Inference in surveys with sequential mixed-mode data collection

Bart Buelens and Jan van den Brakel

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201121)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

© Statistics Netherlands,
The Hague/Heerlen, 2011.
Reproduction is permitted.
'Statistics Netherlands' must be quoted as source.

60083201121 X-10

Inference in surveys with sequential mixed-mode data collection

Bart Buelens and Jan van den Brakel

Mixed-mode surveys are known to be susceptible to selection effects and mode-dependent measurement errors, collectively referred to as mode effects. While the use of different data collection modes within the same survey may reduce selectivity of the overall response, it is characterized by measurement errors differing across modes. Inference in sample surveys generally proceeds by correcting for selectivity – for example by applying generalized regression – and ignoring measurement error. When a survey is conducted repeatedly, such inferences are valid only if the measurement error remains constant between surveys. In sequential mixed-mode surveys, where non-respondents in one mode are re-approached using a different mode, it is likely that the mode composition of the overall response differs between subsequent editions of the survey. Variations in the mode composition lead to variations in the total measurement error, invalidating classical inferences. An approach to inference in these circumstances is proposed. First, it must be ascertained that the response is appropriately corrected for selectivity. Second, the mode composition of the response is calibrated towards fixed levels. Assumptions and risks associated with such a procedure are discussed. An example from the Dutch Crime Survey is used throughout the paper to illustrate the proposed approach.

Key words: generalized regression, survey weighting, mode effects, selection bias, response mode calibration

1 Introduction

For a long period of time uni-mode data collection procedures using computer assisted personal or telephone interviewing have been widely applied by national statistical institutes to produce official statistics. Recently the use of multiple data collection modes receives increasing attention, also by national statistical institutes. In a uni-mode data collection approach, all sampling units complete a questionnaire using the same mode. In mixed-mode data collection different respondents in the sample complete the questionnaire using different data collection modes.

Driving factors behind the rising importance of mixed-mode designs in survey sampling are the increasing pressure to reduce administration costs, attempts to reduce non-sampling errors and new technological developments leading to new data collection procedures. Web interviewing, for example, combines the benefits of traditional self-administered data collection modes through paper and pencil with the power of computer assisted administration, i.e. cost effective, greater sense of privacy for the respondent and the possibility of a more complex questionnaire design (Fricker et al., 2005). Despite these advantages, it is also recognized that mixing data collection modes compromises data comparability since measurement errors become more visible, see e.g. De Leeuw (2005), Voogt and Saris (2005), Jäckle et al. (2010), and Vannieuwenhuyze et al. (2010).

In the literature a lot of empirical research aimed at the quantification of different mode effects can be found, see e.g. De Leeuw (2005). This kind of information is necessary to realize a decrease instead of an increase of the total survey error with the introduction of a mixed-mode design (Voogt and Saris, 2005). Dillman et al. (2009) emphasize that a change-over from a uni-mode to a mixed-mode design in a repeatedly conducted survey results in biased measures for change over time of estimated survey parameters, even if the mixed-mode strategy reduces the total survey error. They also note that this is a practical limitation to consider the introduction of a mixed-mode design. Several methods are available to quantify the systematic effect or discontinuity in the main survey parameters due to such a change-over, see e.g. Van den Brakel et al. (2008), Van den Brakel (2008) and Van den Brakel and Roels (2010). The problem, however, that the introduction of a mixed-mode design results in biased measures for change over time is not confined to the moment of the change-over. The distribution of the respondents over the different data collection modes will generally not be constant at the subsequent editions of a repeatedly conducted survey. It can therefore be anticipated that the systematic effect of the total survey error also varies over time. Time series based on repeatedly conducted surveys that employ a mixed-mode design will therefore reflect a more severely biased estimate of mutations over time of the variables of interest compared to uni-mode surveys. The aforementioned methods to quantify discontinuities are not intended to quantify continuously changing measurement bias in the estimates of repeatedly conducted surveys.

The focus in this paper is on identifying an appropriate inference method that is robust for variations in the distribution of respondents over the different data collection modes.

The immediate cause to develop this inference procedure was the introduction of a sequential mixed-mode design in the Dutch crime and victimization survey known as the Integrated Safety Monitor (ISM). Each year a national sample is observed by Statistics Netherlands (SN), using a mixed-mode strategy of Web Interviewing (WI), Paper and Pencil Interviewing (PAPI), Computer Assisted Personal Interviewing (CAPI) and Computer Assisted Telephone Interviewing (CATI). In addition, local authorities like municipalities and police districts can draw additional samples on a voluntary basis to obtain adequate detailed estimates for their regions. The data collection in these additional samples are for budgetary reasons typically based on WI, PAPI and CATI only. The combination of a sequential mixed-mode design with the regional oversampling results in extreme variations of the distribution of the respondents over the different modes. There are strong indications that the estimates based on the regular linear weighting procedure are implausible because the measurement bias in the successive samples of the ISM varies with the composition of the respondents over the the modes. In an attempt to keep the measurement bias in the survey estimates constant, it is proposed to calibrate the survey weights to a fixed distribution of the population over the modes. This should remove the mode effects from mutations in the survey estimates, at least partially. Although it is preferable to keep the mode composition as constant as possible by the design of the survey and the conduction of the field work, fluctuations cannot be precluded in mixed-mode designs. The proposed method is therefore generally applicable for inference in mixed-mode surveys.

The paper starts with a review of mode effects in section 2. The survey design of the ISM is described in section 3. Section 4 reviews the concept of linear weighting and develops a practical calibration method that is more robust for variations in the mode composition of the response. This approach is applied to the ISM in section 5. The paper concludes with a discussion in section 6.

2 Mode effects in mixed-mode surveys

Data collection modes affect multiple sources of non-sampling errors in the data collection phase of a survey process. The mode determines the coverage of the target population and the response rate of the approached sampling units and therefore the representativeness of the final response. The mode also affects the amount of measurement error that obscures the real values of the variables of interest in the answers produced by the respondents. Mode effects are commonly defined as the observed differences in survey outcomes if the survey is administered using different data collection modes. They are the result of selection effects and measurement effects. Selection effects arise since different modes have different coverage and response rates, and are a form of nonresponse bias. Measurement effects arise since different modes evoke different types of measurement errors during the process of reporting an answer.

In the literature, various mixed-mode strategies are developed. One possible classification is to distinguish between mixed-mode designs where respondents can choose

between different modes, and sequential strategies where multiple contact attempts are implemented using different modes in a prespecified order. Sequential mixed-mode strategies are particularly cost effective since they can start with the self-administered modes that have low administration costs and use interviewer-administered modes to re-approach the remaining nonrespondents. Attempts to reduce administration cost are not a recent development in survey sampling. Hochstim (1967) already conducted a field experiment to compare data quality between different data collection modes with the purpose to consider cost effective data collection strategies.

Mixed-mode designs are useful to improve coverage and response rates because different modes are available to contact different groups of hard to reach respondents, and respondents can choose a preferred mode from several options (De Leeuw, 2005; Voogt and Saris, 2005; Vannieuwenhuyze et al., 2010). Dillman et al. (2009) emphasize that sequential mixed-mode strategies are particularly effective to improve coverage and response rates.

