DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 513–531, http://dx.doi.org/10.1515/JOS-2017-0024

Comparing Two Inferential Approaches to Handling Measurement Error in Mixed-Mode Surveys

Bart Buelens[1] *and Jan A. Van den Brakel*[2]

Nowadays sample survey data collection strategies combine web, telephone, face-to-face, or other modes of interviewing in a sequential fashion. Measurement bias of survey estimates of means and totals are composed of different mode-dependent measurement errors as each data collection mode has its own associated measurement error. This article contains an appraisal of two recently proposed methods of inference in this setting. The first is a calibration adjustment to the survey weights so as to balance the survey response to a prespecified distribution of the respondents over the modes. The second is a prediction method that seeks to correct measurements towards a benchmark mode. The two methods are motivated differently but at the same time coincide in some circumstances and agree in terms of required assumptions. The methods are applied to the Labour Force Survey in the Netherlands and are found to provide almost identical estimates of the number of unemployed. Each method has its own specific merits. Both can be applied easily in practice as they do not require additional data collection beyond the regular sequential mixed-mode survey, an attractive element for national statistical institutes and other survey organisations.

Key words: Generalized regression; mode effects; selection bias; response mode calibration; counterfactuals.

1. Introduction

In mixed-mode sample surveys multiple modes of data collection are combined. Sequential designs apply different modes consecutively, approaching nonrespondents of one mode through a different mode. Each mode of interviewing has its own associated measurement error obstructing unbiased estimation of means or totals of true scores (Jäckle et al. 2010; Schouten et al. 2013; Buelens and Van den Brakel 2015). When different modes are administered in the same survey the total response consists of a mix of interviews obtained through the different modes, and associated therewith, a mix of mode related measurement bias. In surveys that are repeated over time, the mode composition of the mix may vary, and so may the overall measurement bias of estimated means and totals of survey variables. Confounding of true change over time of a survey statistic with change in mode composition limits the usefulness of mixed-mode surveys (Buelens and Van den Brakel 2015; Cernat 2015).

Despite this limitation, conducting surveys using a mix of interview modes has gained popularity in recent years. Benefits include cost – as a substantial number of respondents

[1] Statistics Netherlands, PO Box 4481, 6401 CZ Heerlen, The Netherlands. Email: b.buelens@cbs.nl
[2] Statistics Netherlands and Maastricht University, PO Box 4481, 6401 CZ Heerlen, The Netherlands. Email: ja.vandenbrakel@cbs.nl

are typically interviewed using cheap modes such as the internet – and more representative samples – as respondents who would refuse participation in one mode may be willing to respond in an other mode (De Leeuw 2005; Voogt and Saris 2005). A topical research question in the context of mixed-mode surveys is the influence of mode-specific measurement error on final survey estimates, see for example Lynn (2013); Vannieuwenhuyze and Loosveldt (2013); Schouten et al. (2013); Buelens and Van den Brakel (2015); Klausch et al. (2015).

In the present article two lines of research on measurement error are distinguished and their principles and merits are compared. Both are adaptations of the widely used general regression (GREG) estimator by which survey estimates of totals are expressed as $\sum_k w_k y_k$, a weighted sum of the observations y_k (Särndal et al. 1992). One approach seeks to adjust the survey weights w_k and is aimed at stabilizing total measurement error in repeated surveys (Buelens and Van den Brakel 2015). The other approach leaves the survey weights unchanged and instead proposes adjustments to the observed values y_k in order to remove measurement error (Suzer-Gurtekin et al. 2012; Suzer-Gurtekin 2013). While the two methods are motivated differently, it is shown in this article that both methods are identical for a certain parameterisation when the underlying assumptions are met. The two methods are explained and applied to a series of 36 months of the Dutch Labour Force Survey, in which three interview modes are used. This analysis provides insight into the extent to which sequential mixed-mode surveys that are repeated over time are susceptible to variations in mode composition, and how the estimation method can be adapted accordingly. Both methods are applicable to sequential mixed-mode designs and do not require the collection of additional data either by expanding the questionnaire with additional questions, for example Vannieuwenhuyze and Loosveldt (2013), or by re-interviewing respondents, for example Schouten et al. (2013).

This article contributes to the existing literature on inference with mixed-mode surveys by analytically establishing the conditions under which two different inference procedures for sequential mixed-mode surveys are equivalent. This sheds additional light on the properties of both methods. The results are illustrated by applying both methods to a series of monthly samples of the Dutch Labour Force Survey.

In Section 2 the inference methods under consideration are detailed and their assumptions discussed. Section 3 provides details of the Labour Force Survey (LFS) in the Netherlands. The results of applying the different methods to the LFS are presented in Section 4. Section 5 concludes the article.

2. Methods of Inference

2.1. GREG Estimation

The general regression estimator (GREG) of the total t_u of a variable u can be written as a weighted sum

$$\hat{t}_u = \sum_{k=1}^{n} w_k u_k \tag{1}$$

with u_k the values of u for survey respondents $k = 1, \ldots, n$ and w_k weights. The weights account for unequal inclusion probabilities associated with the sampling design and they

correct for selective nonresponse by calibrating the weights such that the sum over the weighted auxiliary variables equate the known totals in the population. Details of this method including variance estimation can be found in Särndal et al. (1992).

2.2. *Response Mode Calibration*

This paragraph summarizes an approach proposed by Buelens and Van den Brakel (2015) called response mode calibration. When measuring the variable u through a survey mode m, the measurement can be modeled as

$$y_{k,m} = u_k + b_m + \epsilon_{k,m} \tag{2}$$

with $y_{k,m}$ the observations through mode m of the true values u_k, b_m the systematic effect of mode m and $\epsilon_{k,m}$ random mode dependent error components with expected values equal to zero.

Inserting (2) in the GREG estimator for the observed total and taking the expectation with respect to the measurement error model gives

$$\hat{t}_y = \sum_{k=1}^{n} w_k y_k = \hat{t}_u + \sum_{m=1}^{p} b_m \hat{t}_m \tag{3}$$

with $\hat{t}_m = \sum_{k=1}^{n} w_k \delta_{k,m}$ and $\delta_{k,m}$ a dummy indicator equal to one if unit k responded through mode m and zero otherwise.

While the parameter p ordinarily corresponds to the number of modes applied in a survey, other conceptualizations are possible. For example p can refer to the number of interview strategies that are believed to have different associated measurement errors. Additionally, p can refer to a cross-classification of response mode or strategy, and other categorical auxiliary variables; this allows for modeling of a different measurement bias for different population subgroups.

