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

***A case study for the Dutch Health survey and Labour Force survey***

Barry Schouten, Thomas Klausch, Bart Buelens & Jan van den Brakel

*Summary: 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 any setting. In this paper, we perform a cost-benefit analysis for two case studies.*

**1 - Introduction**

With the emergence of the online survey mode, many survey institutes transitioned their surveys to multi-mode designs in which Web is combined with traditional survey modes. Multi-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, multi-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. 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, avoided through data collection design, and adjusted through estimation design. In this paper, we focus on the last two options, although estimates of measurement bias may inform questionnaire redesigns.

The estimation and evaluation of mode-specific measurement bias is inherently hard due to the confounding with mode-specific selection bias. Even with sophisticated designs, it may be hard to separate the two types of bias. 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 re-interview designs 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 (2016). 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 this article.

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. 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 modes 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.

**2 - A motivating example**

Re-interview designs may only be useful in certain circumstances and under certain conditions. We give a motivating example linked to the Dutch Health survey.

The Dutch Health survey 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 effect 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, 2015): 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. For ease of exposition, 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 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 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 [REFS]. They get the same questionnaire with some modifications, an alternative invitation letter is sent and interviewers are informed and receive additional training. 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, and 2) 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. Suppose a re-interview is effective in separating selection from measurement bias on both survey variables.

*Design and/or adjustment*: 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 effects 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.

**3 - Analysis strategy**

**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 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 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 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.

As discussed in section 2.2, the re-interview may serve two purposes: inform survey design and adjust 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.

**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 $π\_{1}$ denote the (constant) inclusion probability for the re-interview and $π\_{2}$ for the follow-up.

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 (2015) 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 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 is the most accurate, We label the three estimators as $\hat{Y}\_{m\_{1}}$, $\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$ and $\hat{Y}\_{INV}$.

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

Obviously, the bias of the three estimators is unknown in real survey settings. Klausch et al (2015) 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.

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 $∆\_{y}$. 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. 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.

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

We discuss optimization under the four scenarios of section 3.1. In deriving optimal re-interview designs, we make one simplification. Klausch et al (2015) 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. 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. We know this is not true. 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.

Schouten, Klausch, Buelens and Van den Brakel (2018) derive optimal subsampling probabilities under the four scenarios. Here, we restrict ourselves to results.

**4 - Application to Dutch Health survey (HS) and Labour Force survey (LFS)**

**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.

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.

*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 |

*Table 2: Selected survey outcome variables with estimates per mode.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *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 |
| % obesitas | HS 2014 | 12.1% | 13.9% | 0.9 |
| % visit to dentist | HS 2014 | 82.3% | 74.5% | 0.7 |

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 (2016); the larger the reliability, the more powerful the re-interview and the smaller the RMSE of the inverse regression estimator. In section 3, for simplicity, we have ignored reliability to derive optimal subsampling probabilities. However, in simulating HS and LFS survey variables and derive properties of estimators, we will include it. Table 3 contains assessments of the reliability. 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 obesitas as relatively stable. We set the reliability of dentist visits in between.

Table 3 displays the three steps of section 3.3 for the pre-assessment of selection and measurement biases per variable. The differences $∆\_{y}$ 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. 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.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | *Sign* | *Interval* | *% measurement bias* |
| $$∆\_{y}$$ | $$SE\_{y}$$ | $$ME\_{y}$$ |
| Unemployment rate | 1.2% | + | (0.0%, 2.3%) | (-1.1%, 1.2%) | (-92%,100%) |
| % good health | -2.4% | - | (-4.1%, 0.0%) | (-2.4%,1.7%) | (-100%, 71%) |
| % smoker | 9.9% | + | (0.0%, 4.0%) | (5.9%, 9.9%) | (60%, 100%) |
| % obesitas | 1.8% | + | (0.0%, 3.3%) | (-1.5%, 1.8%) | (-83%, 100%) |
| % contact dentist | -7.8% | - | (-3.8%, 0.0%) | (-7.8%, -4.0%) | (-100%, -51%) |

In the results, we will consider three relative measurement bias values: the two extreme values in table 3 and the midpoint of the interval. For example, for unemployment rate we consider measurement biases of -92%, 4% and 100% of the total relative bias.

**4.2 - A cost – benefit analysis**

We consider $T=3, 7, 19$ in scenarios 1 and 3, i.e. a one year, two year and five year time horizon.

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

Table 4 presents the RMSE values of the unadjusted response means of the single mode and sequential mode designs and of the adjusted sequential mode design.

*Table 4: 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* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}^{adj}$$ | 3 | 0.5 | 0.5 | 0.5 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.7 | 0.7 | 0.7 | 0.8 | 0.8 | 0.8 |
| 7 | 0.5 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.7 | 0.7 | 0.7 | 0.8 | 0.8 | 0.8 |
| 19 | 0.5 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.8 | 0.8 | 0.8 |
| $$\hat{Y}\_{m\_{1}}$$ | - | <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 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | - | 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* $BM=m\_{2}$*.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *T* | *LFS* | *HS* |
| *Unemployment**ME bias level* | *Health**ME bias level* | *Smoking**ME bias level* | *Obesitas**ME bias level* | *Dentist**ME bias level* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}^{adj}$$ | 3 | 0.6 | 0.6 | 0.6 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.7 | 0.7 | 0.7 | 0.9 | 0.9 | 0.9 |
| 7 | 0.6 | 0.6 | 0.6 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 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 | 0.8 | 0.8 | 0.8 | 0.7 | 0.7 | 0.7 | 0.8 | 0.8 | 0.8 |
| $$\hat{Y}\_{m\_{1}}$$ | - | 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 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | - | 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 |

The highlighted values in table 4 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 relative selection bias is zero and $BM=m\_{1}$, it is not sensible to adjust.

