Measurement error calibration in mixed-mode sample surveys

13

Bart Buelens and Jan van den Brakel

Discussion paper (201304)

Statistics Netherlands

The Hague/Heerlen, 2013

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2012–2013	2012 to 2013 inclusive
2012/2013	average for 2012 up to and including 2013
2012/'13	crop year, financial year, school year etc. beginning in 2012 and ending in 2013
2010/'11– 2012/'13	crop year, financial year, etc. 2010/'11 to 2012/'13 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

60083201304 X-10

Measurement error calibration in mixed-mode sample surveys

Bart Buelens and Jan van den Brakel

Mixed-mode surveys are known to be susceptible to mode-dependent selection and measurement effects, collectively referred to as mode effects. The use of different data collection modes within the same survey may reduce selectivity of the overall response but is characterized by measurement errors differing across modes. Inference in sample surveys generally proceeds by correcting for selectivity – for example by applying calibration estimators – 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 it is likely that the mode composition of the overall response differs between subsequent editions of the survey, leading to variations in the total measurement error and invalidating classical inferences. An approach to inference in these circumstances, which is based on calibrating the mode composition of the respondents towards fixed levels, is proposed. Assumptions and risks are discussed and explored in a simulation and applied to the Dutch crime victimization survey.

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

1 Introduction

Mixing modes of data collection in survey sampling is increasingly attractive. Driving factors are the pressure to reduce administration costs, attempts to reduce non-sampling errors, and technological developments leading to new data collection procedures. Web interviewing, for example, combines the benefits of traditional self-administered data collection through paper with the power of computer assisted administration. It is cost effective, provides a greater sense of privacy for the respondent, and offers the possibility of a more complex questionnaire design (Fricker et al., 2005). Despite these advantages, it is recognized that measurement errors compromise data comparability when mixing data collection modes, see e.g. De Leeuw (2005), Voogt and Saris (2005), Jäckle et al. (2010), and Vannieuwenhuyze et al. (2010).

National Statistical Institutes (NSIs) produce official statistics which are often based on sample surveys repeated at regular intervals. A problem with sequential mixed-mode data collection is that the distribution of the respondents over the different data collection modes will generally not be constant at consecutive editions of a repeatedly conducted survey. This may cause effects associated with these modes, such as measurement bias, to vary over time. For example, a gradual increase in availability of internet access or decrease in the proportion of the population for which a telephone number is available will result in a gradual change of mode-dependent measurement errors in the survey estimates. Time series based on repeatedly conducted surveys employing a mixed-mode design will therefore reflect a more severely biased estimate of changes over time of the variables of interest compared to uni-mode surveys.

In the literature a lot of empirical research aimed at the quantification of different mode effects is found, see e.g. De Leeuw (2005). There is, on the other hand, hardly literature about how to handle fluctuations in measurement bias between subsequent editions of repeated surveys. This is an important limitation of the application of sequential mixed-mode designs in official statistics, since it compromises the comparability of the outcomes between subsequent editions. In addition, cross-sectional comparisons within each edition of a survey may be problematic when the mode composition of the response differs between subpopulations. The focus in this paper is on identifying an appropriate estimation procedure that is robust to variations in the distribution of respondents over the different data collection modes.

At NSIs the general regression (GREG) estimator (Särndal et al., 1992) is widely applied for inference in survey sampling. Under the traditional design-based and model-assisted approach, this estimator is used to correct for unequal selection probabilities of the sample design and selection bias due to nonresponse. In this paper the GREG estimator is used to stabilize the measurement bias between the subsequent editions and the publication domains of a repeated survey by calibrating the response to fixed distributions over the data-collection modes. The use of this predominantly design-based approach is motivated with a measurement error model for the observations obtained in the sample.

The immediate reason for developing 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). The paper starts with a review of mode effects. Next, a calibration method is developed that is robust against variations in the mode composition of the respondents. The properties of this method are studied in a simulation study and illustrated in the ISM context. The paper concludes with a discussion.

2 Mode effects in mixed-mode surveys

The data collection mode affects multiple sources of non-sampling error in the data collection phase of a survey. The mode determines which part of the target population is covered. If not all sample units have access to the mode, then the mode is said to suffer from under coverage. The response rate of the approached sample units depends on the mode, an effect known as mode dependent non-response behavior. The coverage and non-response effects are collectively referred to as the selection effect. In addition, the mode 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 defined as the observed differences in outcomes between identical surveys administered through different modes, and are the combined result of selection and measurement effects.

A possible classification of mixed-mode strategies 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 typically start with the self-administered modes that have low administration costs and use interviewer-administered modes to re-approach the remaining nonrespondents. 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).

Many experiments indicating that different data collection modes have substantial effects on the answers generated by respondents are found in the literature; see De Leeuw (2005) for an overview. Such 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 judgments 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 differences in response between different data collection modes. Several studies indicate that respondents are more likely to offer socially desirable answers and demonstrate acquiescence in the presence of an interviewer than in self-administered surveys

(Dillman et al., 2009; Holbrook et al., 2003). It may also be expected that satisficing (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 telephone interviews, as the speed in such interviews is generally higher compared to face-to-face interviews (Holbrook et al., 2003). 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; 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 associated with differences between visual and aural data collection modes are primacy and recency effects (Krosnick and Alwin, 1987).

The interaction between mode and questionnaire design is discussed by Dillman and Christian (2005) and De Leeuw (2005). 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, 2007). 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 non-sampling 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 increase the mean squared error of the survey estimates. 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. 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 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 face to face and telephone modes.

Although a massive body of literature is available on quantifying and explaining mode effects, no references are available that propose estimation procedures that account for the fluctuations of measurement bias in sequential mixed-mode designs. In this paper a

method is proposed that attempts to keep the measurement bias constant over time, with the purpose to obtain unbiased estimates for changes over time of survey variables.