Equation (3) expresses that the estimate of the true total, $\hat{t}_u$, is observed with error $\sum_{m=1}^{p} b_m \hat{t}_m$, a combination of mode-dependent biases. The quantity $\hat{t}_m$ can be interpreted as the estimated number of units responding through mode m in the population under the given survey design. Of the quantities in Equation (3), only $\hat{t}_y$ and $\hat{t}_m$ are observed, $\hat{t}_u$ and b_m are not.

The issue addressed by the method of response mode calibration is that in repeated surveys the response mode composition may vary between editions, leading to varying $\hat{t}_m$ and hence to a varying bias in the observed totals $\hat{t}_y$. This problem can be prevented if the bias term in Equation (3) is rendered constant. This is achieved by applying a response mode calibration as proposed byBuelens and Van den Brakel (2015). The response mode composition is calibrated to a fixed distribution, effectively requiring the $\hat{t}_m$ to equal given values. As this is exactly what the GREG estimator achieves for the other auxiliary variables, the response mode calibration is straightforwardly implemented by extending the underlying regression model with an additional covariate, response mode, and defining arbitrary but fixed response mode levels $\{\Gamma_m\}_{m=1,\ldots,p}$.

The resulting mode calibrated GREG estimator is

$$\hat{t}_y^c = \sum_{k=1}^{n} w_k^c y_k = \hat{t}_u^c + \sum_{m=1}^{p} b_m \hat{t}_m^c = \hat{t}_u^c + \sum_{m=1}^{p} b_m \Gamma_m \tag{4}$$

with w_k^c the weights resulting from the mode calibrated GREG – compare to expression (1) – and $\hat{t}_u^c = \sum_{k=1}^{n} w_k^c u_k$. By construction of the mode calibrated GREG, $\hat{t}_m^c = \Gamma_m$ for all m. The b's are the regression coefficients of response mode in the GREG weighting model. The variance of the mode calibrated GREG is obtained using the ordinary GREG variance estimation (Särndal et al. 1992), applied as if the calibration levels are known population totals. While the calibration levels Γ_m can be chosen arbitrarily, it is recommended to choose levels close to those realized in the survey. Otherwise the estimator becomes inefficient, inflating the variance unnecessarily as follows from the simulation conducted by Buelens and Van den Brakel (2015). If long-term systematic changes of the realized mode composition occur, the calibration levels Γ_m can be changed and past results can be recalibrated to the new levels to sustain a consistent time series.

A strong assumption of this method is that $\hat{t}_u = \hat{t}_u^c$. This assumption is fulfilled if response mode does not explain any selectivity of the response beyond that explained by the other covariates in the regression model of the GREG. One of the approaches to verify this assumption is suggested by Buelens and Van den Brakel (2015) and consists of applying both the usual and the mode calibrated GREG to register variables known for the survey respondents. As these variables are measured independent of the survey, mode calibration should have no effect as there cannot be a mode-dependent measurement error.

In summary, response mode calibration replaces the original weights w_k in Equation (1) by their mode calibrated version w_k^c and leaves the observations y_k unchanged. Measurement errors are not corrected for, they are merely balanced to render the total measurement bias constant across survey editions.

2.3. *Measurement Error Correction*

When measurement errors are estimated explicitly, estimates can be corrected towards a benchmark survey mode. A model based approach predicting counterfactuals – responses that would have been obtained through another mode than that actually used – has been proposed by Suzer-Gurtekin et al. (2012) and Suzer-Gurtekin (2013). A slightly modified version of their method is implemented here and summarized as follows.

Combining the linear model underpinning the GREG estimator, $u = \beta X + e$, with Equation (2) results in the regression model

$$y_{k,m} = \beta X_k + b_m \delta_{k,m} + \tilde{e}_{k,m} \tag{5}$$

with $\tilde{e}_{k,m} = \epsilon_{k,m} + e_{k,m}$, β a vector of regression coefficients for covariates other than mode, and b_m the regression coefficients for the modes $m = 1, \ldots, p$. If response mode does not explain any selectivity beyond that explained by the other covariates X, the coefficients b_m equal the measurement errors of the modes. This assumption is the same as the assumption required in the mode calibration approach.

In contrast to the mode calibration approach, the correction approach seeks to estimate the unknown parameters b_m explicitly. Fitting Model (5) using least-squares regression

results in estimates $\hat{\beta}$ and $\hat{b}_m$ of the regression coefficients. The estimated regression coefficient $\hat{b}_m$ is at the same time an estimate of the measurement error b_m in Equation (2).

Model (5) is taken to be linear here for fair comparison with the mode calibration method which employs linear models too. If desired, one could choose a generalized linear model such as a logistic regression model.

Suzer-Gurtekin et al. (2012) and Suzer-Gurtekin (2013) propose to use the fitted model to predict individual observations under an alternative mode, counterfactuals,

$$\hat{y}_{k,m}^{m'} = \hat{\beta} X_k + \hat{b}_{m'} \tag{6}$$

which can be calculated for every m' in $1, \ldots, p$. The estimate $\hat{y}_{k,m}^{m'}$ is the predicted outcome of observing unit k through mode m' while it really was observed through mode m. In this article, counterfactuals are instead obtained in a corrective rather than a predictive manner,

$$\hat{y}_{k,m}^{m'} = y_{k,m} - \hat{b}_m + \hat{b}_{m'} \tag{7}$$

which again can be computed for all m and m' in $1, \ldots, p$. The estimated measurement error of the original mode is now removed, and that of the alternative mode is added to the observations. The counterfactuals computed through (7) are closer to the initial observations than those obtained through (6).

Using the counterfactuals, a mode specific estimate of the total is obtained as

$$\hat{t}_y^{m'} = \sum_k \delta_{k,m'} w_k y_{k,m'} + \sum_k (1 - \delta_{k,m'}) w_k \hat{y}_{k,m}^{m'}, \tag{8}$$

the sum over measurements of units observed in mode m' and counterfactuals of units observed in other modes. This estimator would typically be applied if one of the modes is the preferred mode towards which other measurements are benchmarked.

Using the counterfactuals as obtained in (7), Expression (8) can be written as

$$\hat{t}_y^{m'} = \sum_k w_k \hat{y}_{k,m}^{m'}. \tag{9}$$

The variance of $\hat{t}_y^{m'}$ has two sources, associated with the two terms in Equation (8). The first source is the design variance due to sampling. The second is model-based and due to model uncertainty. Suzer-Gurtekin (2013) adopt a multiple imputation approach to capture the model induced variance. Here, a bootstrap approach is followed instead, capturing the design and model variances simultaneously. Through repeated sampling with replacement from the original sample, a bootstrap distribution of $\hat{t}_y^{m'}$ is obtained, from which the total variance is calculated.