In survey literature many references can be found to experiments indicating that different data collection modes have substantial effects on the answers generated by respondents, see De Leeuw (2005) for an overview. These differences can be explained using cognitive models of the survey process (Channel et al., 1981; Tourangeau et al., 2000), which provide a framework for understanding the process by which respondents interpret questions, retrieve the required information, make judgements about the adequate response and come up with an answer. These models are useful to explain how the characteristics of different data collection modes affect this process differently, resulting in different measurement bias.

The presence or absence of an interviewer is one of the most important factors explaining the observed differences in response between different data collection modes. Several studies indicate that respondents are more likely to offer socially desirable answers and demonstrate acquiescence, i.e. the tendency of respondents to agree with statements, in the presence of an interviewer than in self-administered surveys (Dillman et al., 2009; De Leeuw, 2005; Holbrook et al., 2003). It might also be expected that satisficing, i.e. respondents do not expend the maximum amount of cognitive effort required to answer a question to the best of their capabilities (Krosnick, 1991) occurs more frequently in self-administered surveys than in interviewer-administered surveys, depending on the layout of the questionnaire (Fricker et al., 2005). Between interviewer-administered surveys, satisficing will occur more frequently in CATI modes, since the interview speed in telephone interviews is generally higher compared to face-to-face interviews. Self-administered questionnaires have the advantage of evoking a greater sense of privacy than personal interviewing resulting in more openness and self-disclosure and therefore in increased data validity, particularly for sensitive topics (Tourangeau and Smith, 1996; De Leeuw, 2005; Voogt and Saris, 2005; Aquilino and LoSciuto, 1990; Aquilino, 1994).

Several studies demonstrate that respondents tend to answer more positively to opinion

questions in aural orientated modes compared to visual orientated modes (Krysan et al., 1994; Dillman et al., 2009; Christian et al., 2008). Other well known factors that explain differences between visual and aural based data collection modes are recency and primacy effects. Primacy effects refer to the phenomenon that respondents tend to choose items listed first in a list of answer categories in visual based modes. This can be explained because in a visual presentation the items listed first are subjected to deeper cognitive processing (Krosnick and Alwin, 1987; Dillman et al., 2009). Recency effects refer to the phenomenon that respondents tend to choose the last items in a list of answer categories if they are read by an interviewer. The explanation is that in interviewer based modes, there is not enough time to place each answer category into the respondents long-term memory before the next one is read (Krosnick and Alwin, 1987; Dillman et al., 2009).

The interaction between mode and questionnaire design is discussed by Dillman and Christian (2005) and De Leeuw (2005). The literature is not decisive whether the same questionnaire should be used for each mode in a mixed-mode design, or mode-specific questionnaires should be applied. An example of this debate is the problem how to handle the "don't know" option in attitude questions. Generally, interviewer-administered surveys tend to hide the "don't know" option from respondents and interviewers are only allowed to record such answers if a respondent really "doesn't know" (Dillman and Christian, 2005). In self-administered questionnaires, such categories are frequently left out, resulting in more pseudo-attitudes or respondents skipping the question. Alternatively, they may be explicitly listed, which might result in more pseudo-nonattitudes since it offers the respondent an easy way out, (Gilljam and Granberg, 1993), and can be seen as an example of satisficing mentioned before.

In self-administered surveys, the visual layout of the questionnaire is of importance since differences in layout generally result in significant differences in response, (Stern et al., 2007; Toepoel et al., 2009). Tourangeau et al. (2004) and Tourangeau et al. (2007) distinguished five visual interpretive heuristics that respondents follow when completing a self-administered questionnaire. The extend in which these heuristics are obeyed or violated determines the amount of measurement bias in the observed data. Based on this awareness, a growing body of empirical research provides a foundation for what is called visual design theory for self-administered surveys, see e.g. Dillman (2007).

The fact that data collection modes can affect multiple nonsampling errors in different directions, implies that mixed-mode designs do not necessarily improve data quality (Voogt and Saris, 2005). A reduction of selection effects can be counterbalanced by an increase of measurement error, or even worse. Quantitative insight into the different effects of data collection modes on selection effects and measurement errors is therefore required to choose an optimal field work strategy. A massive amount of literature on empirical research on mode effects appeared in the literature, see e.g. De Leeuw (2005) for an overview. Selection and measurement effects are typically strongly confounded when survey outcomes obtained under different modes are compared. Separation of selection effects from measurement

effects in empirical studies requires carefully designed experiments in combination with weighting or regression based inference methods to control for selection effects, see e.g. Jäckle et al. (2010). As an alternative, Vannieuwenhuyze et al. (2010) proposed a method to disentangle measurement and selection effects on the proportions and the mean of multinomial variables. Biemer (2001) applied an interview re-interview approach analyzed with a latent class model to disentangle selection bias and measurement bias in CAPI and CATI modes.

3 The Integrated Safety Monitor

The purpose of the Dutch Integrated Safety Monitor (ISM), is to publish information on crime victimization, public safety and satisfaction with police performance of the Dutch population aged 15 years and over. This survey is conducted in its current form since 2008. Preceding versions of this survey are the Crime Victim Survey (1981-1991), the Legal and Security Survey (1992-1996), the Permanent Survey on Living Conditions (1997-2004) and the National Safety Monitor (2005-2008).

The ISM is based on a stratified two stage sample design of persons aged 15 years or older residing in the Netherlands. The Netherlands is divided into 25 police districts, which are used as the stratification variable in the sample design. In the first stage municipalities are drawn with selection probabilities proportional to the population size. In the second stage, persons are drawn from the municipalities selected in the first stage, with a minimum sample size of one. This results in an optimal regional distribution of the sample.

SN conducts a national sample of the ISM with a size of about 19.000 respondents. The sample is equally divided over the strata resulting in a target response of about 750 respondents in each stratum. Selection probabilities are chosen such that within each stratum a self-weighted sample is drawn. Since police districts have unequal population sizes, inclusion probabilities vary between police districts. This results in an allocation which is optimal for estimating parameters at the level of police districts.

Municipalities and police districts can participate in the survey on a voluntary basis. In these regions, additional samples are drawn with the purpose to provide precise estimates for these regions.

SN collects data for the national sample using a sequential mixed-mode design that is based on WI, PAPI, CAPI and CATI. All persons included in the sample receive an advance letter where they are asked to complete a questionnaire via internet (WI). Persons can receive a paper version of the questionnaire on their request (PAPI). After two reminders, nonrespondents are contacted by telephone if a telephone number is available to complete the questionnaire (CATI). The remaining persons are visited at home by an interviewer to complete the questionnaire face to face (CAPI). The data collection of the additional regional samples is conducted by different marketing research bureaus. For these samples WI, PAPI and CATI modes are mandatory. The use of the CAPI mode is

recommended but not mandatory since this mode is very costly. The additional value of the CAPI mode in the ISM is studied by Buelens and Van den Brakel (2010).