3 Methods

3.1 General regression estimator and measurement error

At NSIs, estimation in person and household sample surveys is commonly conducted using the Horvitz-Thompson (HT) estimator or the general regression (GREG) estimator (Särndal et al., 1992). The HT estimator for the total of a survey variable y is expressed as the sum over the weighted observations in the sample:

$$\hat{t}_y^{HT} = \sum_{k=1}^{n} d_k y_k, \tag{1}$$

with y_k the measure of the target variable y for sample unit k, n the sample size, and d_k the design weight for unit k. The design weight is obtained as the inverse of the probability that the corresponding sampling unit is included in the sample, and accounts for unequal selection probabilities in the sample design. The HT estimator is a design-unbiased estimator for the unknown population total t_y. The GREG estimator improves the precision of the HT estimator by taking advantage of auxiliary information for which the population totals are known exactly from e.g. registrations and can be expressed as:

$$\hat{t}_y = \hat{t}_y^{HT} + \hat{\beta}'(t_x - \hat{t}_x^{HT}). \tag{2}$$

Here $\hat{t}_x^{HT}$ denotes the HT estimator for the known population totals of the auxiliary variables t_x. An expression for $\hat{t}_x^{HT}$ is defined similarly as (1); $\hat{\beta}$ denotes the generalized least squares estimate for the regression coefficients of the linear model

$$y_k = \beta' x_k + e_k, \tag{3}$$

with x_k the auxiliary variables of unit k and e_k a residual. The GREG estimator is an approximately design-unbiased estimator for the population total of y. The second term in (2) is a correction on $\hat{t}_y^{HT}$ that is based on the discrepancy between the HT estimator for the auxiliary variables and the known population values. The correction accounts for skewness of the sample and selectivity due to nonresponse (Särndal and Lundström, 2005; Bethlehem, 1988). Expression (2) can also be written as a weighted sum of the sample observations

$$\hat{t}_y = \sum_{k=1}^{n} w_k y_k \tag{4}$$

with w_k the regression weight for unit k and can be interpreted as a calibration of the design weights such that

$$\hat{t}_x = \sum_{k=1}^{n} w_k x_k = t_x. \tag{5}$$

The regression weights incorporate both the sampling design and the calibration correction using auxiliary information. The weights w_k only involve inclusion probabilities and auxiliary information and do not depend on the target variables. As a result, only one set of weights is required for all survey variables, which is an attractive property of the GREG estimator in multipurpose surveys. See Särndal et al. (1992) for a derivation of this estimator, expressions for $\hat{\beta}$ and w_k, and details about variance estimation.

The use of auxiliary information through the GREG estimator can decrease the design variance of the HT estimator and can correct for selection effects with respect to the survey variables. The better the linear model (3) explains the variation of the target variable in the population as well as the selectivity in the response, the larger the increase in accuracy obtained with the use of auxiliary information through the GREG estimator. Selectivity of response in sample surveys is to some extent determined by the mode of data collection. This is the case in uni-mode surveys as well as in mixed-mode surveys. Both types of surveys are characterized by different selection effects. As long as auxiliary variables that explain any form of selectivity are included, the GREG estimator is applicable in both uni-mode as well as in mixed-mode surveys. Age might for example explain mode related selection effects, since younger groups may be over-represented in web modes and under-represented in telephone administered modes. If age is also correlated with the target variable, then age is a potential auxiliary variable to be included in the weighting model to remove mode dependent selection effects.

GREG estimators remove mode dependent selection effects in the target variables, at least partially. They, however, do not correct for measurement error. In classical sampling theory, the GREG estimator is considered a design-based or model-assisted estimator. Under this approach it is assumed that the observations y_k obtained from the respondents are true fixed values observed without error. This is not tenable when investigating measurement error. In surveys where GREG estimation is used, measurement error is usually ignored. In mixed-mode data collection regimes, measurement errors may differ between modes in the same survey. To investigate how measurement errors affect GREG estimators, a measurement error model explaining the systematic bias due to different modes is required. The following measurement error model is assumed:

$$y_{k,m} = u_k + b_m + \varepsilon_{k,m} \tag{6}$$

with $y_{k,m}$ the observations using mode m of the true intrinsic values u_k, for units $k = 1, ..., n$, b_m the systematic effect of mode m and $\varepsilon_{k,m}$ random error components, with expected values equal to zero. This model suggests that the systematic errors b_m are constant for all units k observed through mode m. However, if the b_m are dependent on a known categorical variable x for a given mode, b can 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 $1, ..., p$. In the remainder of this paper the subscript m is referred to as the mode or mode class, with the understanding that an appropriate cross classification is intended if necessary.

Substitution of the measurement error model in the expression of the GREG estimator and taking the expectation over the measurement error model gives

$$\hat{t}_y = \hat{t}_u + \sum_{m=1}^{p} b_m \hat{t}_m. \tag{7}$$

with $\hat{t}_u = \sum_{k=1}^{n} w_k u_k$, $\hat{t}_m = \sum_{k=1}^{n} w_k \delta_{k,m}$, and $\delta_{k,m}$ a dummy indicator equal to one if respondent k completes the questionnaire through mode m and zero otherwise. Under the measurement error model, the GREG estimator of the measured variable y is equal to the sum of the GREG estimator of the population total of the true values u_k, $\hat{t}_u$, and a measurement bias $\sum_{m=1}^{p} b_m \hat{t}_m$. Note that $\hat{t}_m$ can be interpreted as the GREG estimator of the total number of units in the population that respond through mode m. As a result, the measurement bias is equal to the sum over the estimated population totals of units responding through modes m, $\hat{t}_m$, multiplied with the systematic effect of each mode, b_m.

3.2 Mode calibration

Interest in repeated sample survey outcomes is largely focused on changes over time. Under a sequential mixed-mode design, the distribution of the respondents over the data collection modes will vary between the different editions of a repeated survey. As a result, the $\hat{t}_m$ vary between subsequent editions of a survey, causing variations in the total measurement error of the survey outcomes, which are confounded with true changes over time of the underlying variables.

Assume a mixed-mode survey design, executed twice. The difference between the survey outcomes is ideally unaffected by the differences in measurement bias, i.e. $\hat{t}_y^{(2)} - \hat{t}_y^{(1)} = \hat{t}_u^{(2)} - \hat{t}_u^{(1)}$, where the superscripts (1) and (2) indicate the first and second edition of the survey. For this equality to hold, it is seen from equation (7) that this requires $\hat{t}_m^{(1)} = \hat{t}_m^{(2)}$ for $m = 1, ..., p$, under the assumption that b_m does not change over time. The requirement that $\hat{t}_m^{(1)} = \hat{t}_m^{(2)}$ is not fulfilled when the weighted number of respondents using particular modes varies between editions of the survey, which is usually the case in sequential mixed-mode surveys.