If there is no benchmark mode or preference for one mode specifically, different counterfactuals can be combined. As the models are linear this can be done at aggregate level,

$$\hat{t}_y^{combi} = \sum_{m=1}^{p} \alpha_m \hat{t}_y^m \tag{10}$$

with α_m mixing coefficients summing to one, defining the mode composition of the final estimator. The variance of this combined estimator is again estimated through bootstrapping.

Using (9) and Expressing (10) as

$$\hat{t}_y^{combi} = \sum_k w_k \left(\sum_{m'=1}^{p} \alpha_{m'} \hat{y}_{k,m}^{m'} \right) \tag{11}$$

it is clear that this estimator involves adjustments to the observed values y_k and leaves the original weights unchanged. For the calibration estimator (4) the reverse holds: the weights are adjusted and the measurements are kept unchanged.

Suzer-Gurtekin (2013) propose to choose values for α_m through an optimization procedure, for example minimizing the variance or MSE. In the present study, a comparison with the calibration approach is the primary goal. Therefore the most sensible choice is to choose the mixing proportions α_m such that they correspond to the calibration levels Γ_m in Subsection 2.2. For each mode m, α_m and Γ_m are chosen so that $\alpha_m = \Gamma_m/N$ with N the known population total. With this choice, the calibration estimator (4) and the correction estimator (10) are both composed of the same mixing composition of modes, facilitating comparative analyzes.

2.4. *Relation Between the Two Methods*

When setting the levels in the calibration approach to Γ_m and the mixing proportions in the correction approach to $\alpha_m = \Gamma_m/N$, it can be shown analytically that the two methods are approximately equal. The relation between the two methods has not been addressed before in earlier research.

Using Expression (7), the combined measurement error correction estimator (11) can be written as

$$\hat{t}_y^{combi} = \sum_k w_k \left(\sum_{m'=1}^{p} \alpha_{m'} (y_{k,m} - \hat{b}_{m(k)} + \hat{b}_{m'}) \right). \tag{12}$$

with $\hat{b}_{m(k)}$ denoting the actual response mode of respondent k.

According to measurement error model (2), $y_{k,m} - b_m = u_k + \epsilon_{k,m}$. Expression (12) can be elaborated as

$$\hat{t}_y^{combi} = \sum_k w_k \sum_{m'=1}^{p} \frac{\Gamma_{m'}}{N} (u_k + b_{m(k)} - \hat{b}_{m(k)} + \hat{b}_{m'} + \epsilon_{k,m}).$$

Taking the expectation with respect to the measurement error model gives

$$\begin{aligned} \hat{t}_y^{combi} &= \sum_k w_k \sum_{m'=1}^{p} \frac{\Gamma_{m'}}{N} (u_k + \hat{b}_{m'}) \\ &= \sum_k w_k u_k + \sum_k w_k \sum_{m=1}^{p} \frac{\Gamma_m}{N} \hat{b}_m \\ &= \hat{t}_u + \sum_{m=1}^{p} \Gamma_m \hat{b}_m. \end{aligned}$$

It is assumed that $\alpha_m = \Gamma_m/N$ and that $\sum_{k=1}^{n} w_k = N$. The former is a choice one can make, as said before. The latter equality holds if the weighting model at least uses the target

population size as an auxiliary variable, which is the case if at least one categorical variable dividing the population in two or more poststrata is included – which is almost always the case in practice. Finally the equality holds only approximately since $b_{m(k)} \approx \hat{b}_{m(k)}$.

Comparing Expressions (4) and (13) shows that both estimators are equal if in (4) the assumption holds that $\hat{t}_u^c = \hat{t}_u$, which is the case if the response mode does not explain any selectivity beyond that explained by the other auxiliary variables. In addition, it has been assumed that the GREG models used in both approaches are the same, that the model in Expression (5) is identical to the GREG model extended with response mode, and that it is this model that is used in both the calibration and correction approaches.

3. The Labour Force Survey

Statistics Netherlands conducts the Labour Force Survey (LFS) using a rotating panel design consisting of five waves. Since April 2012, data collection in the first wave follows a sequential mixed-mode strategy. Respondents are invited by regular mail to complete the survey online via the web. Nonrespondents are approached through telephone interviewing if they have a known telephone number and are a household with fewer than three people, and through face-to-face interviewing otherwise. Interviews in the second to fifth waves are conducted by telephone only – a contact telephone number is asked for in the first interview.

The LFS is a household survey. The target population is the non-institutionalized population aged 15 years or over residing in the Netherlands. The sampling frame is obtained from municipal registrations and consists of all known occupied addresses in the country. Each month, a stratified two-stage cluster design of addresses is selected, with strata formed by geographic regions. Municipalities are primary sampling units and addresses secondary. All households residing at an address, up to a maximum of three, are included in the sample and can be regarded as the ultimate sampling units. Each year approximately 140,000 households are in the LFS sample. In 2014, approximately 30,000 households responded via the web, 12,000 via face-to-face interviewing, and 9,000 by telephone. Not all of the web-nonrespondents are re-approached by a different mode; approximately 28,000 addresses are approached for face-to-face interviewing and 24,000 for telephone. These and other details can be found in reports published by Statistics Netherlands, such as the LFS 2014 report (Centraal Bureau voor de Statistiek 2015).

The response data are weighted to account for the survey design and for selective nonresponse using a GREG procedure, see Subsection 2.1. Weighting is conducted for each of the five waves independently. The GREG weighting model used for production of the regular unemployment statistics contains the variables listed in Table 1. All variables are categorical with the number of categories for each variable given in brackets. Age and sex are included as an interaction and the remaining variables as main effects. The variable 'registered unemployed' indicates registration with the Employment Agency and does not coincide with the LFS definition of being unemployed. Registration at the Employment Agency is not compulsory for the unemployed – it is required only to be eligible for unemployment benefits or to receive training or coaching. Given the survey design, it would be sensible to include a dichotomous variable indicating whether households can be

Table 1. Variables used in the regular monthly GREG estimates of the LFS.

Variable (number of categories)	Definition
Sex (2)	Male or female
Age (21)	Age classes
Household type (3)	With children, single-person, other
Region (43)	NUTS-3 areas and largest cities
Registered unemployed (5)	Duration of registration (0 meaning not registered)
Income class (6)	Standardised household income
Income type (3)	Salary, welfare benefit, unknown
Ethnicity (3)	Native, western immigrant, non-western immigrant

reached by telephone. Unfortunately no such population frame data are available; a third-party provides telephone numbers of households in the sample only.