*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 5 for the two benchmarks and the two surveys. The RMSE values of $\hat{Y}\_{m\_{1}}$, $\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$ and $\hat{Y}\_{INV}$ for each of the survey outcome variables are shown in table 6.

*Table 5: Optimal subsampling probabilities for scenario 2 per benchmark and survey.*

|  |  |  |
| --- | --- | --- |
|  | *Health survey* | *Labor Force Survey* |
|
| $$BM=m\_{1}$$ | $$BM=m\_{2}$$ | $$BM=m\_{1}$$ | $$BM=m\_{2}$$ |
| $$π\_{1}$$ | 0.74 | 0.84 | 0.68 | 0.99 |
| $$π\_{2}$$ | 0.56 | 0.51 | 0.68 | 0.54 |

*Table 6: 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* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{INV}$$ | 0.3 | 0.3 | 0.3 | 1.0 | 1.0 | 1.0 | 1.1 | 1.1 | 1.1 | 0.8 | 0.8 | 0.8 | 1.3 | 1.3 | 1.3 |
| $$\hat{Y}\_{m\_{1}}$$ | <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 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | 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* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{INV}$$ | 0.3 | 0.3 | 0.3 | 1.0 | 1.0 | 1.0 | 1.1 | 1.1 | 1.1 | 0.8 | 0.8 | 0.8 | 1.1 | 1.1 | 1.1 |
| $$\hat{Y}\_{m\_{1}}$$ | 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 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | 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 |

The highlighted values in table 6 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. For $BM=m\_{1}$, adjustment is sensible for the HS in almost all cases.

*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. The optimal subsampling probabilities are given in table 7 for the two benchmarks, the two surveys and three time periods. Table 7 also contains the required increase in survey budget to do a re-interview at the fixed precision level. The bias of $\hat{Y}\_{m\_{1}}$, $\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$ and $\hat{Y}\_{INV}$ for each of the survey outcome variables is shown in table 8.

*Table 7: 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* |
| $$π\_{1}$$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $$π\_{2}$$ | 0.76 | 0.76 | 0.76 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $$∆B$$ | 76% | 59% | 41% | 99% | 74% | 51% | 58% | 46% | 32% | 102% | 76% | 55% |

*Table 8: 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* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{INV}$$ | 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 | 0.4 | 0.9 | 0 | 1.0 | 2.0 | 0 | 1.0 | 1.9 | 0 | 0.8 | 1.6 | 0 | 0.9 | 1.8 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | 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* |
| *1.0* | *.04* | *-.92* | *1.0* | *.14* | *-.73* | *1.0* | *.80* | *.60* | *1.0* | *.09* | *-.81* | *1.0* | *.76* | *.51* |
| $$\hat{Y}\_{INV}$$ | 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 |
| $$\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$$ | 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 |

Table 7 tells us that the increase in required budget to keep variances at regular levels is considerable. The smallest increase, 32%, is for the LFS under $BM=m\_{1}$ for a five year period.

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 8 shows the increase in bias, which must be weighed against the increase in budget in table 7.

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

Finally, scenario 4 minimizes bias as scenario 3, but assumes time varying relative measurement bias. The re-interview is done in every wave of the survey. Table 9 displays the optimal subsampling probabilities for all settings and the increase in budget in order to maintain the same precision level for the re-interview. Table 8 has the biases of $\hat{Y}\_{m\_{1}}$, $\hat{Y}\_{m\_{1}\rightarrow m\_{2}}$ and $\hat{Y}\_{INV}$ for each of the survey outcome variables; the remaining biases do not depend on the frequency of the re-interview design.

*Table 9: Optimal subsampling probabilities, sample size scale parameter and corresponding increase in costs for scenario 4 per benchmark and survey.*

|  |  |  |
| --- | --- | --- |
|  | *Health survey* | *Labor Force Survey* |
|
| $$BM=m\_{1}$$ | $$BM=m\_{2}$$ | $$BM=m\_{1}$$ | $$BM=m\_{2}$$ |
| $$π\_{1}$$ | 1 | 1 | 1 | 1 |
| $$π\_{2}$$ | 0.76 | 1 | 1 | 1 |
| $$ΔB$$ | 55% | 93% | 46% | 83% |

From table 9, we can see that again, as expected, the increase in required budget is large; it ranges from 46% up to 93%. When comparing table 9 to table 7, it appears that in some cases a continuous re-interview and adjustment is preferable to a lower frequency re-interview and adjustment under our optimization strategy.

**5 - Discussion**

Adjustment of mode-specific measurement bias may be beneficial when such bias is anticipated to be large, when survey statistics are relatively stable over time (i.e. have a moderate to strong association within respondents) and when re-interview designs do not affect respondent behavior very strongly. Some questions for discussion:

* How can we make (rough) pre-assessments of measurement biases?
* What is more natural: a design perspective or an adjustment perspective?
* What are practical time horizons for which design can be fixed and investments can be made?
* Would re-interview designs be feasible within your institutions?