Thomas Klausch
Department of Epidemiology and Biostatistics
VU University Medical Center, Amsterdam
t.klausch@vumc.nl

Column PhD thesis

A look into the challenges of mixed-mode surveys

Thomas Klausch received the Willem R. van Zwet Award 2014 for his thesis *Informed Design of Mixed-Mode Surveys*. The Willem R. van Zwet Award is the annual prize of the Netherlands Society of Statistics and Operations Research for an excellent PhD thesis in the area of statistics or operations research, in 2014 awarded to two people. In this article Thomas Klausch introduces us to his research.

Policy makers in governments, businesses, and NGOs need information on many aspects of our society and economy to be able to take decisions effectively. National statistical institutes, like Statistics Netherlands (*Centraal Bureau voor de Statistiek*), collect data and produce official statistics for these actors. Concern for the quality of the published estimates is high and research into improving quality is actively followed for this reason.

Mixing modes: the new face of survey research

In my PhD thesis at Utrecht University, I studied methods for evaluating the quality of data collected in a new type of survey design, the so-called 'mixed-mode' survey. Traditionally, a survey uses one way of communication with persons in a sample (the 'mode'), in particular asking questions in person (face-to-face), on the phone, or on paper questionnaires. In the past two decades, traditional surveys have come under increased pressure. For one, the number of persons willing to participate in surveys and thus provide personal data has steadily decreased. Furthermore, available budgets for data collection have shrunk, which made it difficult to still use costly modes in interviewing, especially face-to-face or telephone. In addition, the internet has made available a new and particularly cost-efficient way to collect survey data. Contrary to all traditional modes, administering web questionnaires involves only very small additional costs per sample unit (e.g., for sending a letter with a hyperlink by mail to a home address). Despite this advantage, response to web surveys is, unfortunately, slim. Surveys at Statistics Netherlands, for example, can obtain data from 60 to 70 per cent of sampled persons when using face-to-face interviewing, but usually not more than 20 to 30 per cent when using the internet.

The idea of a mixed-mode survey is getting the best of both worlds: saving on costs while increasing response. Figure 1 shows a so-called 'sequential' design as it was fielded yearly from 2008 until 2013 by Statistics Netherlands for the Dutch Crime Victimization Survey (CVS). This procedure increased response rates approximately to the level of face-to-face surveys. However, costs were reduced compared to a single-mode face-to-face survey, because a large share of respondents in the mixed-mode design was interviewed in the web mode.

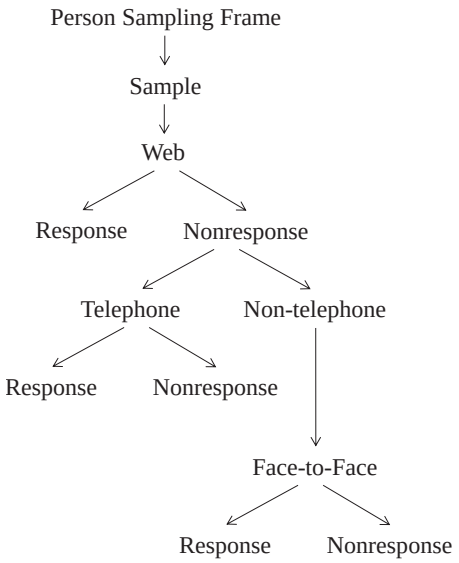

Figure 1 Illustration of the sequential design of the Crime Victimization Survey (in use from 2008 to 2013). A sample is drawn from a list of all population units (sampling frame). After a first attempt to complete the survey in the web mode, non-respondents and non-telephone households are approached either by telephone or face-to-face.