The GREG results are used as input for a structural time series model. Through the use of such model, the precision of the estimates is increased as the model allows for borrowing strength from previous time periods. In addition, the model takes into account rotation group bias and discontinuities due to the survey redesigns in 2012 and before, see Van den Brakel and Krieg (2015). The structural time series model explicitly accounts for the systematic differences between the first and subsequent waves by benchmarking the outcomes for the second, third, fourth and fifth waves to the level of the first. The level estimates resulting from the first wave of the survey are therefore crucial. To avoid

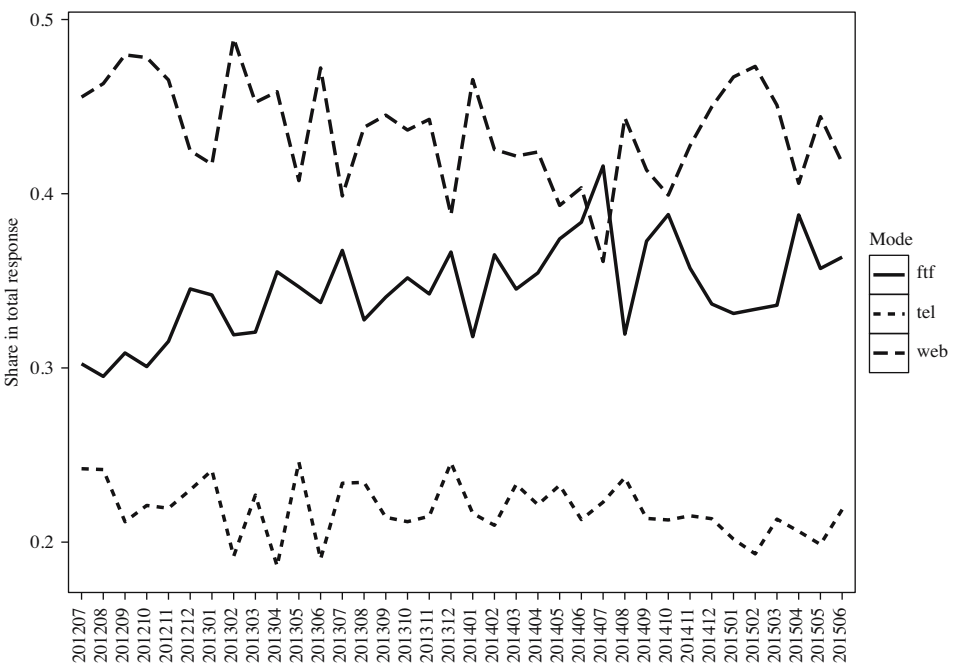

Fig. 1. Response mode composition of the LFS response during the 36 month study period; the three modes are face-to-face (ftf), telephone (tel), and web.

additional technical complications with this time series modeling approach, only the level estimates obtained in the first wave are used in this research.

In this article, first wave GREG weighted estimates from the LFS from July 2012 through June 2015 are studied, a period of 36 months. In the remainder of this article, this series is referred to as the regular approach – not applying any of two adjustment methods. Data collected in the subsequent telephone-only waves are not used in the present research. Issues pertaining to the redesigns of April 2012 and earlier are not discussed as they precede the study period. Executing the sequential mixed-mode strategy and applying the GREG procedure results in a weighted survey response composed of a mix of three modes, web, telephone and face-to-face. The composition varies from month to month and is shown in Figure 1. The share of telephone is rather constant. Face-to-face and web are exchanged in that months with relatively low web shares exhibit relatively high face-to-face shares and vice versa. The average mode composition over the study period is web 44%, telephone 22%, and face-to-face 34%.

4. Results

4.1. Response Mode Calibration

The calibration method of Subsection 2.2 is applied to the LFS, independently for each month of the 36 month study period. Four different calibration schemes are executed. The first, calBalanced, is the scheme that would ordinarily be applied based on recommendations in earlier research (Buelens and Van den Brakel 2015), taking the proportions for the three modes to be the averages over the study period, 44% web, 22% telephone, and 34% face-to-face interviews. The other three schemes are more extreme, each suppressing the contribution of one of the modes: two modes are calibrated to 45% each, and the third mode to ten per cent. These alternative schemes are executed to assess robustness and to illustrate the mode calibration technique.

The resulting estimates of the number of unemployed are shown in Figure 2. The mode calibrated estimates are presented relative to the number of unemployed estimated using the regular approach, which consists of the GREG estimates obtained from the survey weights, without applying mode-related calibration adjustments. The calBalanced alternative does not deviate a lot from the regular approach. The more extreme alternatives exhibit larger deviations. Estimates that are five per cent higher or lower than the regular estimates occur often. Table 2 lists the estimated monthly number of unemployed averaged over the whole study period. The calLessWeb and calLessFtf approaches result in systematically lower estimates, while the calLessTel results in a systematically higher estimate. Under the assumptions of the method, these differences are due to measurement error. In this case the telephone mode must measure lower than the other two modes.

The estimated standard errors of the point estimates are obtained with the standard analytic approximation for the variance of the GREG estimator and are shown in Figure 3 and are relative to the standard errors of the regular approach. The errors of the calBalanced approach are similar to those of the regular approach. The alternative approaches have larger standard errors, as expected, as they use the sample in a less efficient manner due to up or down weighting of respondents of certain modes. Of the

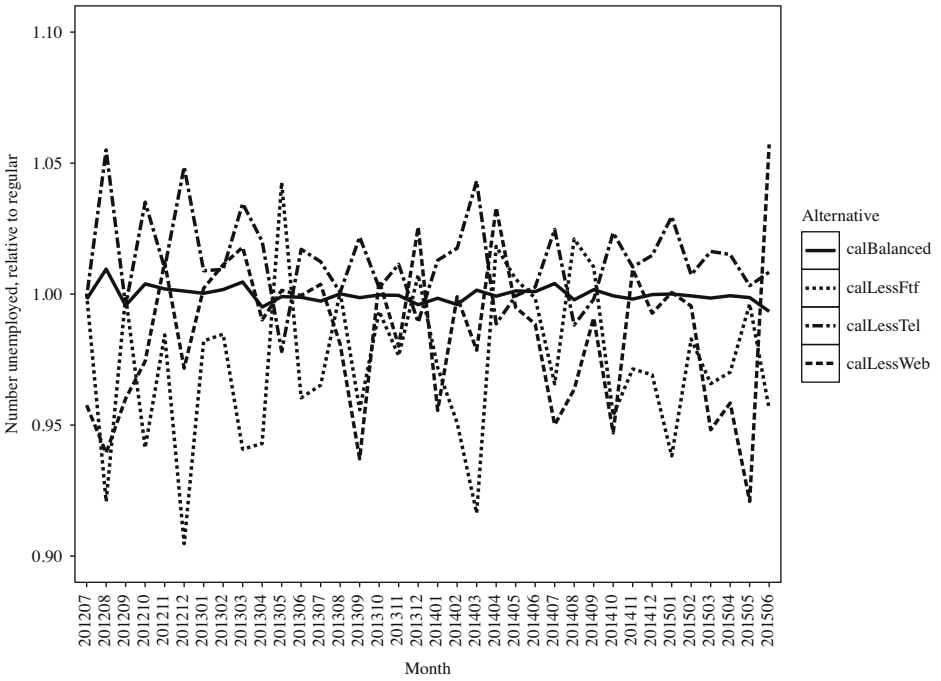

Fig. 2. Estimates of the total number of unemployed obtained through the calibration approach, relative to the regular approach.

three alternatives, the calLessWeb is the least efficient. This is expected, as the share of Web respondents is largest, so suppressing them has the most extreme adverse effect on the efficiency.