Statistical inference is based on the generalized regression (GREG) estimator (Särndal et al., 1992). The size of the SN-sample allows adequate estimation of population parameters at the national level and at the level of police districts. For regions where additional local samples are selected, parameters are estimated at a more detailed level, e.g. municipalities and even neighborhoods.

During the three editions of the ISM, additional responses are collected in the regional samples. The number of responses obtained in the regional samples fluctuated with a factor four between the successive editions of the ISM, and results in a strong variation in the distribution of the response over the different data collection modes. There are strong indications that this causes unstable results for estimates of levels as well as mutations over time. It is also recognized that the estimates based on the national sample of SN are rather stable, since the distribution of the respondents in the national sample over the modes as well as the sampling fractions are rather stable in the different editions of the ISM. The response rate of the national sample is significantly higher compared to the regional samples. In combination with the use of the CAPI mode, it can therefore be anticipated that with the national survey a more representative sample is obtained, which can be used as a benchmark.

The purpose of this paper is to develop a calibration estimator that is more robust for fluctuations in the response composition. The details are developed in section 4. Results for the ISM are given in section 5.

4 Methods

4.1 Classical GREG estimation

At SN and many other national statistical institutes, inference in person and household sample surveys is commonly conducted using generalized regression or GREG-estimation (Särndal et al., 1992). GREG-estimators are regarded as model-assisted estimators. The motivating model is a linear relationship between the target variable y and a vector of auxiliary variables x for which the population totals t_x are known,

$$y_k = x_k^t \beta + e_k, \tag{1}$$

with regression coefficients β and errors e_k. The associated GREG estimator for t_y, the total of y, is

$$\hat{t_y} = \hat{t}_y^{HT} + (t_x - \hat{t}_x^{HT})^t \hat{\beta}^t \tag{2}$$

with $\hat{t}_y^{HT}$ and $\hat{t}_x^{HT}$ the Horwitz-Thompson estimators for the totals of y and x respectively, and $\hat{\beta}^t$ estimates of the regression coefficients, see (Särndal et al., 1992) for details.

GREG-estimators can be expressed as weighted sums of the sample observations. This is particularly useful in multipurpose surveys, which most common person and household surveys are. Writing the GREG estimator in this way gives

$$\hat{t_y} = \sum_{k=1}^{n} w_k y_k \tag{3}$$

with n the sample size, and weights independent of y,

$$w_k = \frac{1}{\pi_k}(1 + \frac{x_k^t}{v_k}(\sum_{k=1}^{n} \frac{x_k x_k'}{v_k \pi_k})^{-1}(t_x - \hat{t}_x^{HT})), \tag{4}$$

with π_k the probability that unit k is included in the sample and v_k the variance of the errors in the linear model (1).

If (1) fits the data in the finite target population reasonably well, then the use of auxiliary information through the GREG estimator reduces the sampling variance and corrects for selective nonresponse, at least partially (Särndal and Lundström, 2005; Bethlehem, 1988).

4.2 Mode effects

As explained in section 2, by using multiple modes of data collection, two effects can be present: selection effects, and mode-dependent measurement errors. Selection effects arise when particular subpopulations exhibit a preference for one or more particular modes or through different coverage of the modes. For example, Buelens and Van den Brakel (2010) show that in the ISM internet response rates are lower for elderly people and non-western immigrants. In general, respondents observed through different modes differ with respect to auxiliary variables x, survey variables y, or both. Selectivity with respect to x is easily corrected for, by including these x variables in the GREG weighting model. Selectivity with respect to y is more difficult to assess, as it cannot be determined from survey data typically at hand. In fact, this situation is no different to weighting as correction for non-response. The statistician must strive to remove selectivity with respect to y by including covariates x in the model explaining the selectivity with respect to y. Analogously, mode related selection effects with respect to y can be corrected for by including appropriate covariates in the weighting model. In what follows, the assumption is made that the covariates x fully explain the mode related selectivity with respect to y, see section 4.6.

In classical sampling theory, the GREG estimator is considered as a design-based or model-assisted estimator. Parameter and variance estimators are derived under the concept of repeatedly drawing samples from a finite population according to the applied sampling design with all population parameters held fixed. Under this approach it is assumed that the observations obtained at the sampling units are true fixed values observed without error. The weights attached to the observations in (4) are derived from the probability structure of the sampling design and the available auxiliary information for which population totals are known. These weights are intended to correct for unequal

selection probabilities, skewness of the sample and selectivity due to nonresponse and coverage errors. It can therefore be expected that GREG estimators remove mode dependent selection effects in the target variables, at least partially. It cannot be expected, however, that GREG estimators correct for measurement error, not even partially. The traditional notion, followed in classical sampling theory, that observations obtained form the sampling units are true fixed values observed without error is not tenable if the effect of measurement errors due to mixed-mode data collection in design-based inference is investigated. Therefore a measurement error model explaining the systematic bias due to different modes is required. The measurement error model is written as

$$y_{k,m} = u_k + b_m + \varepsilon_{k,m} \tag{5}$$

with $y_{k,m}$ the observations using mode m of the true intrinsic values u_k, for units $k = 1, ..., n$, and b_m and $\varepsilon_{k,m}$ systematic and random error components respectively. It is assumed that $E_e(\varepsilon_{k,m}) = 0$, where the subscript e stands for the expectation with respect to the measurement error model. It is further assumed that b_m is constant for all units k observed through mode m. If, however, the systematic errors b_m are dependent on a known categorical variable x for a given mode, for example due to different response propensities, b must be defined for the cross classification of mode with x. Without loss of generality it is further assumed that the subscript m refers to an appropriate cross classification, and takes on values in the range $1, ..., p$.

The expression of the GREG estimator, eq. (3), then becomes

$$\hat{t}_y = \sum_{k=1}^{n} w_k(u_k + b_m + \varepsilon_{k,m}), \tag{6}$$

with expectation

$$E_s E_e(\hat{t}_y) = E_s E_e(\sum_{k=1}^{n} w_k(u_k + b_m + \varepsilon_{k,m})) = t_u + E_s(\sum_{k=1}^{n} w_k b_m). \tag{7}$$

The expectations E_s and E_e are with respect to the sample and the measurement error model respectively. The quantity $\sum_{k=1}^{n} w_k b_m$ is the total measurement error. It is the difference between the expected value of the estimator, and the total of u. In (7) it is assumed that $E_s\hat{t}_u = t_u$, is not biased with selection effects. The validity of this assumption depends on the extent to which the weighting model corrects for selection effects and the extent to which the field work approach results in a representative sample.

It is useful to introduce notation to simplify subsequent expressions. The total measurement error is written as

$$\begin{aligned} \sum_{k=1}^{n} w_k b_m &= \sum_{m=1}^{p}\sum_{k=1}^{n} w_k \delta_{mk} b_m \\ &= \sum_{m=1}^{p} b_m \sum_{k=1}^{n} w_k \delta_{mk} \end{aligned}$$

$$\equiv \sum_{m=1}^{p} b_m \hat{t}_m \tag{8}$$

with δ_{mk} equal to 1 if unit k is observed using mode m, and zero otherwise. The $\hat{t}_m$ are the estimated population totals of units responding through mode m.