To avoid confounding of true changes over time with varying mode compositions, the mode calibrated GREG estimator is proposed. This is a GREG estimator which additionally calibrates the distribution of the respondents in the sample over response modes to fixed levels. The GREG estimator formulated in expression (2) is therefore extended as

$$\hat{t}_y^c = \hat{t}_y^{HT} + \hat{\beta}'(t_x - \hat{t}_x^{HT}) + \hat{\beta}'_{\mathrm{M}}(\Upsilon_{\mathrm{M}} - \hat{t}_{\mathrm{M}}^{HT}) \tag{8}$$

with $\hat{\beta}_{\mathrm{M}} = \{\hat{\beta}_m\}_{m=1...p}$ a vector of additional regression coefficients related to mode. The $\hat{t}_{\mathrm{M}}^{HT}$ and Υ_{M} are respectively the HT estimators and the calibration levels for the total number of people in the population that respond through each of the modes. The $\Upsilon_{\mathrm{M}} = \{\Upsilon_m\}_{m=1...p}$ formally play the role of population totals, but are arbitrarily chosen levels as they do not correspond to some a priori known true population characteristics. Through inclusion of the additional term, this estimator achieves that the calibrated total

for each mode m is equal to the corresponding calibration level: $\hat{t}_m^c = \Upsilon_m$, as follows from restriction (5). This offers control over the response mode composition of the weighted sample. Similar to the classic GREG estimator, weights can be derived for the mode calibrated version. Through the mechanism of the GREG estimator, this estimator simultaneously achieves calibration for all x variables to correct for selectivity, and calibration of the response mode levels to stabilize the measurement error. Expressing (8) in terms of the error model gives the mode calibrated equivalent of equation (7),

$$\hat{t}_y^c = \hat{t}_u^c + \sum_{m=1}^{p} b_m \hat{t}_m^c = \hat{t}_u^c + \sum_{m=1}^{p} b_m \Upsilon_m. \tag{9}$$

If the measurement errors b_m do not vary between the different editions of a survey and if the size of the population remains constant, the measurement bias $\sum_{m=1}^{p} b_m \Upsilon_m$ in the mode calibrated GREG estimator is constant. It cancels out in the difference between the estimates of two consecutive survey editions, despite variations in mode compositions between the editions. The difference of the estimated totals is now an approximately unbiased estimator of the change in the totals of the underlying variable u. This is a desirable outcome, as changes over time are not affected by changing measurement bias. It is important to note that the total measurement bias as well as the measurement errors b_m have not been quantified, nor has the bias been removed from the level of the survey estimates.

The assumption that the b_m do not vary over time is made explicit through the adoption of the measurement error model. In uni-mode surveys where the classic GREG estimator is used and measurement errors are ignored, the assumption of constant measurement error is made too, albeit implicit. Changing measurement errors in repeated uni-mode surveys would invalidate any inference about change over time of survey variables. Because in such settings measurement errors are not accounted for, confounding of varying measurement errors with true changes over time remains invisible. While in the present setting this assumption is made explicit, it is no stronger than what is typically assumed in survey sampling.

The proposed mode calibration does not only improve comparability of outcomes of subsequent editions of a repeated survey, but it can also improve comparability between publication domains. Sample survey estimates for target variables are usually not only produced at a national level but also for different subpopulations or domains, and are defined by socio-demographic background variables. Examples include regions, gender and age classifications. The distributions of the respondents over the modes within such domains can differ between domains when using sequential mixed mode designs. The proposed mode calibration can improve the comparability between the publication domains through applying the mode calibration at the domain level. This is achieved by including the interaction of mode and domain in the model.

3.3 Conditions and assumptions

Mode calibration is intended to affect the bias term in expression (7), and to leave the estimate $\hat{t}_u$ unchanged. Using the mode calibrated GREG estimator, change over time is estimated as $\hat{t}_y^{c(2)} - \hat{t}_y^{c(1)}$. This is an approximately unbiased estimator for $t_u^{(2)} - t_u^{(1)}$ only if $\hat{t}_u^c = \hat{t}_u$, with $\hat{t}_u^c$ the mode calibrated GREG estimator for the total of u. This condition is fulfilled if the additional weighting term in the GREG estimator that concerns response mode, does not explain any mode dependent selectivity with respect to u beyond that explained by the other auxiliary variables x. Alternatively this condition expresses that the additional calibration to fixed mode levels should have no effect in the absence of measurement errors. Since only $\hat{t}_y$ and $\hat{t}_y^c$ can be obtained, it is very difficult to test fulfilment of this condition. Hence, this condition must be assumed fulfilled in order to appropriately use the mode calibrated GREG estimator. Applying mode calibration in settings where the assumption does not hold may have undesirable effects. In particular, varying mode compositions in this case may reflect true variations in the underlying target variable u. Mode calibration will suppress true differences and hence introduce selection bias.

One possibility for checking this assumption is to seek additional variables z, that are available from a register and therefore 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 $\hat{t}_z$ is not significantly different from t_z. Since z is a register variable, it has no associated measurement error. Hence $\hat{t}_z^c$ should not differ significantly from $\hat{t}_z$. Substantial differences between $\hat{t}_z$ and $\hat{t}_z^c$ are an indication that, conditionally on x, the additional term for the mode calibration explains mode dependent selection bias. If not all three t_z, $\hat{t}_z$ and $\hat{t}_z^c$ are approximately equal, the covariates x are not sufficient in removing mode dependent selectivity with respect to z. A straightforward action in this case is to include z in the weighting model. Checking many variables in this way builds confidence that the weighting model removes all selectivity.

In addition, the sensitivity to the choice of calibration levels Υ_m can be analyzed. 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. In the ideal case, the differences between subsequent survey estimates remain unaffected under varying calibration levels. This is a strong indication that the assumption holds. Then the effect of mode calibration is the leveling out of measurement bias, as intended.

3.4 Choosing appropriate calibration levels

The aim of the calibration is to neutralize changes in the total measurement error. Therefore, it is not critical what the precise levels for the modes are set to in a particular survey. It is important to chose the calibration levels such that they do not incur negative weights

or large fluctuations in the dispersion of the weights, since this will inflate the variance of the GREG estimates. Intuitively, this is accomplished by choosing the calibration levels Υ_m as close as possible to the observed $\hat{t}_m$. If a survey has taken place more than once, the observed levels of $\hat{t}_m$ can be used as a guide. The expected trends in these levels can be taken into account when choosing the levels Υ_m. It can for example be anticipated that in future years the proportion of web interviewing will increase at the cost of telephone and face-to-face interviewing.