If one were to apply the mode calibration method to the LFS for production purposes, the recommendation would be in accordance with Buelens and Van den Brakel (2015) to use calibration levels that are close to the levels realized in the survey. In this case, this would be the calBalanced approach.

4.2. *Measurement Error Correction*

The measurement error correction approach presented in Subsection 2.3 is applied to the same LFS data. Measurement errors are estimated using a regression model with survey

Table 2. Number of unemployed averaged over the 36 month study period, under the various schemes. The composition is the percentage share of Web-Tel-Ftf.

Scheme	Mode composition	Unemployed	SE
regular	variable	678,126	5,211
calBalanced	44-22-34	677,863	5,202
calLessWeb	10-45-45	668,539	6,482
callLessTel	45-10-45	686,634	5,555
calLessFtf	45-45-10	660,369	5,847

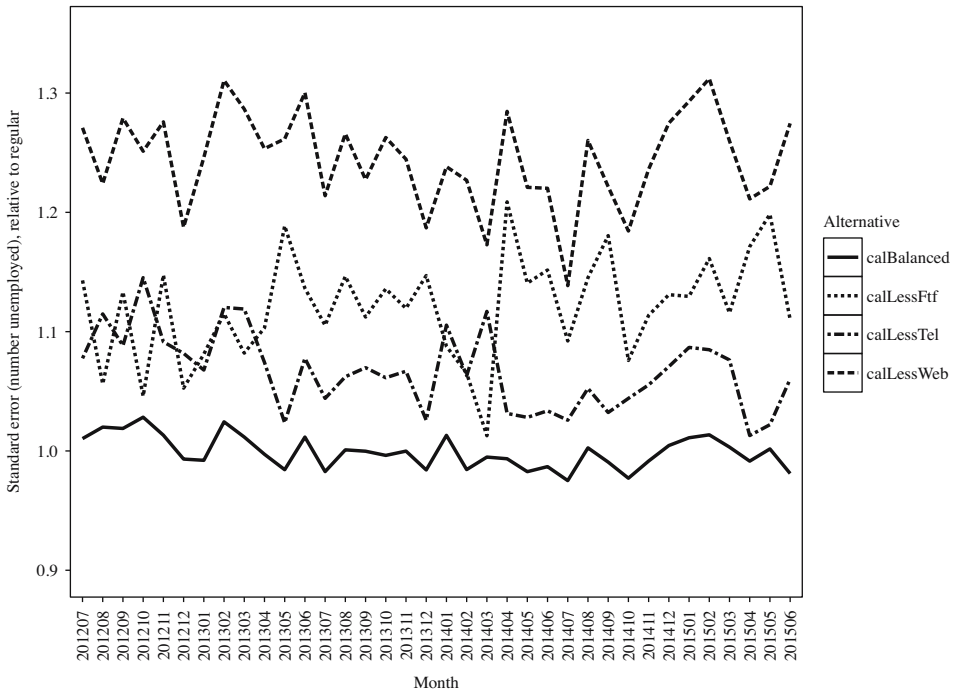

Fig. 3. Standard errors of estimates of the total number of unemployed obtained through the calibration approach, relative to the regular approach.

mode as an explanatory variable in addition to the variables in the GREG model (see Table 1). Since it can be expected that the measurement error does not change during the study period the model is fitted with all data pooled. To allow for between-month variance not explained by the other covariates, month itself is added to the model as a covariate. Corrections are applied in an additive manner using the estimated regression coefficients, which correspond to estimates of the measurement errors, see Equation (7).

Four estimators are considered. One for each mode, corFtf, corTel, and corWeb, which correct the measurements towards face-to-face, telephone, and web modes respectively. A combined correction estimator, corCombi, is a mix of the other three with mixing coefficients in line with the calibration levels of the calBalanced estimator, ie. 44% web, 22% telephone, and 34% face-to-face.

The resulting estimates are shown in Figure 4, again relative to the level of the regular approach. The corCombi estimates are almost equal to the regular estimates. The corFtf and corWeb estimates are higher and the corTel estimates are lower than the regular estimates. Under the assumptions of the applied method, these level differences are due to relative measurement bias between the modes. The finding that telephone interviewing measures at a level below that of the other modes confirms the results of the calibration approach.

The standard errors of these estimates are obtained with a bootstrap procedure and are shown in Figure 5. They are all relatively small compared to the standard errors of the calibration estimators other than the balanced version, see Figure 3. The corFtf and corTel

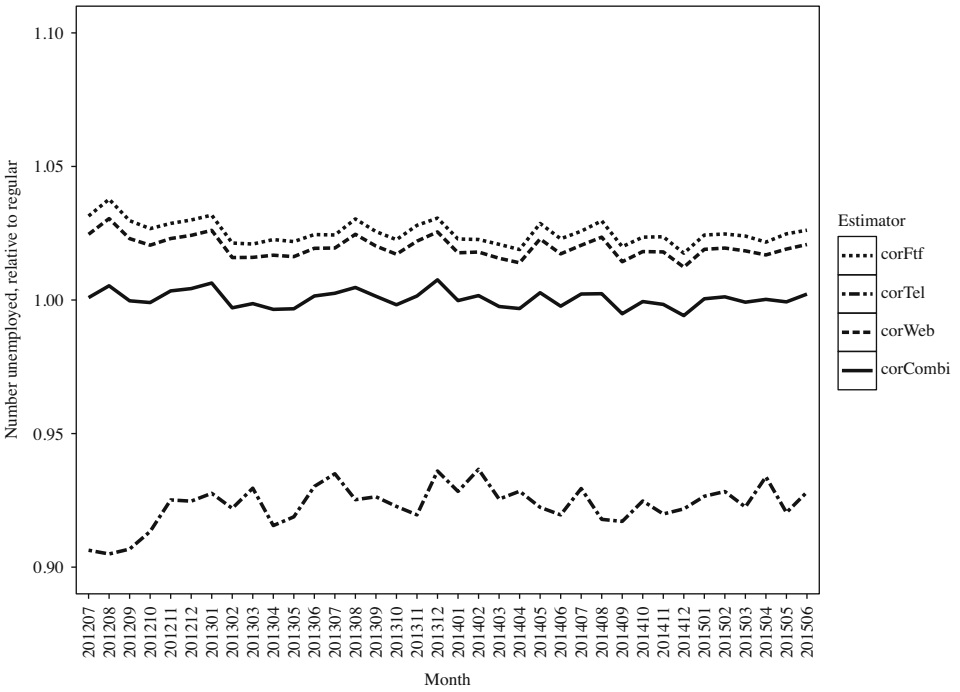

Fig. 4. Estimates of the total number of unemployed obtained through the correction approach, relative to the regular approach.

standard errors are largest as they both require more unit observations to be corrected. The corWeb estimates have standard errors that are only marginally larger than the corCombi estimates, which are similar to the standard errors of the regular approach.