The expectation of the total measurement error as occurring in eq. (7) is now written as

$$E_s(\sum_{k=1}^{n} w_k b_m) = E_s(\sum_{m=1}^{p} b_m \hat{t}_m) = \sum_{m=1}^{p} b_m E_s(\hat{t}_m) \tag{9}$$

since the b_m are constants. In regular surveys, not incorporating dedicated experiments, the measurement errors b_m cannot be identified. The total measurement error remains included as an unidentified bias in the survey estimates.

The total measurement error is zero only if $b_m = 0$ for all m, in which case the levels of the target variables are estimated unbiasedly. If $b_m = b$ for all m, then the total error is expected to be constant, and is effectively independent of the data collection modes. In this case the real mutations of the target variables are estimated unbiasedly. Since these are desirable circumstances, common approaches are to design the survey, questionnaire and data collection process in such a way that the b_m are small, or at least that they do not vary a lot across modes. This is difficult to achieve and test. Pilot studies and experiments can be conducted to establish this.

In the general case, where the b_m do vary across modes – but are constant between surveys, and over time – the expected value of the total measurement error is constant between surveys only if the $E_s(\hat{t}_m)$ are constant. This is the case, for example, in survey designs where sample units are randomly assigned to data collection modes, according to a fixed distribution. The $E_s(\hat{t}_m)$ then correspond to this design distribution.

Under a sequential mixed-mode design, however, the composition of the response in terms of data collection mode is left free, and no natural population totals exist. Under such designs, the $E_s(\hat{t}_m)$ can vary between surveys. Such fluctuations cause variations in the total measurement error of the survey outcomes, causing varying biases. This variation is confounded with true mutations over time in the underlying variables, and cannot be isolated. This is the core problem addressed in the present paper.

4.3 Mode calibration

Interest in sample survey outcomes is largely focused on mutations over time. In sequential mixed-mode surveys, mutations in $\hat{t}_y$ are confounded with mutations in the response mode composition $\hat{t}_m$, compromising the validity of the results. In this section a practical approach is developed, aimed at stabilizing the total measurement error.

Assume a sequential mixed-mode survey design, executed twice, simultaneously, but independently. Assume hereby different realizations of the response mode compositions.

The expected difference between the survey outcomes is ideally zero, unaffected by the differences in mode composition:

$$E_s E_e(\hat{t_y}^{(2)} - \hat{t_y}^{(1)}) = 0 \tag{10}$$

For this to hold, using equations (7) and (8), it is required that

$$\begin{aligned} E_s(\sum_{m=1}^{p} b_m \hat{t}_m^{(2)} - \sum_{m=1}^{p} b_m \hat{t}_m^{(1)}) &= 0 \\ \Leftrightarrow \sum_{m=1}^{p} b_m (E_s(\hat{t}_m^{(2)}) - E_s(\hat{t}_m^{(1)})) &= 0 \end{aligned} \tag{11}$$

For this to hold, with $b_m \neq 0$ in general, it is required that

$$E_s(\hat{t}_m^{(2)}) - E_s(\hat{t}_m^{(1)}) = 0 \tag{12}$$

for all $m = 1, ..., p$. For a given m, the quantity $E_s(\hat{t}_m^{(i)})$ is the expected number of people to respond to the survey using data collection mode m in survey $i = 1, 2$. This quantity is undefined in sequential mixed-mode surveys, as there is no naturally expected number of people to respond through a particular mode.

The fulfilment of the requirement in expression (12) can be achieved by explicitly imposing conditions on the weights, such that $\hat{t}_m^{(i)}$ is constant for all m in both surveys $i = 1, 2$.

Imposing such conditions on the weights with regard to the response mode distribution is an artificial calibration. The calibration levels can be chosen arbitrarily. With $\Upsilon = (\Upsilon)_m$ the chosen levels for modes m, the conditions to be imposed on the weights are

$$\hat{t}_m = \sum_{k=1}^{n} w_k \delta_{mk} = \Upsilon_m \tag{13}$$

for all m. These expressions are of exactly the same form as the conditions typically imposed on the GREG weights in terms of the usual auxiliary variables x. Consequently, including mode calibration in a GREG estimator is achieved by including the response mode as an explanatory variable in the weighting model, and using arbitrarily chosen calibration levels Υ_m as if it were population quantities.

With weights calibrated in this way, requirement (12) holds, and therefore the expected values of the estimates of the two totals are equal. The effect of variations in mode compositions are neutralized. It is important to note that the total measurement error has not been quantified, nor has it been removed. It is merely kept constant, and is now equal to

$$\sum_{k=1}^{n} w_k b_m = \sum_{m=1}^{p} b_m \hat{t}_m = \sum_{m=1}^{p} b_m \Upsilon_m, \tag{14}$$

with – as discussed above – the b_m unidentified, but assumed constant.

4.4 Change over time

If two surveys are conducted in parallel, under the same conditions, then the measurement bias in $\hat{t}_y$ can be expected to be constant. If the survey is conducted repeatedly over time, more variation in the measurement bias in $\hat{t}_y$ can be expected, for example due to a gradual increase of internet access or decrease of the population for which a telephone number is available.

Consider a situation similar to the one discussed above, but now with the second survey conducted at a later time, after the first survey. Interest is in the change over time in the value of t_y, estimated by

$$\hat{\Delta}(t_y) = \hat{t}_y^{(2)} - \hat{t}_y^{(1)} \tag{15}$$

The expected value of this quantity is

$$E_s E_e(\hat{\Delta}(t_y)) = E_s E_e(\hat{t}_y^{(2)}) - E_s E_e(\hat{t}_y^{(1)}). \tag{16}$$

Applying the proposed mode calibration, this is written as

$$E_s E_e(\hat{\Delta}(t_y)) = (t_u^{(2)} + \sum_{m=1}^{p} b_m \Upsilon_m) - (t_u^{(1)} + \sum_{m=1}^{p} b_m \Upsilon_m) \tag{17}$$

Assuming that the systematic components b_m are constant over time, and that the modes are calibrated to the same levels Υ_m,

$$E_s E_e(\hat{\Delta}(t_y)) = t_u^{(2)} - t_u^{(1)}. \tag{18}$$

The difference of the estimated totals is now an unbiased estimator of the change in the totals of the underlying variable u. This is a desirable outcome, as this quantity is no longer affected by measurement errors.

Without the proposed calibration, $\hat{t}_m^{(1)} \neq \hat{t}_m^{(2)}$ generally, and part of the observed change would be caused by changes in the response mode compositions, causing a bias in the estimated temporal change.

A potential risk of mode calibration is that real mutations are suppressed if the composition of the respondents of the different modes changes gradually over time. If such selective changes with respect to t_m occur, it is assumed that the socio-demographic variables used in the weighting model correct for this selectivity, and hence that the mode calibration has no adverse effect by suppressing true mutations. See also section 4.6 below.

4.5 Choosing appropriate calibration levels

It is important to avoid extreme variations in the weights in GREG estimation, as they can inflate the variance. Hence calibration levels for the modes are preferably chosen such that they do not incur large inflations in the dispersion of the weights.