If a survey is conducted for the first time, then the Υ_m can be taken close to the observed levels of $\hat{t}_m$. This is a safe approach since it will not affect the GREG estimates of the survey variables. The levels of Υ_m can eventually be adjusted when information from subsequent editions of the survey becomes available. Results observed in pilot studies or related surveys employing a similar sequential mixed-mode strategy, and expectations about how the proportions of the different modes may develop in the future can also be taken into account in the choice of Υ_m.

It is necessary to decide with which categorical variables the modes need to be crossed, keeping in mind that the number of calibration levels should be as parsimonious as possible. A practical approach is to consider classification variables (i) where the distributions of the respondents over the modes differ substantially, and (ii) which are used in the main output tables of the survey. As explained in subsection 3.2, this avoids that differences between the survey variables in the output tables are partly due to differences in measurement errors.

As time proceeds, the differences between Υ_m and the actual $\hat{t}_m$ can become too large so that it will be inevitable to adjust the levels of Υ_m to the actual situation. With such an adjustment, a discontinuity in the observed time series is introduced, which can be quantified and adjusted by recalculating the survey estimates with the revised weighting model.

3.5 Quantifying and correcting for relative mode effects

The mode effects b_m in measurement error model (6) denote the systematic deviations that occur if mode m is used to measure the true intrinsic values u. It is very difficult, or even impossible to quantify these mode effects as long as no a priori information about the values of u is available, e.g. through a register. The generalized least square (GLS) estimates for the regression coefficients $\hat{\beta}_m$ from the linear model underlying the mode calibrated GREG estimator, however, provide quantitative information about systematic differences between the modes applied to administer the survey. The auxiliary variable that specifies the mode used to complete the questionnaire, possibly crossed with other relevant background variables, is a categorical variable. To identify the model, one of the p modes must be used as the reference category. It is understood that more reference categories are required if mode is crossed with other known categorical variables. Under the aforementioned assumption that the other auxiliary variables in the linear model

explain mode-dependent selectivity, the GLS estimates for the regression coefficients of the remaining $p-1$ modes can be interpreted as the systematic differences between the measurement errors of a mode and the mode of the reference group. These differences are referred to as the relative mode effects.

After having estimated the relative mode effects, it becomes possible to adjust the mode calibrated GREG estimates, such that systematic differences in measurement errors with respect to the mode that is used as the reference mode, are removed from the estimates. If e.g. face-to-face interviewing is used as the benchmark and corresponds to $m = 1$, then the mode calibrated GREG estimator adjusted for measurement bias of the modes with respect to face-to-face interviewing could be defined as

$$\tilde{t}_y^c = \hat{t}_y^c - \sum_{m=2}^{p} \hat{\beta}_m' \Upsilon_m. \tag{10}$$

For estimates of domain totals not distinguished in the classification of m, the estimator would be defined as

$$\tilde{t}_{y,d}^c = \hat{t}_{y,d}^c - \sum_{m=2}^{p} \hat{\beta}_m' \hat{t}_{m,d}^c, \tag{11}$$

with

$$\hat{t}_{y,d}^c = \sum_{k=1}^{n} w_k^c y_k \delta_{k,d}, \qquad \hat{t}_{m,d}^c = \sum_{k=1}^{n} w_k^c \delta_{k,m} \delta_{k,d}, \tag{12}$$

and $\delta_{k,d}$ an indicator that equals one if sampling unit k belongs to subpopulation d and zero otherwise. The mode adjusted GREG estimator (10) or (11) is equal to the mode calibrated GREG estimator (8) were the original observations y_k are replaced by mode-corrected imputations $\tilde{y}_k = y_k - \hat{\beta}_m$. It follows indeed from (4) that $\hat{t}_y^c$ equals $\tilde{t}_y^c$, if y_k is replaced by $\tilde{y}_k$ and using mode calibrated weights.

The mode calibrated GREG estimator attempts to stabilize the total measurement bias in the GREG estimates in the subsequent editions of a survey without adjusting the observations obtained in the sample. The mode adjusted GREG estimator $\tilde{t}_y^c$ goes beyond this goal by applying a synthetic adjustment of the mode calibrated GREG $\hat{t}_y^c$ with the purpose to obtain an estimate as if the survey is administered through a uni-mode design using the mode that serves as the reference mode. This approach has several drawbacks. The synthetic adjustment clearly has an additional risk of introducing bias in the estimates if the linear model underlying the mode calibrated GREG estimator is misspecified. This risk could be partially reduced by assigning part of the sampling units randomly over the modes. With such built-in experiments, estimates for relative measurement bias could be obtained that rely less heavily on the assumption that selection bias is removed conditionally on other auxiliary variables. An other issue is that there is no clear choice for a preferred reference mode since each mode potentially suffers from measurement bias while the true underlying values remain invisible. Finally the mode adjusted GREG estimator (10) or (11) is dependent on the target variable. It is not possible to derive one set of

weights that can be used for the estimation of all target variables. As the primary goal in the present paper is the stabilization of the total measurement bias, the adjusted GREG estimator is not investigated further in this paper.

4 Simulation

The proposed mode calibration is based on assumptions that cannot be tested directly. Therefore the robustness of the proposed mode calibration is evaluated in a simulation. From an artificially constructed population, samples are drawn which suffer from mode dependent nonresponse bias and mode dependent measurement bias. Both sources of bias depend on the same auxiliary variable. As a result the selection bias and the measurement bias are correlated, which is generally the case in real life applications. It is investigated to which extent the proposed mode calibration succeeds in stabilizing measurement bias in the situation where the assumption is met that the auxiliary variable explains selective nonresponse. It is also tested how the mode calibration performs if this estimator fails to correct for selection bias, i.e. the auxiliary variable explaining selective nonresponse is not used in the GREG estimator. Finally it is tested how the mode calibrated GREG estimator performs if there is only mode dependent selection bias but no measurement bias.

A population of size $N = 100,000$ is created, with a single known binary background variable x taking on the values 1 and 2, with associated subpopulations of equal size. A continuous target variable u is assumed, normally distributed with standard deviation 3 and means 20 and 30 for units of class $x = 1$ and $x = 2$ respectively. The population mean of u is 25. One sample of size $10,000$ is drawn from this population using simple random sampling. This sample is observed using a mix of two modes, $m = 1$ and $m = 2$. Mode 2 is assumed to have a measurement error of 5 relative to mode 1. Denoting with y the observed value of u, $y(m = 1) = u$ and $y(m = 2) = u + 5$. In order to create confounding of mode dependent measurement bias with selection effects, unequal response probabilities are assumed for units of different classes. Units of class $x = 1$ have a response probability of 0.8 when approached in mode 1, and 0.2 for mode 2. Class $x = 2$ units have response probabilities of 0.4 and 0.6 respectively for modes 1 and 2. The mean response probabilities for each mode are 0.5 when applied to a sample with equal proportions of units of classes $x = 1$ and $x = 2$. The expected number of responses is therefore $5,000$.