Similar to the annual results for the calibration estimator (see Table 2), the annual results for the correction estimators are shown in Table 3. Of the three estimators that are corrected towards a single mode, the web and face-to-face estimators give comparable results, while the telephone estimator results in a substantially lower estimated number of unemployed. Consequently, the combined estimator results in a level estimate above telephone and below web and face-to-face. The combined estimate is almost equal to the estimate obtained with the regular approach. It is important to stress again that selection bias that is not explained by the model might contribute to the differences seen in Table 3.

It is an empirical result that the estimates corFtf and corWeb are comparable and that both are higher than corTel. The differences are due to mode-dependent measurement errors. The difference between telephone and face-to-face interviewing found here is in line with earlier research; with a randomized experiment embedded in the Dutch LFS, Van den Brakel (2008) showed that the unemployment rate under telephone interviewing is significantly lower than under face-to-face interviewing. The Dutch LFS is a household survey where a response is required from all adult household members. Proxy responses are allowed and are much more frequent in telephone interviews than in face-to-face, which may explain at least a part of the observed differences. Other explanations could be offered by cognitive models of the survey response process. Such models provide a

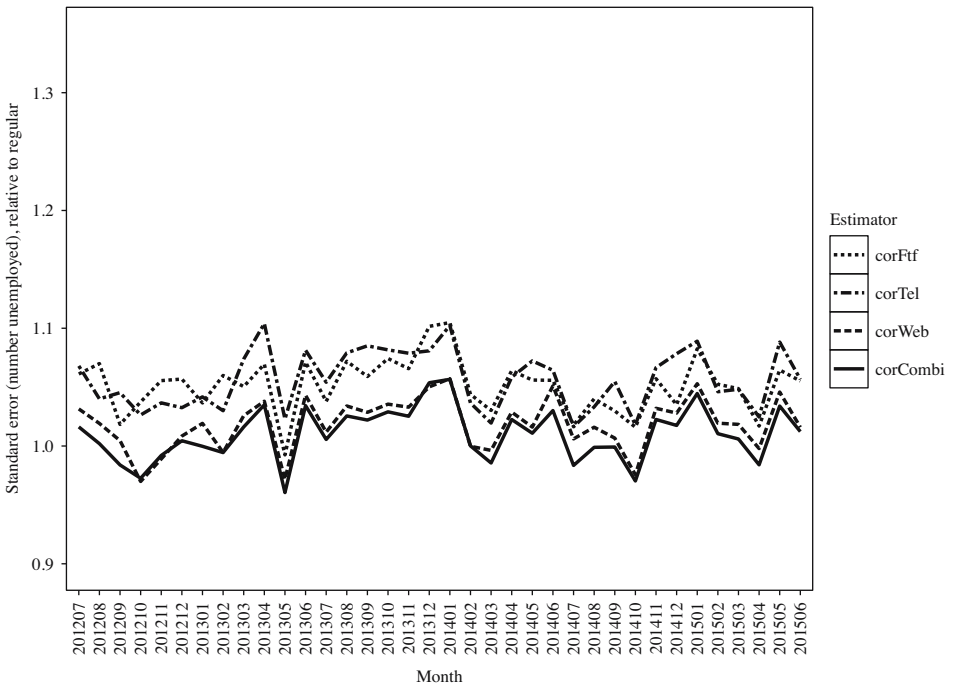

Fig. 5. Standard errors of estimates of the total number of unemployed obtained through the correction approach, relative to the regular approach.

framework for describing the process by which respondents interpret questions, retrieve the required information, make judgements about an adequate response, and provide an answer (Cannel et al. 1981; Tourangeau et al. 2000). A complicating factor in understanding the effects seen in the present analysis is that labour status is derived from a set of questions to determine whether a respondent is working, or willing to work, and is actively looking for work, among other elements. Respondents are generally more likely to give socially desirable answers and demonstrate acquiescence in the presence of an interviewer than in self-administered modes (Dillman et al. 2009; Holbrook et al. 2003). Satisficing (Krosnick 1991) occurs more frequently in self-administered modes than in interviewer modes, and within interviewer modes satisficing occurs more in telephone interviews than in face-to-face interviews, due to the higher speed of the former (Holbrook

Table 3. Number of unemployed averaged over the 36 month study period, using the various correction estimators. The composition is the percentage share of Web-Tel-Ftf.

Estimator	Mode composition	Unemployed	SE
regular	variable	678,126	5,211
corCombi	44-22-34	678,394	5,267
corWeb	100-0-0	691,374	5,311
corTel	0-100-0	626,581	5,507
corFtf	0-0-100	695,122	5,482

et al. 2003). Primacy and recency effects are factors that may explain differences between visual and aural modes (Krosnick and Alwin 1987). They do not completely explain the observed differences, since for some of the questions used to derive the labour market status of respondents, the answer categories are not read out loud by the interviewer. In these cases the interviewer asks an open question and chooses an appropriate answer category based on the answer provided by the respondent. Under the web mode, the respondent can read the different answer categories.

The explanations for the differences between the modes offered by these theories are tentative only. It is not possible to draw conclusions about the validity of the estimates under the different modes, or to choose one of the modes as the benchmark best approximating the true level of unemployment.