Objectives when designing mixed-mode surveys

An important quality criterion of any survey statistic is its bias. We can distinguish two main sources of bias, selection bias and measurement bias, which we illustrate in the following by a simple example for the estimator of a population mean. In a population of size N let $Y = [y_1, ..., y_i, ..., y_N]$ denote the true scores of a survey target variable Y. Assume a 'pattern mixture model' for Y which stratifies its distribution into respondents and nonrespondents, where $\bar{Y}_{r_m}$ denotes the population response mean and $\bar{Y}_{nr_m}$ the non-response mean. Furthermore, a question asked in some survey mode m is observed with systematic measurement error $\mu^{(m)}$ leading to mode-specific measurement error model $y_i^{(m)} = y_i + \mu^{(m)}$. A simple random sample (SRS) now results in a subset of all N units being approached, indicated by random variable $S = [s_1, ..., s_i, ..., s_N]$, where $s_i = 1$ if unit i is selected and 0 otherwise. Depending on the mode, selected unit i may then either respond or not respond to the survey indicated by $R_m = [r_{m1}, ..., r_{mi}, ..., r_{mN}]$, where $r_{mi} = 1$ if i responds and 0 otherwise. A simple estimator of the population mean $\bar{Y}$ from the response sample of size $n_r = \sum_i s_i r_{mi}$ is

$$\hat{\bar{Y}}_{r_m}^{(m)} = n_r^{-1} \sum_i s_i r_{mi} y_i^{(m)},$$

which has expectation

$$E(\hat{\bar{Y}}_{r_m}^{(m)}) = \bar{Y}_{r_m} + \mu^{(m)}.$$

It can be seen that its total bias is $B_t = \mu^{(m)} + \bar{Y}_{r_m} - \bar{Y}$, where $B_{s_m} = \bar{Y}_{r_m} - \bar{Y}$ represents the selection bias and $\mu^{(m)}$ measurement bias.

The first objective in designing mixed-mode surveys is achieving a reduction of the selection bias B_{s_1} of the initial mode in the mixed-mode design (e.g., web) by adding the follow up of respondents in the second or third mode (e.g., telephone or face-to-face) to the response set of the first mode. This objective can be stated as minimizing criterion $\Delta_s = |B_{s_{mm}}| - |B_{s_1}|$ by design, where $B_{s_{mm}}$ is the selection bias of the mixed-mode design. We require at least $\Delta_s < 0$ and, ideally, $\Delta_s = -|B_{s_1}|$. Now we extend the population response model for the mixed-mode response mean $\bar{Y}_{r_{mm}} = P_1 \bar{Y}_{r_1} + P_2 \bar{Y}_{r_2}$ with non-response stratum mean $\bar{Y}_{nr_{mm}}$, where $\bar{Y}_{r_1}$ and $\bar{Y}_{r_2}$ are the response means in the initial and second mode of the mixed-mode design and $P_1 + P_2 = 1$ the relative sizes of response groups of mode one and two (limiting the illustration here to two modes). It can be seen that the change in B_{s_1} by the follow up is $B_{s_{mm}} - B_{s_1} = P_2(\bar{Y}_{r_2} - \bar{Y}_{r_1})$, where the contrast $SE = \bar{Y}_{r_2} - \bar{Y}_{r_1}$ is called a relative selection effect between modes. It follows that necessarily, but not sufficiently, $\bar{Y}_{r_2} \neq \bar{Y}_{r_1}$ (i.e., presence of a selection effect) for $\Delta_s < 0$. Furthermore, in the absence of a selection effect ($\bar{Y}_{r_2} = \bar{Y}_{r_1}$ or, equivalently, $\Delta_s = 0$) the mixed-mode design surely misses its objective of reducing selection bias of the initial mode and, strictly speaking, it is not needed.

The second objective in designing mixed-mode surveys concerns the size of mode-specific measurement bias $\mu^{(m)}$, which is strongly influenced by the topic of the question and how it is posed. A good example for measurement bias is 'socially desirable answering', which is more common in interviewer-administered than in self-administered modes. When answering desirable the respondent biases the answer in the direction of what (s)he perceives as the social norm. For example, when asked about smoking behaviour, a respondent may perceive less or no smoking as the desirable answer. A strong smoker may then choose to under-report the behaviour to an interviewer. This causes a measurement error where $\mu^{(m)}$ is the average measurement error in the population.