In the simulation, the composition of the modes used to observe the sample is varied. This can be thought of as the survey agency varying the data collection modes over time, or between regions. In the first run, 10% of the sample is observed using mode 1 and 90% using mode 2. The fraction of the sample observed through mode 1 is increased in steps of 5%, up to 90% in the 17th run. The fraction observed through mode 2 is decreased accordingly.

This establishes a situation where the mode composition that is used suffers from selection effects and from measurement effects; the more mode 2 is used the more respondents

will be of class $x = 2$ and the more units are observed with measurement error. As a result, selection and measurement effects are confounded as units of class 2 are more likely to respond through the mode suffering from measurement error. The varying proportions of the modes cause varying proportions of x in the sample, and varying proportions of units that are measured with error.

For each of the combinations of the two modes, four estimates of the mean value of y are obtained:

1. The sample mean, which is obtained from the HT etimator under simple random sampling assuming equal selection probabilities. This estimator ignores selective nonresponse and measurement error. Differences between estimates obtained with the simulations are the result of selective nonresponse and fluctuations in measurement bias.

2. The GREG-estimate based on the covariate x. This estimator corrects for selective nonresponse, but ignores the measurement bias. Differences between estimates obtained with different mode compositions are the result of fluctuations in measurement bias.

3. The GREG-estimate based on the covariate x additionally calibrated to a fixed mode proportion, which is taken to be 50% (mode 1) / 50% (mode2). This is the situation were the assumption underlying the mode calibration is met, i.e. the other auxiliary variables of the GREG estimator explain the selection bias in the response. In this situation the estimator should stabilize the measurement bias, resulting in more or less equal estimates for the different simulation runs.

4. The mode-calibrated GREG estimator not using x as an auxiliary variable. This is in fact a mode-calibrated sample mean. This estimator is included to investigate the situation in which no covariate explaining selectivity is available, but mode-calibration is applied nevertheless. This tests the performance of the the mode-calibrated GREG estimator in the situation were the assumption that the estimator accounts for selection bias is not met.

While the theory developed in the previous section applies to totals, here means are considered. This is equivalent since the population size is known and not estimated. Figure 1 (top panel) shows the four estimates for each of the 17 simulations. The share of the sample that is observed through mode 1 increases from left to right.

The sample mean is affected most, showing large differences for the different mode compositions. The GREG-estimate correcting for x reduces this effect, by removing selection bias but fails to remove differences in measurement bias.

The mode calibration implemented here uses equal fractions of modes 1 and 2, and the population consists of equally many class $x = 1$ and $x = 2$ units. Therefore the GREG estimate obtained in that simulation where the fractions of modes are equal, can be seen

as the reference estimate of the mean of y. Knowing the relative measurement error of the modes, it is easily obtained that the expected value of this GREG estimate for y, the measured value of u, is 27.5.

The mode-calibrated GREG succeeds in rendering the estimates stable, producing estimates that are unaffected by the mode composition. The level of the estimates corresponds to that of the standard GREG obtained in the simulation with equal mode composition, with expected value 27.5. The mode-calibrated sample mean also results in rather stable estimates, but it is biased compared to the appropriately mode calibrated GREG estimate since it fails to remove selectivity with respect to x.

Next, a similar analysis is conducted for the case where u is measured without error. The results are shown in the bottom panel of Figure 1. The sample mean is again sensitive to the mode composition, but the GREG-estimate no longer is, as selectivity is corrected for appropriately and measurement error does not play a role. Calibrating the GREG-estimate to mode has no effect, and hence is harmless in the absence of measurement bias. Calibrating to mode without having the covariate x again leads to stable yet biased estimates. In this case the expected value is 25, the mean of u measured without error.

This simulation shows that the proposed mode calibration stabilizes the measurement bias in survey estimates in samples with large fluctuations in the distribution of respondents over different modes. The method is rather robust against violation of the assumption that the GREG estimator must correct for selection bias. It also appears that the application of mode calibration is harmlessness when applied in situations without measurement error. The availability of covariates explaining the selectivity is important, as without such variables mode calibration introduces bias in the levels.

5 Application

5.1 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. The ISM is based on a stratified two stage sample design of persons aged 15 years or older residing in the Netherlands. The country is divided into 25 police districts, which are used as the stratification variable in the sample design. Municipalities are used as primary sampling units in the first stage.

Statistics Netherlands (SN) conducts a national sample of the ISM with a size of about 19,000 respondents, which is equally divided over the strata, and is called the SN-sample. 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 local estimates. The combination of the SN-sample and these local samples is referred to as the ISM-sample.

All persons included in the sample receive an advance letter where they are asked