4.3. Calibration versus Correction

Comparing the preferred calibration approach, where a mode composition is chosen that resembles that actually realized in the survey, to the correction approach with mixing coefficients that are chosen accordingly, gives rise to Figures 6 and 7. All three estimation methods result in virtually the same series of unemployed (Fig. 6) with very similar standard errors (Fig. 7). This is in agreement with the established relation between the two methods, see Subsection 2.4. The empirical outcome that the calibration and correction methods give the same results is reassuring as they are largely based on the same assumptions and models, albeit motivated differently. The small differences between both

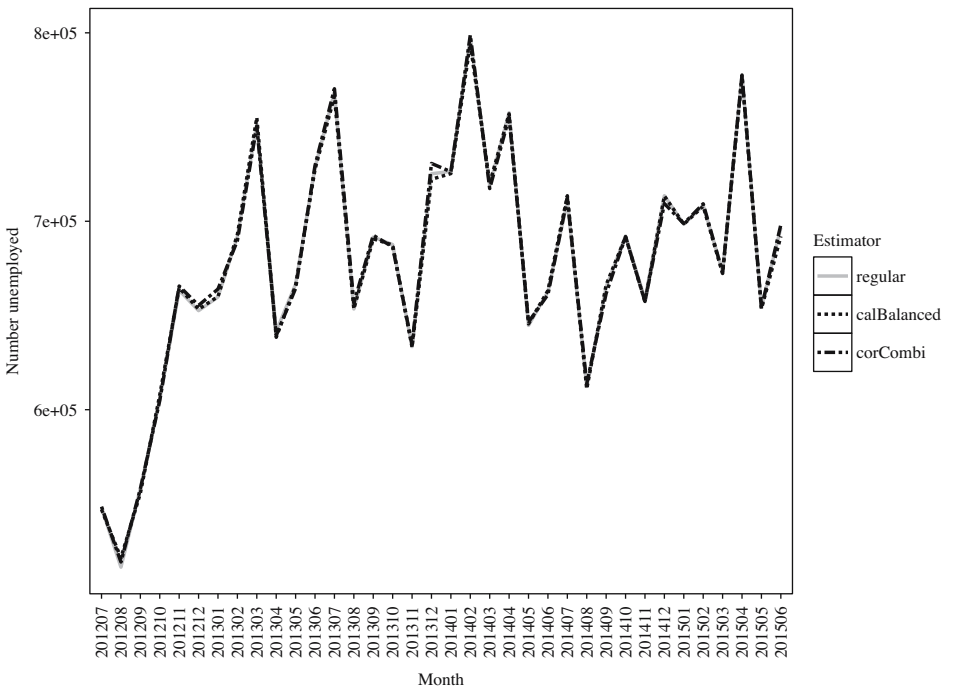

Fig. 6. Estimates of the total number of unemployed obtained through the regular, calibration and correction approaches.

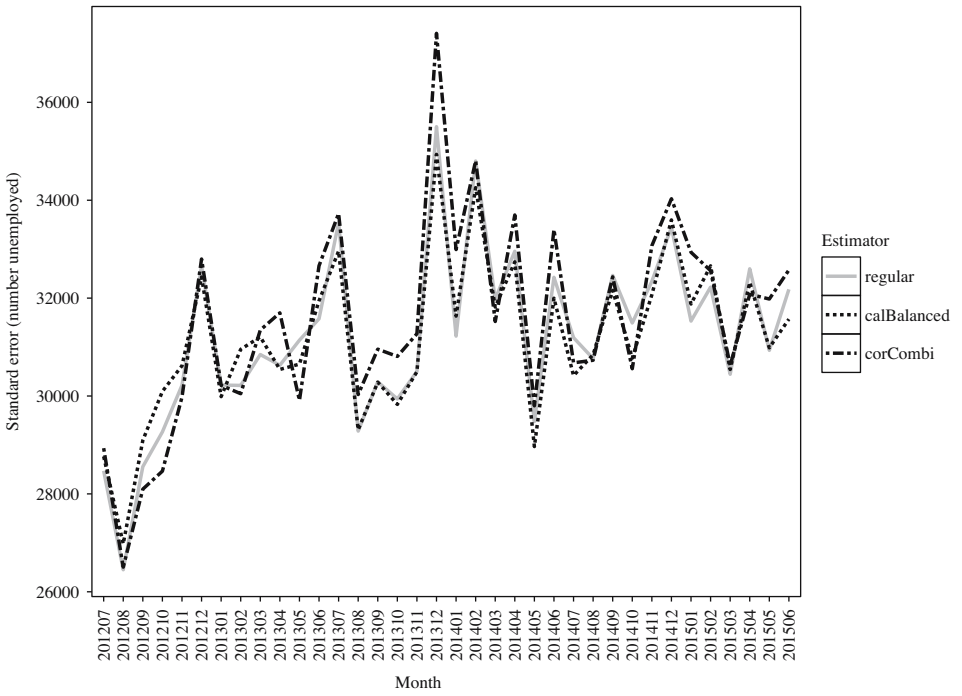

Fig. 7. Standard errors of estimates of the total number of unemployed obtained through the regular, calibration, and correction approaches.

approaches observed in the application can be explained by the fact that the underlying assumption that the auxiliary variables in the GREG estimator apart from the response mode do not completely correct for selective nonresponse. The fact that both almost coincide with the original series is specific to the case at hand, and is due to the relative insensitivity of the results to the realized variations in the mix of survey modes in the LFS. In this specific case, there is no pressing need to apply any of the two methods. However, since there are no adverse effects of the methods, it might be desirable to apply one of the methods nevertheless, as a protective measure against potential future instabilities in the mode composition.

In survey statistics where change over time is strongly confounded with changes in survey mode composition, the calibration and correction methods have a stabilizing effect. This is the case specifically for survey variables that suffer from large mode-dependent measurement effects, such as attitudes or answers to questions susceptible to social-desirability bias. An example where the mode composition varies extremely is the Crime Victimization Survey in the Netherlands, discussed in Buelens and Van den Brakel (2015).

5. Discussion

Estimates from repeated mixed-mode sample surveys can be unstable when the mode composition of the response varies over time. Two recently proposed methods of inference are compared in the present article. The calibration method adjusts the survey weights to

balance the response with respect to the survey modes, while the correction approach adjusts measurements using predicted counterfactuals. While motivated differently, it is shown that both estimators are equal if the mixing parameters for the combined measurement error correction approach mirror the mode distribution assumed for the mode calibration estimator; the remaining auxiliary variables of the weighting schemes of both estimators must be equal too. The two methods rely on the following assumptions (Buelens and Van den Brakel 2015):

i) the weighting model removes mode-dependent selectivity with respect to the survey variables;
ii) time-indepence of the measurement error model;
iii) constant population size – only required when estimating population totals.

When (iii) does not hold, a residual measurement error bias remains, which is not affected by fluctuations in mode composition. Condition (iii) is not required for population means. Violations of (i)–(iii) will lead to biased estimates. It must be emphasized that this would be the case too in uni-mode designs employing a single mode of data collection. Some issues could be resolved by more advanced modeling, for example allowing for time-dependent measurement errors.