In a mixed-mode survey, it is a threat that some modes may cause larger $\mu^{(m)}$ than others. It can be seen that the measurement bias of a mean estimated from mixed-mode data has form $P_1 \mu^{(1)} + P_2 \mu^{(2)}$. Often, however, it is more practically relevant to assume that one of the modes measures at 'ideal' level, that is, it provides the optimal combination of question, format, and mode. We then may set this mode as 'golden standard' with $\mu^{(m)} = 0$ and express measurement bias with respect to this mode, also called the 'benchmark'. The second objective consequently is to design mixed-mode questionnaires that minimize $P_j \mu^{(j)}$ for all modes j that are not the benchmark.

Problems in assessing the objectives in practice

In designing mixed-mode surveys it is important to estimate selection and measurement biases and the change one may expect when using a mixed-mode instead of a single-mode design. If the size of all biases was known, it would be simple to decide on the benefits of a mixed-mode survey, for example in comparison to only using a web survey. Unfortunately, assessing the objectives is problematic in practice. We take a look at the complications. Figure 2 shows a schematic pattern of available and missing data in a sample surveyed by a sequential mixed-mode design with two modes. White areas indicate data that is observed and grey areas indicate unavailable data. True scores on variable Y are fully unavailable for the whole sample, which frankly is the reason we conduct the survey in the first place. We, however, do obtain measurements of Y from respondents in the initial mode $Y^{(1)}$ with mean $\bar{Y}_{r_1}$ (field A), but there is also some non-response (fields B and C). Non-respondents are followed up in the second mode leading to measurements $Y^{(2)}$ with $\bar{Y}_{r_2}$ (E) and again some nonresponse (F) with $\bar{Y}_{nr_{mm}}$.

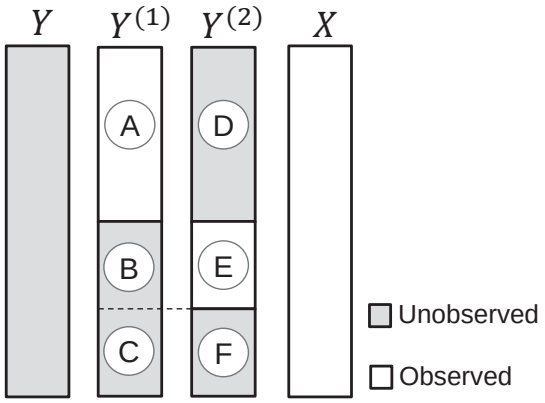

Figure 2 Illustration of the missing data pattern of a sequential design with two modes. The true score vector Y is unobserved and instead measurements $Y^{(1)}$ and $Y^{(2)}$ are observed from respondents to the survey. Some institutes, like Statistics Netherlands, have available sampling frame information (X) on all units.

Due to the missing data it is impossible to estimate any of the biases including the total bias B_t of the survey. At best, the relative difference between mode-specific sample means can be estimated (difference in means of fields A and E). However, this difference amounts in expectation to $SE + \mu^{(j)}$ (where $\mu^{(j)}$ denotes measurement bias of the mode that is not the benchmark). This difference is sometimes called the relative total effect. Statistically, the total effect confounds the relative selection effect between modes with the difference in measurement biases (between benchmark mode and focal mode). Taken by itself, the total effect is quite uninformative. However, if we can disentangle (estimate) both of its components, we can say more about the two design objectives. We would know whether one of the modes may have higher measurement bias than the benchmark ($\mu^{(j)} \neq 0$) and we would know if the design is capable of reducing selection bias of the initial mode ($SE \neq 0$).