In the absence of mode calibration, denote with $\hat{t}_m^{GREG}$ the GREG estimates of the mode composition as they arise from the sample weighted using the regular model. Intuitively, if these are close to the calibration levels Υ_m, the mode calibration will not have a large effect on the weights, and therefore on the GREG estimates of survey variables.

Since the levels Υ_m can be set arbitrarily, it is advisable to choose them such that they are roughly within the range of actually occurring levels. If a survey has taken place once or more, the GREG levels of these surveys can be used as a guide. An alternative is to estimate response propensities, and use these to derive estimates of the mode distribution in the population. In practice, the same auxiliary variables would be used as those included in the weighting model, and the results would be in line with those found based on the weighted sample.

Using calibration levels Υ_m that are very different from the $\hat{t}_m^{GREG}$ may cause a large increase in the dispersion of the weights, with possibly even negative weights as a result.

Since the aim of the calibration is to neutralize changes in the total measurement error, it is not critical what the precise levels are set to in a particular survey.

4.6 Assumptions and risks

The standard GREG estimator is given in eq. (6), and

$$\hat{t}_y^c = \sum_{k=1}^{n} w_k^c (u_k + b_m + \varepsilon_{k,m}) \tag{19}$$

is the mode-calibrated GREG estimator, with the superscript c indicating mode calibrated versions of the parameters. Generally, $\hat{t}_y \neq \hat{t}_y^c$.

The key assumption underpinning the presented approach is that

$$\sum_{k=1}^{n} w_k u_k = \sum_{k=1}^{n} w_k^c u_k. \tag{20}$$

This is equivalent to stating that response mode calibration has no effect in the absence of measurement errors.

Mode calibration is aimed at altering the measurement error, rendering it constant. Consequently it will generally be the case that

$$\sum_{k=1}^{n} w_k b_m \neq \sum_{k=1}^{n} w_k^c b_m. \tag{21}$$

It is this inequality causing the standard estimate $\hat{t}_y$ to differ from its calibrated version $\hat{t}_y^c$.

However, since the u_k are not known explicitly, and the b_m are unknown, it is impossible to guarantee that an observed difference between $\hat{t}_y$ and $\hat{t}_y^c$ is due only to the difference as in eq. (21). It can be the case too that assumption (20) does not hold. This can happen in situations where the response mode has some additional explanatory power in predicting y after controlling for the auxiliary variables x. Including response mode as an independent variable in the regression model in this case will affect the estimates. Using

arbitrary population levels in the calibration will introduce a bias in the estimates, as the GREG is asymptotically unbiased only if the true population totals of the covariates in the model are used, i.e. the population totals in t_x are defined exactly the same as the covariates x_k of the individual respondents.

If constant calibration levels for the response modes are used, true changes in the target variable y may be suppressed by the calibration. Since this is a serious risk, it is advisable to establish validity of the assumption. One could seek for an additional variable z, known for the sampled units as well as for the population, such that z correlates with y and m. Weighting the survey response using covariates x, these x remove selectivity with respect to z, if

$$\hat{t}_z^{GREG} \approx t_z, \tag{22}$$

as z has no associated measurement error because it is a known register variable. Condition (22) can be formalized by significance testing of the difference of the two quantities. If the GREG weighted sample remains selective with respect to z, a straightforward action is to including z in the weighting model. Checking many variables in this way builds confidence that the GREG weighting model removes all selectivity.

In addition, the results of applying different calibration levels can be analyzed. This may provide insight into the effect of mode calibration. If the final estimates do not vary with varying calibration levels, the measurement errors of the target variable do not depend on response mode. In this case, applying mode calibration is not necessary. If the results do depend on the chosen calibration levels, mode calibration does have an effect. If the assumption holds, the effect of mode calibration is the leveling out of measurement bias, as intended. If the assumption does not hold, mode calibration is introducing a bias.

Not applying response mode calibration in the presence of mode effects gives rise to a bias due to varying measurement errors, while applying weight calibration without the assumption being fulfilled leads to a bias too. External results from other sources or expert intuition can provide assistance in the approach to be followed. Truly validating the assumption may be achieved by conducting an experiment designed and analyzed such that selectivity effects and measurement bias can be separated (Jäckle et al., 2010; Vannieuwenhuyze et al., 2010).

5 Results

5.1 Integrated Safety Monitor

The ISM is a sequential mixed-mode survey, introduced in section 3. It has been executed three times. As discussed earlier, the extent to which local authorities participate in the oversampling scheme affects the mode composition of the integrated sample. Table 1 gives an overview of the oversampling and the number of respondents for the years 2008, 2009 and 2010. Part of the ISM response consists of data collected by SN, and is aimed at

producing statistics at the national level as well as at the level of police districts. Hence, the SN-sample is a subset of the ISM-sample, the latter being the combination of the SN-sample and the local samples. Associated with the variations in oversampling in the three years, are variations in the composition of the response modes, see Table 2. In 2009, when the population in oversampled areas accounted for 65% of the total population, 69% of the responses were obtained through internet, with only 1.5% through personal interviews.

Inclusion probabilities vary extremely within single editions of the ISM. Not only are there differences between units inside and outside oversampled areas, but inclusion probabilities vary greatly between oversampled areas, as local authorities are free to choose desired sample sizes for their areas. For example, in the ISM 2009, inclusion probabilities varied from 1/1053 in larger police districts not oversampled, to 1/6 in intensively oversampled neighborhoods. The smallest inclusion probability in oversampled areas was 1/390. This has an adverse effect on the variance of the estimates, rendering the estimators inefficient as illustrated with Basu's circus elephants (Basu, 1971) . The confidence intervals of country level estimates based on the ISM-sample turn out not to be much smaller than those based on the SN-sample only (see below), despite the much larger size of the ISM-sample, especially in 2009.

	2008	2009	2010
Number of oversampled municipalities	77	239	21
Size response SN-sample	16,964	19,202	19,238
Size response local sample	45,839	182,012	19,982
Percentage of population in oversampled areas	29%	65%	16%

Table 1. Overview of oversampling in ISM surveys 2008 - 2010.

		2008	2009	2010
SN-sample	WI	40.1%	47.3%	49.5%
	PAPI	15.4%	16.1%	13.2%
	CATI	34.0%	25.3%	23.8%
	CAPI	10.5%	11.3%	13.6%
ISM	WI	55.9%	68.9%	61.4%
	PAPI	11.4%	12.3%	12.0%
	CATI	26.6%	17.3%	19.6%
	CAPI	6.0%	1.5%	7.0%

Table 2. Overview of response mode composition in ISM surveys 2008 - 2010, not weighted.

5.2 GREG estimation

The so-called SN-sample is independent of the local samples, and can be used for estimation, without using the locally collected data. The advantage of doing so is that this provides reference figures at the national level, which are obtained in a similar manner in all three years. However, the officially released figures are those obtained from the integrated sample, i.e. the combination of the SN-sample and the local samples into one single data set. In what follows the distinction shall be made between results based on the SN-sample and on the ISM-sample.