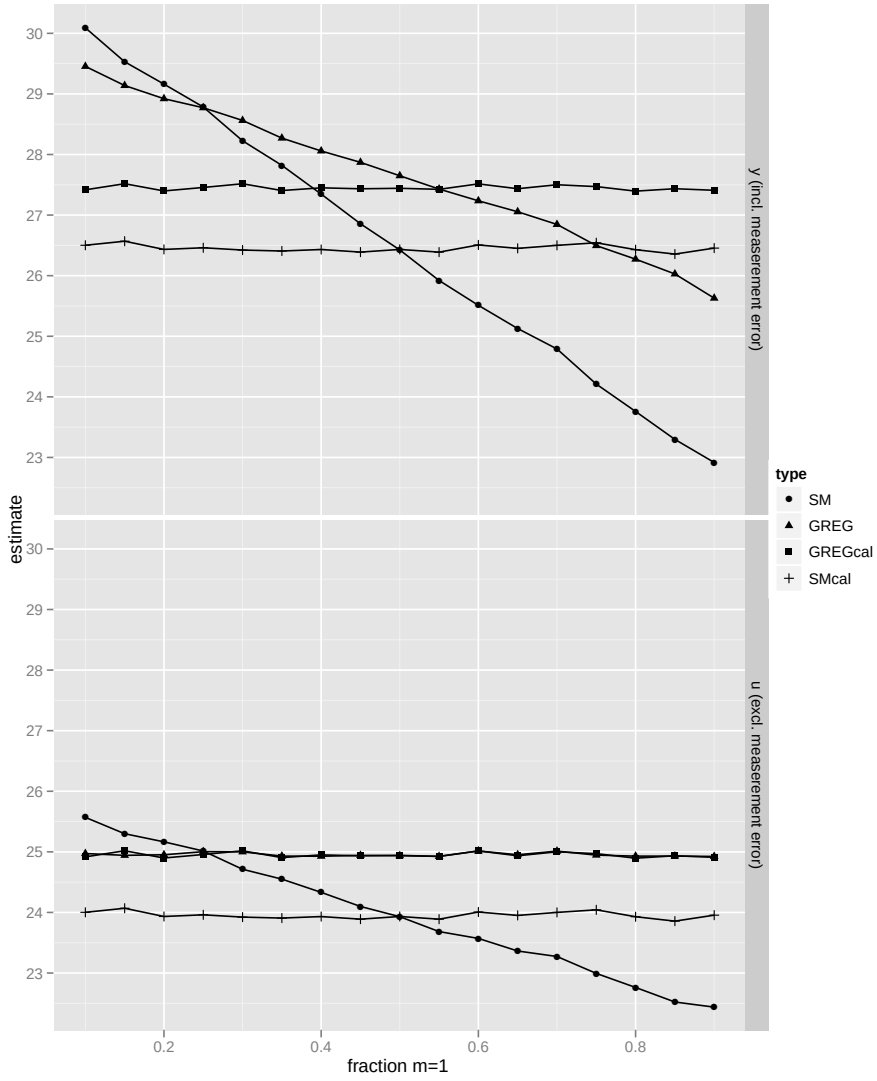

Figure 1. The sample mean (SM), the GREG-estimate based on the covariate x (GREG), this GREG-estimate additionally mode calibrated (GREGcal), and the mode-calibrated sample mean (SMcal) for 17 simulations, with the fraction of mode $m = 1$ responses varying from 0.10 to 0.90. The top panel shows y, affected by mode dependent measurement error. The bottom panel shows u, the case without measurement error.

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). For the data collection of the additional regional samples the WI, PAPI and CATI modes are mandatory. The use of the CAPI mode is recommended but not mandatory since this mode is very costly.

Table 1 gives an overview of the oversampling and the number of respondents for the

years 2008, 2009, 2010 and 2011. The extent to which local authorities participate in the oversampling affects the mode composition of the ISM-sample and results in strong fluctuations of the distribution of the response over the different data collection modes between the successive editions of the ISM, see Table 2. There are strong indications that this causes unstable results for estimates of levels as well as changes over time.

	2008	2009	2010	2011
Number of oversampled municipalities	77	239	21	225
Size response SN-sample	16,964	19,202	19,238	20,325
Size response local sample	45,839	182,012	19,982	203,621
Percentage of population in oversampled areas	29%	65%	16%	66%

Table 1. Overview of oversampling in ISM surveys 2008 - 2011.

		2008	2009	2010	2011
SN-sample	WI	40.1%	47.3%	49.5%	44.1%
	PAPI	15.4%	16.1%	13.2%	11.8%
	CATI	34.0%	25.3%	23.8%	30.2%
	CAPI	10.5%	11.3%	13.6%	13.9%
ISM	WI	55.9%	68.9%	61.4%	73.2%
	PAPI	11.4%	12.3%	12.0%	12.0%
	CATI	26.6%	17.3%	19.6%	13.5%
	CAPI	6.0%	1.5%	7.0%	1.3%

Table 2. Overview of response mode composition in ISM surveys 2008 - 2011, not weighted.

The estimates based on the national sample of SN are rather stable, as the mode distributions of the respondents in the national sample, as well as the sampling fractions, are comparable in the four editions of the ISM. The response rate of the national sample (approx. 60%) is significantly higher compared to the regional samples (approx. 45%). 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 (Buelens and Van den Brakel, 2010). The SN-sample therefore provides benchmark figures at the national level, which are obtained in a similar manner in all four years. However, the officially released figures are those obtained from the ISM-sample.

5.2 GREG estimation

The inclusion probabilities in de ISM are determined by the sampling desing, accounting for stratification and oversampling at regional levels. The GREG estimator uses the

auxiliary variables age, gender, ethnicity, urbanization, household size, police district, and the strata used in the regional oversampling scheme. The weighting model includes interactions between these variables up to the third order. Variance estimates are obtained with the standard Taylor series approximation of the GREG estimator, see Särndal et al. (1992) ch. 6.

The analysis conducted in this section is implemented in the statistical software package R, using the package for complex surveys, Lumley (2010). An alternative software package that can be used to conduct the same analysis is Bascula. This software is developed for weighting sample surveys, and is available as a component of the Blaise survey processing software system, developed by Statistics Netherlands (2002).

Variable	Description of statistic
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	Mean satisfaction with functioning of police (on a scale 1-10)
antisoc	Mean suffering from anti-social behavior (on a scale 1-10)

Table 3. Overview of key ISM variables and their associated statistics.

For the purpose of this study five key survey variables are considered, see Table 3. For clarity, means are analyzed rather than totals. Estimates of totals are simply divided by the population size, which is known. Table 4 presents annual changes in GREG estimates for the key variables. Estimates based on the SN and ISM samples show diverging developments for some variables, which is not plausible. The changes 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. For 2010-2011, a decrease is observed in offences, based on the SN sample, while the ISM sample indicates an increase. 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. The number of registered offences, for example show a continuously decreasing trend over this time period.

The changes in unsafe in the ISM and SN-samples are very similar for 2008-2009 and 2009-2010, but diverge for 2010-2011. The changes of the two other variables differ less between the two samples. External validation is difficult for these variables, as they concern attitudes and opinions.

