subsampling and the follow-up subsampling.

We use the following notation: $\pi_1$ is the subsampling probability in the re-interview, $\pi_2$ is the subsampling probability in the follow-up, $n$ is the sample size, $N$ is the population size, $p_A$ is the mode $m_1$ response rate, $p_C$ is the re-interview response rate, and $p_E$ is the follow-up response rate. $P_1$ is the proportion of response in mode $m_1$, i.e. $P_1 = \frac{p_A}{p_A+(1-p_A)p_E}$, and $P_2$ is the proportion of response in mode $m_2$, i.e. $P_2 = \frac{(1-p_A)p_E}{p_A+(1-p_A)p_E}$. We assume that all response rates and the proportions $P_1$ and $P_2$ are fixed and non-random. For the expected sizes of the areas in figure 1, we have

- Expected size of $m_1$ response: $np_A$
- Expected size of $m_2$ re-interview sample: $np_A\pi_1$
- Expected size of $m_2$ follow-up sample: $n(1-p_A)\pi_2$
- Expected size of $m_2$ re-interview response: $np_A\pi_1p_C$
- Expected size of $m_2$ follow-up response: $n(1-p_A)\pi_2p_E$

We further simplify by assuming that the variances of a survey variable $Y$ under the two modes, $\sigma^2(Y^{(1)})$ and $\sigma^2(Y^{(2)})$, are the same, so that $\sigma^2\left(Y^{(1)}\right) = \sigma^2\left(Y^{(2)}\right) = \sigma^2$. Let $\rho_{1,2}$ denote the correlation between the two measurements.

The variance of the mode $m_1$ response mean is

$$S_{m_1} = \frac{1}{np_A}\sigma^2, \tag{B1}$$

and is independent of the subsampling probabilities.

The variance of the sequential design $m_1 \rightarrow m_2$ response mean with fixed mode proportions is

$$S_{m_1\rightarrow m_2} = P_1^2\frac{1}{np_A}\sigma^2 + P_2^2\frac{1}{n(1-p_A)\pi_2p_E}\sigma^2, \tag{B2}$$

which is dependent on the follow-up subsampling.

The variance approximation for the inverse regression estimator depends on the choice of benchmark and requires more work. The inverse regression estimator has the following form

$$BM = m_1\colon \hat{\bar{Y}}_{IREG}^{(1)} = P_1\bar{Y}_A^{(1)} + P_2(\bar{Y}_C^{(1)} - \frac{1}{\nu_1}\bar{Y}_C^{(2)} + \frac{1}{\nu_1}\bar{Y}_E^{(2)}) \tag{B3}$$

$$BM = m_2\colon \hat{\bar{Y}}_{IREG}^{(2)} = P_2\bar{Y}_E^{(2)} + P_1(\bar{Y}_C^{(2)} - \frac{1}{\nu_2}\bar{Y}_C^{(1)} + \frac{1}{\nu_2}\bar{Y}_A^{(1)}) \tag{B4}$$

In expressions (B3) and (B4), the subscript 1 or 2 indicates the mode of measurement and the superscript A, C or E the set of sample units as in figure 1. As regression coefficients lead to higher order effects, we treat the regression coefficients $\nu_1$ and $\nu_2$ as fixed in the following and ignore their contributions to sampling variation. The regression coefficients are equal to

$$\nu_1 = \frac{cov(Y^{(1)},Y^{(2)})}{\sigma^2} \tag{B5}$$

$$\nu_2 = \frac{cov(Y^{(1)},Y^{(2)})}{\sigma^2}. \tag{B6}$$

We have that $\nu_1 = \nu_2 = \rho_{1,2}$.

For the variances in (B3) and (B4), we have

$$var(\bar{Y}_A^{(1)}) = \frac{1}{np_A}\sigma^2 \tag{B7}$$

$$var(\bar{Y}_C^{(m)}) = \frac{1}{np_A\pi_1 p_C}\sigma^2 \tag{B8}$$

$$var(\bar{Y}_E^{(2)}) = \frac{1}{n(1-p_A)\pi_2 p_E}\sigma^2, \tag{B9}$$

and for the covariances

$$cov(\bar{Y}_A^{(1)}, \bar{Y}_C^{(1)}) = \frac{1}{np_A}\sigma^2 \tag{B10}$$

$$cov(\bar{Y}_A^{(1)}, \bar{Y}_C^{(2)}) = \frac{\rho_{1,2}}{np_A}\sigma^2 \tag{B11}$$

$$cov(\bar{Y}_A^{(1)}, \bar{Y}_E^{(2)}) = 0 \tag{B12}$$

$$cov(\bar{Y}_E^{(2)}, \bar{Y}_C^{(2)}) = 0 \tag{B13}$$

$$cov(\bar{Y}_E^{(2)}, \bar{Y}_C^{(1)}) = 0 \tag{B14}$$

$$cov(\bar{Y}_C^{(1)}, \bar{Y}_C^{(2)}) = \frac{\rho_{1,2}}{np_A\pi_1 p_C}\sigma^2 \tag{B15}$$

Now, combining (5) to (13), we get for $BM = m_1$

$$S_{IREG}^{m_1} = P_1^2\frac{\sigma^2}{np_A} + P_2^2\left(\frac{\sigma^2}{np_A\pi_1 p_C}\left(1+\frac{1}{\nu^2}\right) + \frac{1}{\nu^2}\frac{\sigma^2}{n(1-p_A)\pi_2 p_E} - \frac{\rho_{1,2}}{\nu}\frac{\sigma^2}{np_A\pi_1 p_C}\right) + 2P_1P_2\frac{\sigma^2}{np_A}\left(1-\frac{\rho_{1,2}}{\nu}\right). \tag{B16}$$

(B16) can be simplified to

$$S_{IREG}^{m_1} = \frac{\sigma^2}{np_A}\left(P_1^2 + P_2^2\frac{1}{\rho_{1,2}^2}\left(\frac{1}{\pi_1 p_C} + \frac{p_A}{(1-p_A)\pi_2 p_E}\right)\right). \tag{B17}$$

Similarly, for $BM = m_2$, we get

$$S_{IREG}^{m_2} = P_2^2\frac{\sigma^2}{n(1-p_A)\pi_2 p_E} + P_1^2\left(\frac{\sigma^2}{np_A\pi_1 p_C}\left(1+\frac{1}{\rho_{1,2}^2}\right) + \frac{1}{\rho_{1,2}^2}\frac{\sigma^2}{np_A} - \frac{2\sigma^2}{np_A\pi_1 p_C} + \frac{2\sigma^2}{np_A} - \frac{1}{\rho_{1,2}^2}\frac{2\sigma^2}{np_A}\right), \tag{B18}$$

which simplifies to

$$S_{IREG}^{m_2} = \frac{\sigma^2}{np_A}\left(P_2^2\frac{p_A}{(1-p_A)\pi_2 p_E} + P_1^2\left(2 - \frac{1}{\pi_1 p_C} + \frac{1}{\rho_{1,2}^2}\left(\frac{1}{\pi_1 p_C} - 1\right)\right)\right). \tag{B19}$$

We let $S_{IREG}^{BM}$ denote the variance of the inverse regression estimator for the benchmark mode BM.