In 2008, when the ISM was first conducted, a GREG weighting model was established. This was the result of experiences with earlier crime surveys, and a preliminary methodological study. The model includes background variables age, gender, ethnicity, urbanization, household size, police district, and the strata used in the regional oversampling scheme.

For the purpose of this study five key survey variables are considered, see Table 3.

Variable	Description
victim	Percentage of people indicating to have been a victim of crime in the last 12 months
offences	Number of offences per 100 inhabitants in the last 12 months
unsafe	Percentage of people feeling unsafe at times
funcpol	Satisfaction with functioning of police (on a scale 1-10)
antisoc	Suffering from anti social behavior (on a scale 1-10)

Table 3. Overview of key ISM variables.

Table 4 presents annual mutations in estimates for the key survey variables. In general the mutations in the ISM-sample are larger than in the SN-sample. The mutations in victim and offences are suspicious, as the ISM-sample shows a significant increase in 2008-2009 and a significant decrease in 2009-2010, while the changes are not as extreme, and not significant, when based on the SN-sample. The three other variables are less suspect, in particular the mutations in unsafe in the ISM and SN-samples are very similar, though rather large. Due to the inefficiency of the oversampling, the variances of the ISM estimates are not much smaller than those of the SN estimates, despite the large differences in sample sizes.

The variables relating to feeling unsafe, satisfaction with the police, and suffering from anti social behavior describe attitudes and opinions. The variables relating to victimization and offences are factual, and can be validated to some extent by comparison with other data sources. Expert assessment, using data from the Police on the number of registered offences, shows that the ISM peak in 2009 in the variables victim and offences is not plausible. An increase of 5% and 9% respectively from 2008 to 2009, followed by a decrease

Variable	Sample	2008-2009	p	2009-2010	p
victim	ISM	+5.0%	<.001	-6.7%	<.001
	SN	+0.9%	0.347	-3.0%	0.055
offences	ISM	+9.2%	<.001	-12.5%	<.001
	SN	+3.3%	0.146	-6.1%	0.011
unsafe	ISM	+7.0%	<.001	-1.4%	0.159
	SN	+7.7%	<.001	+0.2%	0.126
funcpol	ISM	-3.5%	<.001	+1.5%	<.001
	SN	-1.9%	<.001	0.0%	0.467
antisoc	ISM	+5.2%	<.001	-1.3%	0.149
	SN	+3.5%	0.034	+0.5%	0.388

Table 4. Annual mutations in key ISM variables using the standard GREG estimator.

of 6% and 12% in 2010, is unlikely. The SN-sample shows no significant change in these time periods, a finding confirmed through a confrontation with Police register data.

The reason for these implausible changes measured by the ISM must be sought in differences in the different ISM editions themselves. From Tables 1 and 2, it is seen that 2009 is an unusual year in comparison with 2008 and 2010, in that 2009 was characterized by massive oversampling, and associated with that, the collection of a large proportion of WI. Prior to elaborating on this, it must be established that the ISM-samples of all three years are representative with respect to background variables, in particular those correlating with victimization.

5.3 Correcting selection effects

The response weighted using the standard model may still be selective with respect to variables correlated with the key target variables (see paragraph 4.6). In order to investigate this further, a number of additional variables are identified (i) that correlate well with the key target variables, (ii) that are known for the whole population, and (iii) that are not already included in the standard weighting model.

There is generally a negative correlation between socio-econmic status and both victimization rates and response rates. Therefore variables indicative of socio-economic class are used for this analysis. Unfortunately, the variables are not available at the unit level, but only in an aggregated form. The aggregation level is that of post code areas, of which there are approx. 4,000 in the Netherlands. The size of the target population of people aged 15 and over is approx. 3,250 per area on average. The additional variables are average house value, percentage of inactive people (i.e. not in the labor force), the number of with the police reported crimes per 1,000 inhabitants, and degree of urbanization (at a more detailed level than that included in the standard weighting model). These variables

are all categorized, see Table 5. This table shows the estimates for these variables, both using the standard weighting model and the model including the calibration (details of mode calibration are given in the next section). The known population quantities are shown in the last column in Table 5. Almost all are within two standard errors from their corresponding estimates. Category 4 of the inactive population is the only exception, with areas with many inactive people slightly underrepresented in the weighted samples. This is not considered a cause for concern, as the distributions for this variable are unaffected by including mode calibration in the weighting model. A similar argument could be made for urbanization, where rural areas are somewhat underrepresented, although not severely, and the estimated distribution is again not affected by including the mode calibration.

This analysis provides support for the standard weighting model, in it being sufficient in correcting for selective nonresponse. However, selectivity with respect to other variables may remain, in particular selectivity with respect to survey variables is difficult to assess.

Variable	Category	ISM standard	ISM calibrated	Population
House value	Cat. 1 (low)	31.3% (0.2)	31.3% (0.2)	31.3%
	Cat. 2	25.2% (0.2)	25.4% (0.2)	25.4%
	Cat. 3	22.5% (0.2)	22.4% (0.2)	22.2%
	Cat. 4 (high)	20.9% (0.2)	20.9% (0.2)	21.0%
Inactive pop.	Cat. 1 (few)	20.0% (0.2)	20.0% (0.2)	19.9%
	Cat. 2	20.2% (0.2)	20.2% (0.2)	19.9%
	Cat. 3	25.3% (0.2)	25.3% (0.2)	25.2%
	Cat. 4 (many)	34.5% (0.2)	34.5% (0.2)	35.0%
Reported crime	Cat. 1 (little)	9.5% (0.1)	9.6% (0.1)	9.6%
	Cat. 2	14.5% (0.2)	14.5% (0.2)	14.6%
	Cat. 3	25.1% (0.2)	25.1% (0.2)	24.9%
	Cat. 4	25.8% (0.2)	25.7% (0.2)	25.8%
	Cat. 5 (much)	25.1% (0.2)	25.2% (0.2)	25.1%
Urbanization	Cat. 1 (urban)	16.7% (0.1)	16.7% (0.2)	16.7%
	Cat. 2	22.3% (0.2)	22.3% (0.2)	22.3%
	Cat. 3	15.9% (0.2)	15.9% (0.2)	15.8%
	Cat. 4	22.8% (0.2)	22.8% (0.2)	22.4%
	Cat. 5 (rural)	22.3% (0.2)	22.3% (0.2)	22.7%

Table 5. Estimates and population totals of four known register variables. The ISM 2009 sample is used with the standard and calibrated weighting models. Standard errors are given in brackets.

5.4 Mode calibrated ISM

The mode calibration introduced in section 4.3 is now applied to the ISM. Since the oversampling is regionally clustered, the composition of data collection modes varies across police districts. It is therefore important to cross data collection mode with district. The four modes used in the ISM are aggregated to two types: the modes with interviewer (telephone and face-to-face) and those without interviewer (internet and paper). Interviewer presence is an important characterizing factor of response modes (see section 2). In addition, this is done to avoid extreme weights for the face-to-face respondents in districts where hardly any face-to-face interviews are held. Furthermore, internet and paper are the response modes offered initially, whereas telephone and face-to-face are the follow-up modes.