From Tables 1 and 2, it is seen that 2009 and 2011 are unusual years in comparison with 2008 and 2010, in that the former are characterized by massive oversampling, and associated with that, the collection of a large proportion of responses through WI. Prior to elaborating on this, it is determined that the ISM-samples of all four years are

Variable	Sample	2008-2009	p	2009-2010	p	2010-2011	p
victim	ISM	+5.0%	<.001	-6.7%	<.001	-0.8%	0.291
	SN	+0.9%	0.347	-3.0%	0.055	-2.5%	0.099
offences	ISM	+9.2%	<.001	-12.5%	<.001	+3.1%	0.051
	SN	+3.3%	0.146	-6.1%	0.011	-4.6%	0.046
unsafe	ISM	+7.0%	<.001	-1.4%	0.159	-1.5%	0.149
	SN	+7.7%	<.001	+0.2%	0.126	-8.2%	<.001
funcpol	ISM	-3.5%	<.001	+1.5%	<.001	+1.0%	0.005
	SN	-1.9%	<.001	0.0%	0.467	+2.7%	<.001
antisoc	ISM	+5.2%	<.001	-1.3%	0.149	-0.7%	0.286
	SN	+3.5%	0.034	+0.5%	0.388	-3.6%	0.014

Table 4. Annual changes in key ISM variables using the standard GREG estimator. Differences are tested with a design-based version of Welch's t-test statistic.

representative with respect to background variables.

5.3 Correcting selection effects

The GREG estimates obtained with the standard model may still be selective with respect to variables correlated with the key target variables. 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-economic status and both victimization rates and response rates. Therefore the following variables indicative of socio-economic class and of crime are used for this analysis; average house value, percentage of inactive people (i.e. not in the labor force), degree of urbanization (at a more detailed regional level than that included in the standard weighting model), and the number of police reported crimes per 1,000 inhabitants. These variables are available at the aggregated level of post code areas and are all categorized, see Table 5. This table shows for 2009 the estimates for these variables, both using the standard weighting model and the model including the mode calibration (details of mode calibration are given in the next section). The known population quantities are almost all 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 mode calibration.

This analysis provides support for the standard weighting model, in it being sufficient in correcting for selective nonresponse.

Variable	Category	ISM standard	ISM mode 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 mode calibrated weighting models. Standard errors are given in brackets.

5.4 Mode calibrated ISM

The mode calibration introduced in section 3 is 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 police district. The four modes used in the ISM are aggregated to two types: modes with and without interviewer. This is done since interviewer presence is an important characterizing factor of response mode effects and to avoid extreme weights for the CAPI respondents in districts where hardly any CAPI interviews are held. Furthermore, WI and PAPI are the response modes offered initially, whereas CATI and CAPI are the follow-up modes. Strictly spoken,

this is not a pure mode calibration, but rather a calibration of data collection strategies, of which the mode is an essential part.

Not weighted		2008	2009	2010	2011
ISM-sample	With interviewer	32.6%	18.7%	26.6%	14.9%
	Without interviewer	67.4%	81.3%	73.4%	85.1%
SN-sample	With interviewer	44.5%	36.6%	37.3%	44.0%
	Without interviewer	55.5%	63.4%	62.7%	56.0%
Weighted (standard model)					
ISM-sample	With interviewer	40.9%	25.6%	35.5%	25.8%
	Without interviewer	59.1%	74.4%	64.5%	74.2%
SN-sample	With interviewer	46.8%	37.4%	38.6%	44.7%
	Without interviewer	53.2%	62.6%	61.4%	55.3%

Table 6. Overview of response composition in ISM surveys 2008 - 2011, not weighted and weighted, for modes with and without interviewer.

Table 6 lists the composition of the respondents according to modes with and without interviewer. The distributions for unweighted and weighted data are shown for the four years. Weighting the sample results in an increase in the share of the respondents interviewed in an interviewer administered mode. These are clearly underrepresented, in particular in the ISM-sample. While the data collection of the SN-sample was conducted identically in each of the three years, the number of responses collected by interviewer administered modes is higher in 2008 and 2011 than in the other years. The larger number of interviews collected using modes without interviewer in the ISM-sample in 2009 and 2011 is due to the large oversampling that took place in these years.

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 are chosen based on the distributions in this sample. The levels for the modes with and without interviewer are set to 40% and 60% respectively, crossed with police district. This implies the inclusion of a term to the weighting model were the 40/60 split is applied to the population size of each separate police district.

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 and 2011 (see Table 6).

A comparison of standard and mode calibrated GREG estimates is shown in Figure

2. 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 changes 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	2010-2011	p
victim	ISM	+2.9%	0.031	-5.3%	<.001	-2.0%	0.085
	SN	-0.2%	0.456	-2.8%	0.074	-2.4%	0.108
offences	ISM	+5.5%	0.005	-10.4%	<.001	+0.9%	0.313
	SN	+1.4%	0.329	-5.8%	0.015	-4.4%	0.053
unsafe	ISM	+1.4%	0.192	+2.0%	0.088	-4.1%	0.002
	SN	+2.7%	0.112	+2.7%	0.083	-7.6%	<.001
funcpol	ISM	-1.8%	<.001	+0.3%	0.107	+2.1%	<.001
	SN	-1.0%	0.026	-0.1%	0.391	+2.6%	<.001
antisoc	ISM	+0.8%	0.269	+1.6%	0.107	-2.4%	0.028
	SN	+1.1%	0.282	+0.8%	0.314	-3.2%	0.023

Table 7. Annual changes in key ISM variables using the mode calibrated GREG estimator. Differences are tested with a design-based version of Welch's t-test statistic.

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 imply 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 changes between 2009 and 2010 are shown, but the pattern of changes between the other years are similar. The column 40/60 contains the results presented before, in Table 7. The other columns show outcomes for some alternative mode compositions; 20/80, 30/70, and 50/50, with the first number referring to the share of responses obtained through interviewer administered modes, and the second number to the share of those without interviewer. While the choice of the calibration