Some national offices of statistics including Statistics Netherlands have a set of background information from a register (X), such as socio-demographics, which is available for the full sample or the full population. This information could be used for addressing the missing data problem in two ways. First, let vector R_{mm} describe the mixed-mode response set with element i equal to 1 if $r_{1i} = 1$ (response to the first mode in the design) or $r_{2i} = 1$ (response to second mode in the design) and o otherwise. If we assume conditional independence $P(Y^{(m)} \mid R_1, R_{mm} = 1, X) = P(Y^{(m)} \mid R_{mm} = 1, X)$, also called *missing at random* (MAR) data Y^m in the mixed-mode response set [5], a method for adjusting missing data, such as weighting or imputation, can be used for unbiased estimation of the unobserved ('potential outcome') means $\bar{Y}^{(1)}_{r_2}$ or $\bar{Y}^{(2)}_{r_1}$ in fields B and D. It can be shown that these means allow direct estimation of SE and $\mu^{(j)}$ assuming one of the modes as the benchmark. Second, if we assume $P(Y^{(m)} \mid R_{mm}, X) = P(Y^{(m)} \mid X)$, we may extrapolate the observed data and arrive at an estimate of $\bar{Y}$ as a basis for quantifying total bias.

Whereas MAR assumptions have a strong tradition in statistics they are, unfortunately, hardly ever testable. However, at least for the case of social surveys the assumption often is not plausible. The auxiliary data, X, from population registers at Statistics Netherlands is limited to basic socio-economic data, such as sex, household size, and income, and the observed correlations between response mechanisms R and auxiliary information are often very weak. Although this is not a test of MAR, it seems necessary to find alternative approaches for solving the confounding and extrapolation problem.

The MEPS experiment: an innovative study into mixed-mode design

In my PhD thesis, I developed a framework, outlined partly above, for describing biases and effects between modes in mixed-mode surveys and studied alternative ways of causal inference about these parameters. For this purpose, a large-scale mode experiment was designed and implemented for the case of the Dutch Crime Victimization Survey (CVS) in collaboration with Statistics Netherlands in 2011 [1], called the MEPS experiment (in Dutch: *Mode-effecten in persoonsstatistieken*). The goal of the empirical study was to estimate measurement and selection effects as good as possible.

In a first wave, the four major contemporary modes were administered in parallel to independent samples: face-to-face, telephone, mail, and web. Subsequently, the non-respondents in all modes were re-approached after some weeks' time as in a sequential mixed-mode survey. The follow-up mode was face-to-face in all cases. However, contrary to a standard sequential mixed-mode design also the respondents in all modes were followed up a second time leading to a repeated measurement in face-to-face of many of the CVS target variables.

The missing data pattern of this extended mixed-mode design is shown in Figure 3 for two of the four samples (web and face-to-face modes). It can be seen that the repeated measurement leads to overlap (fields A and E) between the partly observed response vectors $Y_1^{(web)}$, $Y_1^{(f2f)}$ and $Y_2^{(f2f)}$, where indices denote measurement in the first and second wave, respectively. In several empir-

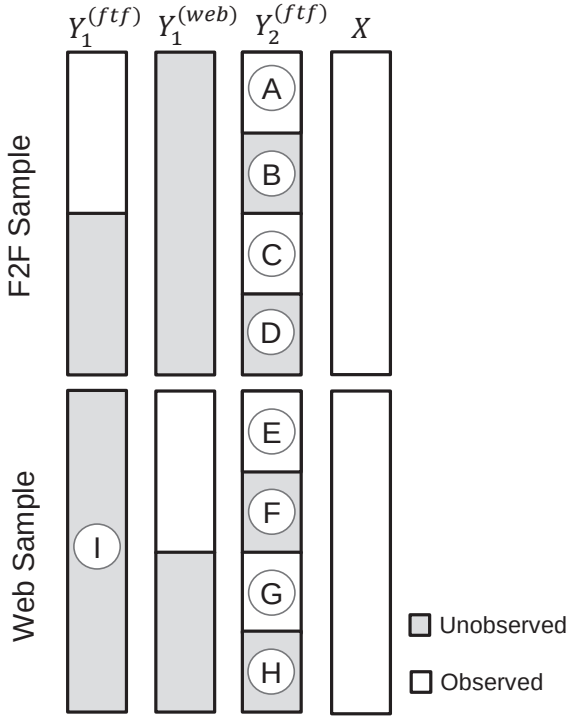

Figure 3 Illustration of the missing data pattern of a sequential design with re-interview. The repeated measures of respondents in the first wave (fields A and E) create overlap between the partly observed response vectors $Y_1^{(web)}$, $Y_1^{(f2f)}$, and $Y_2^{(f2f)}$.