Table 6 lists the composition of the response according to modes with and without interviewer. The distributions for unweighted and weighted data are shown for the three years. Weighting the sample results in an increase in the share of the respondents interviewed in modes with interviewer. These are clearly underrepresented, in particular in the ISM-sample. While the data collection of the SN-sample has happened identically in each of the three years, the number of interviews collected by interviewers is higher in 2008 than in the other years. The larger number of interviews collected using modes without interviewer in the ISM-sample in 2009 is due to the large oversampling that took place that year (see section 5.1).

Not weighted		2008	2009	2010
ISM-sample	With interviewer	32.6%	18.7%	26.6%
	Without interviewer	67.4%	81.3%	73.4%
SN-sample	With interviewer	44.5%	36.6%	37.3%
	Without interviewer	55.5%	63.4%	62.7%
Weighted (standard model)				
ISM-sample	With interviewer	40.9%	25.6%	35.5%
	Without interviewer	59.1%	74.4%	64.5%
SN-sample	With interviewer	46.8%	37.4%	38.6%
	Without interviewer	53.2%	62.6%	61.4%

Table 6. Overview of response composition in ISM surveys 2008 - 2010, not weighted and weighted, for modes with and without interviewer.

Levels for the proposed mode calibration were chosen when only the 2008 and 2009 surveys were available. Since the SN-sample can be considered as the reference sample,

which crime statistics would be based on in the absence of oversampling, the calibration levels were chosen based on the distributions of the SN-sample in these years. The levels for the modes with and without interviewer were set to 40% and 60% respectively, crossed with police district. The distribution realized in 2010 in the weighted SN-sample is again close to the 40/60 composition.

This mode calibration is added to the standard weighting model, resulting in the calibrated survey outcomes listed in Table 7. This table is to be compared with the standard GREG results shown in Table 4. The effect of the calibration is largest when the realized response composition deviates most from the applied calibration levels. In the ISM, this is the case in 2009 (see Table 6). A comparison of standard and mode calibrated GREG estimates is shown in Figure 1. The variables concerning victimization and offences are affected less by the calibration than the three other variables. A possible explanation is that the former are more factual variables, having similar or even smaller measurement errors for the different modes. Nevertheless, year-to-year mutations of calibrated results are less extreme for these variables too. The calibration has a stabilizing effect. In addition, the differences between the results based on the calibrated ISM and SN-samples are smaller than those based on the standard weighted samples.

Variable	Sample	2008-2009	p	2009-2010	p
victim	ISM	+2.9%	0.031	-5.3%	<.001
	SN	-0.2%	0.456	-2.8%	0.074
offences	ISM	+5.5%	0.005	-10.4%	<.001
	SN	+1.4%	0.329	-5.8%	0.015
unsafe	ISM	+1.4%	0.192	+2.0%	0.088
	SN	+2.7%	0.112	+2.7%	0.083
funcpol	ISM	-1.8%	<.001	+0.3%	0.107
	SN	-1.0%	0.026	-0.1%	0.391
antisoc	ISM	+0.8%	0.269	+1.6%	0.107
	SN	+1.1%	0.282	+0.8%	0.314

Table 7. Annual mutations in key ISM variables using the mode calibrated GREG estimator.

The differences between the ISM and SN results for the variables relating to victimization and number of offences are only partly removed by the calibration. The remaining differences are difficult to explain. A potential but unverifiable reason is that the ISM-sample is still selective compared to the SN-sample, with respect to victimization. In such a scenario, the ISM-sample would contain more victims of crime, for reasons unrelated to response mode or other variables included in the weighting model. It can for example be anticipated that respondents are more likely to participate with a crime survey, if they have been victimized recently. The fact that the overall response rate achieved in the SN-

sample is higher than that in the local samples may mean that the local samples contain more victims of crime, regardless of the mode the data were collected in.

Lastly, the sensitivity of the ISM outcomes to the chosen calibration levels is analyzed. Table 8 contains the percentage changes in the ISM survey outcomes for different calibration proportions of the response mode. Only mutations between 2009 and 2010 are shown, but the pattern of mutations between 2008 and 2009 is similar. The column 40/60 contains the results presented before, in Table 7. The other columns show outcomes for mode compositions of 20/80, 30/70, and 50/50, with the first number referring to the share of responses obtained through modes with an interviewer present, and the second number to the share of those without interviewer. While the choice of the calibration levels does affect the outcomes, as expected, the conclusions based on this survey would not change when using different levels. From Table 7 it is seen that the decreases in victim and offences are significant when using the 40/60 levels, and that the increases in the other three variables are not. The same applies to the mutations found when using any of the other levels shown in Table 8. The choice of which levels to use is arbitrary, with the limitation that the same levels should be used in all editions of the surveys. The choice for the 40/60 level for the ISM is motivated above.

Variable	20/80	30/70	40/60	50/50
victim	-6.1%	-5.7%	-5.3%	-5.0%
offences	-11.4%	-10.9%	-10.4%	-10.1%
unsafe	+3.1%	+2.6%	+2.0%	+1.4%
funcpol	+0.3%	+0.3%	+0.3%	+0.3%
antisoc	+1.2%	+1.4%	+1.6%	+1.7%

Table 8. Sensitivity of ISM outcomes to choice of calibration levels. Annual mutations are shown for the key variables for the ISM-sample, for the years 2009-2010.

While not all issues are explained by the mode calibration, and while this calibration is based on strong assumptions that cannot be validated, it was decided that the mode calibrated results are preferable rather than the standard GREG estimates. Not calibrating the GREG estimates for mode results in year-to-year mutations in survey outcomes that are almost certainly, at least partially, due to differences in the composition of the response modes. The mode calibrated results have been officially published by SN.

6 Conclusion

This paper presents a method of response mode calibration for mixed-mode surveys where the composition of response modes is not fixed. The Dutch ISM is an example of such a survey, where different response modes are used sequentially. This survey has been executed three times, with large differences in the mode compositions between years. Ap-

plying the proposed mode calibration stabilizes the survey results, and renders the results based on the full ISM-sample comparable to those based on the more stable SN-sample. It is hereby assumed that the GREG weighting model removes selectivity with respect to the survey variables. This assumption cannot be truly validated, though analysis of some correlating background variables provides supporting evidence. A second assumption is that the underlying measurement error model is constant over time.

The proposed methodology is applicable generally, to surveys in which the composition of response modes can vary. Measurement errors related to response mode are calibrated to constant levels. The errors are not quantified, nor are they removed. Resulting estimates can still be biased. The calibration causes the bias due to measurement error to be constant. As a result, different editions of repeated surveys provide outcomes that are comparable, with mutations in the survey estimates that are no longer confounded with mutations in response mode composition. This is true only under the aforementioned assumptions. Since they are difficult to verify, the proposed method can be considered as a practical solution to improve stability of survey outcomes over time.

However, preventing instabilities by choosing stable survey designs is recommendable. Extreme variations in inclusion probabilities, differences in data collection strategies between agencies or regions, and varying achieved response percentages are to be avoided.