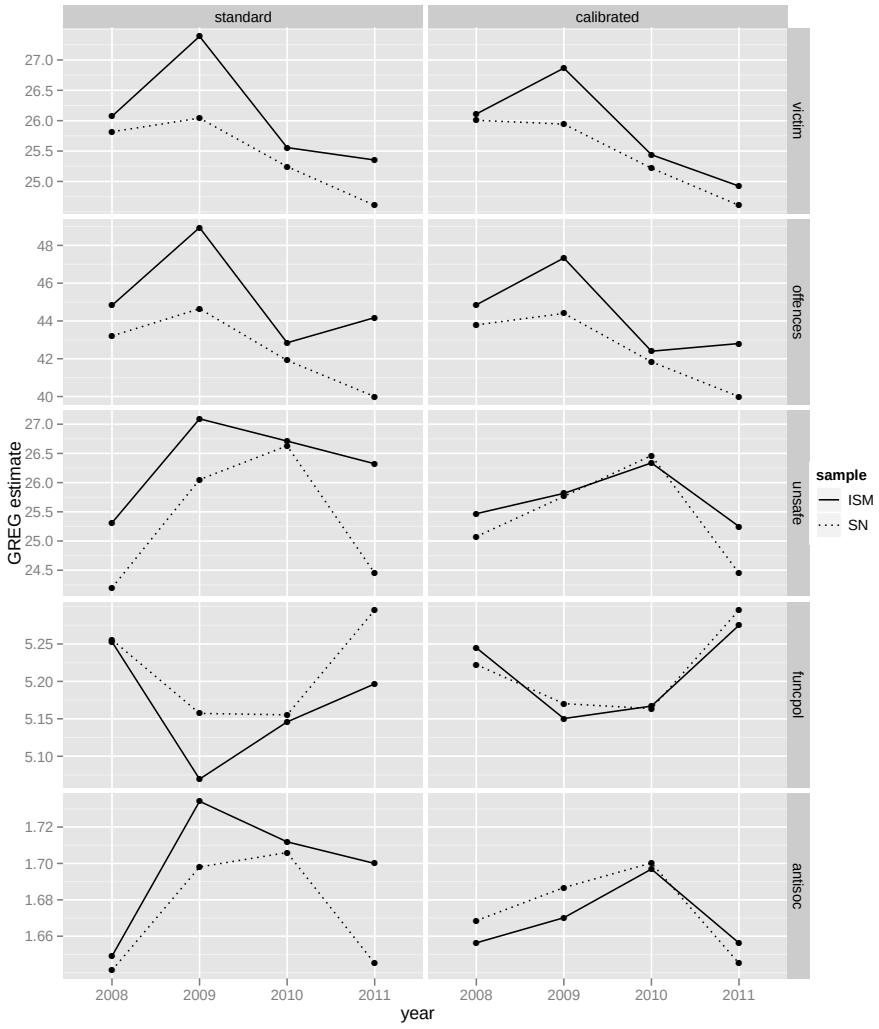

Figure 2. Standard (left) and mode calibrated (right) GREG estimates for the 5 key ISM variables.

levels does affect the outcomes, as expected, the conclusions based on this survey would not change when using different levels. The significance of the changes in the key variables does not depend on the mode composition used in the calibration, apart for the variable unsafe, where for the most extreme case of 20/80 the observed increase is just significant at the 5%-level (p=0.041). This is an indication that the weighting model may not be fully removing all selectivity with respect to this survey variable, and that there is a risk of obtaining biased estimates for this variable when applying mode calibration.

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

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 changes are shown for the key variables for the ISM-sample, for the years 2009-2010. Significance is indicated by *** ($p<.001$), ** ($p<.01$) and * ($p<.05$), with unmarked changes not significant ($p>.05$).

6 Conclusion

Mode related measurement bias is not constant between different editions of repeated surveys that apply mixed-mode data collection, making results of the successive editions of the survey incomparable. This is an important limitation of the use of mixed-mode data collection in official statistics. In this paper the GREG estimator is applied to calibrate the response mode, possibly crossed with other covariates, to predetermined fixed levels. The proposed methodology relies on the assumption that the weighting model removes mode-dependent selectivity with respect to the survey variables and that the underlying measurement error model is constant over time. Note that changes over time in uni-mode surveys too are measured unbiasedly only when the latter assumption holds. Calibrating the distribution of the respondents over the modes to fixed levels causes the bias in the survey estimates due to measurement error to be constant. As a result, different editions of repeated surveys provide outcomes that are comparable, with changes in the survey estimates that are no longer confounded with measurement bias induced by changes in response mode composition. Calibrating the response to fixed mode levels that are crossed with covariates that specify the publication domains of the output tables of the survey also assures that these domain estimates remain comparable.

A simulation study demonstrates that calibration to fixed response levels stabilizes outcomes of surveys subject to varying response mode compositions. Stabilization occurs even if the weighting model does not correct for mode dependent selectivity, although in such cases a bias is introduced. The simulation study also indicates that the mode calibration is not harmful in the absence of measurement errors in the survey variables.

The practical value of the proposed method is that it attempts to stabilize measurement bias without requiring dedicated built in experiments to quantify differences in measurement bias under different data collection modes. Since such experiments are costly and hard to combine with the field work of sample surveys, this approach offers a manageable solution to an important limitation of the use of sequential mixed-mode designs in the

daily practice of official statistics. The Dutch ISM is an example of a survey where different response modes are used sequentially with large differences in the mode compositions between successive editions. Applying the proposed mode calibration clearly stabilizes the survey results. The assumption that the weighting model corrects for mode dependent selectivity can be difficult to validate in practice. The analysis of several correlating background variables provides supporting evidence for the application in the ISM, as does the sensitivity analysis of the chosen calibration levels. Finally the simulation provided evidence that the method improves the stability and comparability of the survey outcomes over time.

The proposed method to improve stability of survey outcomes over time is applicable to surveys in which the mode composition can vary. Preventing such instabilities by choosing stable survey designs is recommendable. This emphasizes the urgency of research into questionnaires that yield comparable outcomes under different data collection modes. Local oversampling in combination with different sets of data collection modes resulting in large variations of mode compositions, as in the case of the ISM, is to be avoided.

Applying the standard variance approximation of the GREG estimator (Särndal et al. (1992), ch. 6) implicitly assumes that the distributions over the modes in the population, used in the weighting model of the GREG estimator, are known exactly from external sources. It is not obvious to which extent the property is violated that the GREG estimator is asymptotically design-consistent by including an arbitrary chosen response distribution in the weighting model. Additional research might focus on variance approximations that reflect the additional uncertainty of choosing arbitrary levels for these mode distributions, and on an assessment under which conditions the mode calibrated GREG estimator is design consistent.

The methodology developed in this paper is aimed at population totals using the GREG estimator. They apply directly to population means and ratios as long as the denominator is not subject to different measurement errors under different modes. The methods are not directly applicable to the estimation of more complex parameters such as regression coefficients, or ratios where the denominator is sensitive to differences in measurement error. Inference procedures for such parameters requires additional research, probably outside the traditional design-based framework followed in this paper.

Another line of research is the development of sample designs with dedicated in-built experiments that are aimed at quantifying measurement errors. If the assumed measurement error model is not constant over time, more advanced estimation techniques are required to stabilize the measurement error. This may require a model-based inference approach that takes advantage of the experimental information about the change of the measurement error over time.

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.

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.

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.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley, New York.

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.

Statistics Netherlands (2002). *Blaise developer's guide.* Statistics Netherlands, Heerlen (Available from www.Blaise.com).

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.

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.