ical studies, the repeated measures were used in different ways to disentangle the biases on CVS target variables. In a study published in *Journal of the Royal Statistical Society*, for example, the second wave face-to-face measurements were considered benchmark data [2]. The authors completed the missing data points in $Y_2^{(f2f)}$ using multiple imputation (fields B, D, F and H). Subsequently, differences in response distributions on $Y_2^{(f2f)}$ were studied per mode, and the change introduced by the face-to-face follow up was evaluated. Innovative in this approach was that the repeated measure could be used like register information. That is, it had the same measurement bias (of face-to-face) in all modes thus avoiding the confounding problem. The authors found that selection bias was about equal in all modes and that it was only marginally impacted by the follow up.

Using an alternative approach, the single-mode face-to-face sample was considered to give the best benchmark measurements ($Y_1^{(f2f)}$) and ideal selection bias [4]. The face-to-face estimate of $\bar{Y}$ thus becomes unbiased by assumption and all other biases are estimated against the face-to-face benchmark estimate. This approach allows quantifying the total bias B_t in a straight-forward way, but it requires estimating unobserved ('potential') benchmark outcomes in field I for units in the comparison mode. Again the repeated measures were used as a basis for this inference. Because the empirical correlations between initial and repeated measures were moderate to large a model of $Y_1^{(f2f)}$ using $Y_2^{(f2f)}$ was anticipated to be stronger than 'usual' models using weak register information X only.

However, also this approach could not identify strong selection bias or relative selection effects between modes. Instead the majority of the total bias was attributed to measurement bias ($\mu^{(m)}$). Here differences were partly very strong, in particular when comparing the self-administered (web, mail) modes with the interviewer-administered modes.

A major conclusion from the MEPS experiment was that it matters chiefly which mode is considered to give 'benchmark' measurements. Depending on this choice either only interviewer-administered or only self-administered modes should be used in the mixed-mode survey. After the MEPS experiment, Statistics Netherlands chose to redesign the CVS using only web and mail in the design.

Where do we go from here?

Mixed-mode surveys have become ever more important in international survey research and they are probably here to stay. The next step in innovation is data collection on 'mobile' devices, such as smartphones or tablet PCs. These devices present new modes and will be used simultaneously in the future.

Methodological research currently progresses in two directions. First, social researchers try to find better questionnaire designs that avoid mode differences in measurement bias optimizing measurement at the level of the 'best' mode. Second, statisticians try to find ways for adjusting measurement bias in mixed-mode surveys. Controlling for the confounding problem of selection effects and measurement bias between modes continues to pose a problem in these endeavours. Building on the PhD thesis, researchers at Utrecht University and Statistics Netherlands, for example, have developed a simulation to investigate under which practical circumstances, such as different measurement error models and strengths of selection effects, re-interview data can lead to better adjusted estimates than unadjusted estimators do [3]. This research may finally lead to important quality indicators and more precise estimates in future mixed-mode surveys.

References

1 B. Buelens, J. van der Laan, B. Schouten, J. van den Brakel and T. Klausch, Disentangling mode-specific selection and measurement bias in social surveys (Discussion paper No. 201211), Statistics Netherlands. The Hague, 2012.

2. T. Klausch, J. Hox and B. Schouten, Selection error in single- and mixed mode surveys of the Dutch general population, *Journal of the Royal Statistical Society, Series A (Statistics in Society) (2015)*, doi: 10.1111/rssa.12102.

3. T. Klausch, B. Schouten, B. Buelens and J. van den Brakel, Adjusting measurement bias in sequential mixed-mode surveys using re-interview data (Discussion paper No. 201523), Statistics Netherlands. The Hague, The Netherlands, 2015.

4. T. Klausch, B. Schouten and J.J. Hox, Evaluating Bias of Sequential Mixed-mode Designs Against Benchmark Surveys, *Sociological Methods & Research* (2015), doi: 10.1177/0049124115585362.

5. R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, Wiley, Hoboken, 2nd ed., 2002.