References

Aquilino, W. (1994). Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly* 58, 210–240.

Aquilino, W. and LoSciuto, L. (1990). Effect of interview mode on self-reported drug use. *Public Opinion Quarterly* 54, 362–395.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of statistical inference*, pp. 203–242. Toronto: Holt, Rinehart and Winston.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4, 251–260.

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

Buelens, B. and Van den Brakel, J. (2010). On the necessity to include personal interviewing in mixed-mode surveys. *Survey Practice* 4.

Channel, C., Miller, P., and Oksenberg, L. (1981). *Research on interviewing techniques.* in S. Leinhardt (Ed.) Sociological Methodology, Jossey-Bass, San Fransisco, pp. 389-437.

Christian, L., Dillman, D., and Smyth, J. (2008). *The effects of mode and format on answers to scalar questions in telephone and web surveys.* in J. Lepkowski, C. Tucker,

M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, M. Link, R. Sangster (Eds.) Advances in Telephone Survey Methodology, John Wiley, New York, pp. 250-275.

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

Dillman, D. (2007). *Mail and internet surveys: The tailored design method.* John Wiley, New York.

Dillman, D. and Christian, L. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods* 17, 30–52.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response and the internet. *Social Science Research* 39, 1–18.

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

Gilljam, M. and Granberg, D. (1993). Should we take don't know for an answer? *Public Opinion Quarterly* 57, 348–357.

Hochstim, J. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association* 62, 976–989.

Holbrook, A., Green, M., and Krosnick, J. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly* 67, 79–125.

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collecetion mode on measurement. *International Statistical Review* 78, 3–20.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 213–236.

Krosnick, J. and Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly* 51, 201–219.

Krysan, M., Schuman, H., Scott, L., and Beatty, P. (1994). Response rates and response content in mail versus face to face surveys. *Public Opinion Quarterly* 58, 381–399.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* John Wiley, New York.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Stern, M., Dillman, D., and Smyth, J. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Journal for Survey Research Methods* 1, 121–138.

Toepoel, V., Das, M., and Van Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of Official Statistics* 25, 509–528.

Tourangeau, R., Couper, M., and Conrad, F. (2004). Spacing, position and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly* 68, 368–393.

Tourangeau, R., Couper, M., and Conrad, F. (2007). Color, labels and interpretive heursitics for response scales. *Public Opinion Quarterly* 71, 91–112.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge.

Tourangeau, R. and Smith, T. (1996). Asking snesitive questions: The impact of data collection, question format, and question context. *Public Opinion Quarterly* 60, 275–304.

Van den Brakel, J. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Forcs Survey. *Journal of the Royal Statistical Society, Series A* 171, 581–613.

Van den Brakel, J. and Roels, J. (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Annals of Applied Statistics* 4, 1105–1138.

Van den Brakel, J., Smith, P. A., and Compton, S. (2008). Quality procedures for survey transitions: experiments, time series and discontinuities. *Journal for Survey Research Methods* 2, 123–141.

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly* 74, 1027–1045.

Voogt, R. and Saris, W. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of official statistics* 21, 367–387.

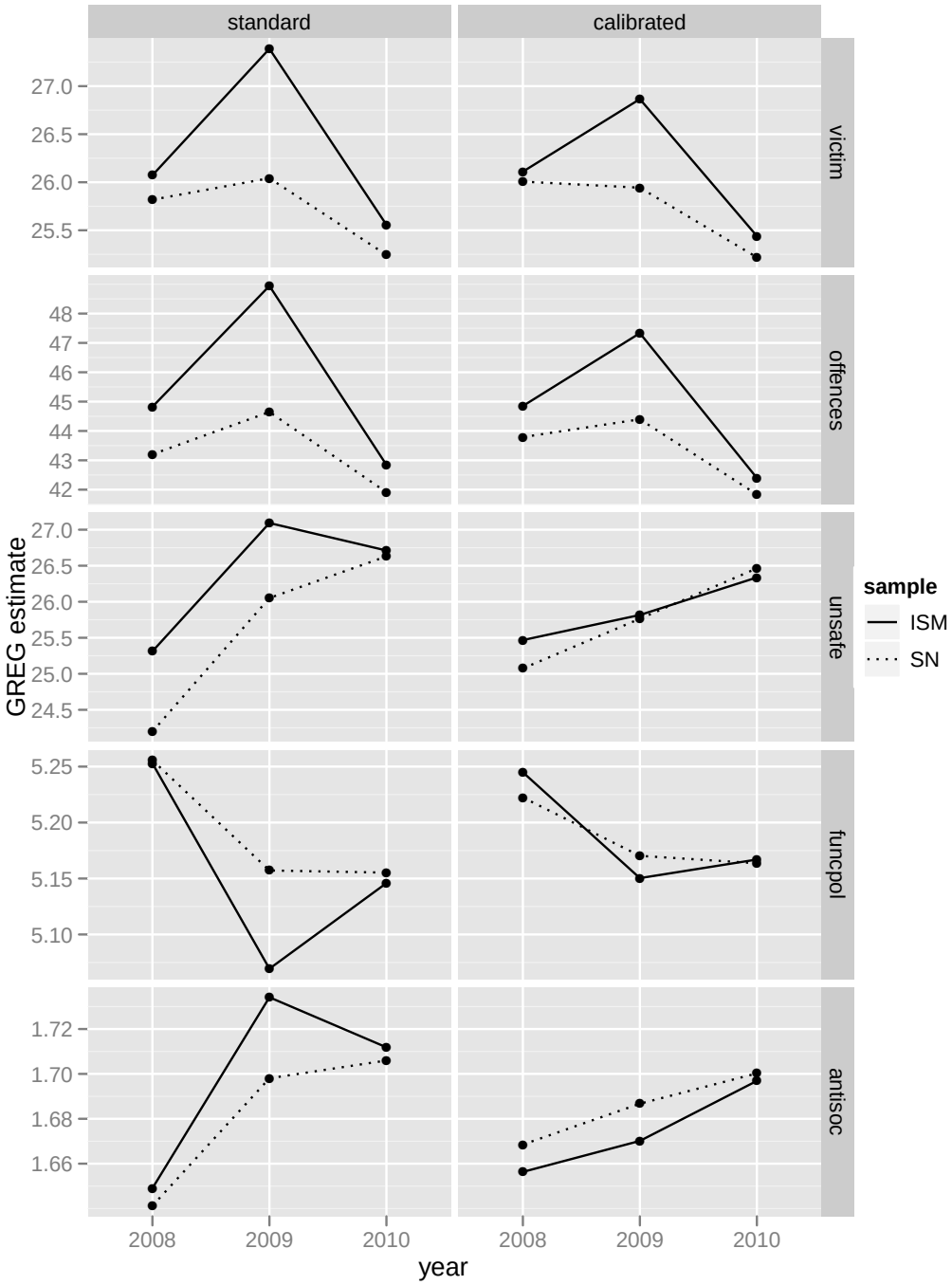

Figure 1. Standard (left) and mode calibrated (right) GREG estimates for the 5 key ISM variables.