In the present research, thirty-six monthly editions of the Dutch LFS are used as a case study. Small deviations between both approaches are observed and can be explained by not meeting the underlying assumption that the auxiliary variables in the weighting model, apart from the mode distribution, completely correct for selective nonresponse. Both approaches produce similar standard errors for the unemployed labour force in the case that the mixing parameters for the combined measurement error correction approach resemble the distribution of the respondents over the modes observed in the sample. In the case of extreme distributions, where the contribution of one of the modes is suppressed, the differences in standard errors under the two approaches are large. The standard error of the mode calibration estimator increases rapidly with increasing discrepancies between the distribution in the sample and in the population. Under the measurement error correction approach, the standard errors increase only slightly, even when the outcomes are corrected to a single mode. The explanation for this difference between the two methods is that the measurement error correction estimator uses additional information by explicitly relying on Model (7) to correct the actual observations for a measurement error component. Unlike the calibration method, the measurement error correction method does not have a built-in protection against strong deviations of the sample and population distributions, unless the mixing coefficients are chosen by minimizing the MSE as proposed by Suzer-Gurtekin (2013), or by choosing them close to the observed mode distribution, as proposed in this article.

The results in Subsection 4.2 indicate that if the LFS were conducted by telephone and the same respondents were reached as currently with the mixed-mode strategy, the estimated average unemployed during the study period would drop from 678000 to 627000. Had the same respondents been interviewed face-to-face, the estimated average would have been 695000. It is a disconcerting thought that the true number of unemployed could be anywhere in this range, or even outside the range, as all three modes can be biased with only relative bias observable. This stresses the inadequacy of traditional measures of

uncertainty only taking into account the uncertainty due to random sampling. This issue is also present in single-mode surveys where it is not as manifestly visible as in mixed-mode surveys. Further research into quantifying measurement related uncertainty is important and could possibly follow the strand of research of the Total Survey Error paradigm, see, for example Groves and Lyberg (2010) for a review.

The observed differences between the three modes are in line with the results of a mode experiment with the LFS obtained in the past and can be partially explained with cognitive models of the survey process. Observed relative mode-effects are nevertheless empirical results and explaining their direction or making statements which mode can be used as a benchmark remains highly speculative. The two methods studied in this article are intended to stabilize the mode distribution in repeated surveys to avoid fluctuations in mode-dependent measurement bias obscuring measurements of change over time. As is the case in single mode surveys, mixed-mode surveys may measure at a level different from the true level in the population. As long as the level difference remains constant through time, change over time can be estimated without bias, both in single mode and mixed-mode surveys. It is recommended to choose the distribution for the mode calibration or the mixing proportions for the correction approach close to the observed distribution of the respondents over the modes in the samples. This avoids unnecessary increase of fluctuations in the weights and in the standard errors. The techniques applied in this article are practically useful as they do not require additional questions, questionnaires, or repeated interviewing.

6. References

Buelens, B. and J.A. Van den Brakel. 2015. "Measurement Error Calibration in Mixed-Mode Sample Surveys." *Sociological Methods & Research* 44(3): 391–426. Doi: http://dx.doi.org/10.1177/0049124114532444.

Cannel, C., P. Miller, and L. Oksenberg. 1981. "Research on Interviewing Techniques." In *Sociological Methodology*, edited by S. Leinhardt. 389–437. San Fransisco: Jossey-Bass.

Centraal Bureau voor de Statistiek. 2015. *Methoden en Definties Enquête Beroepsbevolking 2014*. Technical report, Statistics Nederlands, Heerlen. Available at: https://www.cbs.nl/NR/rdonlyres/1BB3C645-47CC-4F58-9031-89F490AEE981/0/methodenendefinitiesebb2014.pdf (accessed March 2017).

Cernat, A. 2015. "Impact of Mode Design on Measurement Errors and Estimates of Individual Change." *Survey Research Methods* 9(2): 83–99. Doi: http://dx.doi.org/10.18148/srm/2015.v9i2.5851.

De Leeuw, E. 2005. "To Mix or not to Mix data Collection Modes in Surveys." *Journal of Official Statistics* 21: 233–255.

Dillman, D., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B. Messer. 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. Doi: http://dx.doi.org/10.1016/j.ssresearch.2008.03.007.

Groves R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Holbrook, A., M. Green, and J. Krosnick. 2003. "Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires." *Public Opinion Quarterly* 67: 79–125. Doi: http://dx.doi.org/10.1086/346010.

Jäckle, A., C. Roberts, and P. Lynn. 2010. "Assessing the Effect of Data Collection Mode on Measurement." *International Statistical Review* 78: 3–20. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00102.x.

Klausch, T., J. Hox, and B. Schouten. 2015. "Selection Error in Single- and Mixed Mode Surveys of the Dutch General Population." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(4): 945–961. Doi: http://dx.doi.org/10.1111/rssa.12102.

Krosnick, J. 1991. "Response strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. Doi: http://dx.doi.org/10.1002/acp. 2350050305.

Krosnick, J. and D. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." *Public Opinion Quarterly* 51: 201–219. Doi: http://dx.doi.org/10.1086/269029.

Lynn, P. 2013. "Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs." *Journal of Survey Statistics and Methodology* 1(2): 183–205. Doi: http://dx.doi.org/10.1093/jssam/smt015.

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

Schouten, B., J. van den Brakel, B. Buelens, J. van der Laan, and T. Klausch. 2013. "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys." *Social Science Research* 42(6): 1555–1570. Doi: http://dx.doi.org/10.1016/j.ssresearch.2013.07.005.

Suzer-Gurtekin, Z.T. 2013. *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys*. Ph.D. thesis, University of Michigan. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/102471/tsuzer_1.pdf (accessed March 2017).

Suzer-Gurtekin, Z.T., S. Heeringa, and R. Vaillant. 2012. "Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys." In Proceedings of the JSM, Section on Survey Research Methods, San Diego, July 28–August 2, 2012. American Statistical Association, 4711–25.

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

Van den Brakel, J. 2008. "Design-Based Analysis of Embedded Experiments with Applications in the Dutch Labour Force Survey." *Journal of the Royal Statistical Society, Series A* 171: 581–613. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00532.x.

Van den Brakel, J.A. and S. Krieg. 2015. "Dealing with Small Sample Sizes, Rotation Group Bias and Discontinuities in a Rotating Panel Design." *Survey Methodology* 41(2): 267–296. Available at: http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-eng.pdf (accessed March 2017).

Vannieuwenhuyze, J.T.A. and G. Loosveldt. 2013. "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement

Effects." *Sociological Methods & Research* 42(1): 82–104. Doi: http://dx.doi.org/10.1177/0049124112464868.

Voogt, R. and W. Saris. 2005. "Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21: 367–387.

Received January 2016
Revised February 2017
Accepted March 2017