Journal of Official Statistics
vol. 39, 1 (March 2023)

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 1–26, http://dx.doi.org/10.2478/JOS-2023-0001

Characteristics of Respondents to Web-Based or Traditional Interviews in Mixed-Mode Surveys. Evidence from the Italian Permanent Population Census

Elena Grimaccia[1], *Alessia Naccarato*[2], *Gerardo Gallo*[1], *Novella Cecconi*[1], *and Alessandro Fratoni*[1]

In order to provide useful tools for researchers in the design of actions to promote participation in web surveys, it is key to study the characteristics that define the profile of a "web respondent", so that specific interventions can be planned. In this contribution, which draws on data collected during the 2019 housing population census in Italy, we define the set of familial and geographical characteristics that correspond to a greater probability that the interviewed household will choose to respond online, by estimating a multilevel model. The profile of a "computer-assisted web interview household" (CAWI-H) is then defined, on the basis of the structural characteristics of this population. Moreover, the geographical distribution of households is studied according to their distance from the CAWI-H profile. The results show that households that are more distant from the CAWI-H profile have characteristics that correspond to segments of the population generally affected by economic and social fragility; they are mainly elderly, foreigners, residents in small towns, and people with a low level of education. It is to these households in particular that survey designers can address specific actions that can enhance their willingness to participate in web surveys.

Key words: Mixed-mode surveys; respondent profiling; multilevel models; computer assisted web interview (CAWI).

1. Introduction

In recent decades, developments in information technology and the ever-increasing availability of administrative data have led several European countries to develop innovative methods for their population censuses (Eurostat 2020).

Mixed-mode surveys have been adopted by necessity in survey practice (Biemer 2010; De Leeuw 2005) and are, nowadays, also essential for National Statistical Institutes (NSIs). According to Tourangeau (2017) their usage is only expected to increase over time.

Online surveys have become part of mixed-mode data collection strategies, since web data collection presents several advantages, like a reduction in cost, a general improvement in timeliness (since the higher the share of web respondents, the shorter the time devoted to data collection as in Dillman et al. 2014 and in De Leeuw 2005), and a potential to use complex questionnaires (De Leeuw and Berzelak 2016), and to reduce

[1] Istat, Italian National Institute for Statistics, via Cesare Balbo 16, 00184 Rome, Italy Emails: elgrimac@istat.it, gegallo@istat.it, ncecconi@istat.it, and fratoni@istat.it
[2] Department of Economics, Roma Tre University, via Silvio D'Amico 77, 00145 Rome, Italy. Email: alessia.naccarato@uniroma3.it
Acknowledgments: We would like to thank the anonymous reviewers whose suggestions and comments helped improve and clarify this manuscript.

coverage and nonresponse error (Bianchi et al. 2017; Brick and Tourangeau 2017; Cobben and Bethlem 2013; De Leeuw 2018; De Leeuw et al. 2019; Luiten et al. 2020).

In particular, sequential mixed-mode strategies (which offer one mode at a time, starting with the cheapest and quickest) improve response rates when different modes acquire different types of respondents: typically, younger respondents respond through the web, and respondents who are older or first-generation immigrants respond in interviewer administered modes (Benzeval et al. 2021; De Leeuw and Berzelak 2016; Kappelhof 2015). Indeed, including a mail survey for those who do not have internet access and as a follow-up for web nonrespondents also improves representativity on demographic and attitudinal variables (Bandilla et al. 2014; Messer and Dillman 2011).

However, there are some disadvantages because of the comparability of data collected via different modes, which could lead to different measurement errors. For instance, past research has shown that self-administered forms (e.g., both paper mail surveys and web surveys) perform better when more sensitive questions are asked allowing less social desirable answers than interview surveys (De Leeuw and Berzelak 2016; De Leeuw et al. 2008, 299).

In this framework, no one could have predicted that the coronavirus pandemic emergency and the impossibility of carrying out field operations would have boosted the need for web-based interviews with such strength and urgency.

However, even if the use of the internet is quite widespread in Europe (Eurostat 2020), respondents' attitudes toward web surveys cannot be taken for granted. Therefore, planning interventions targeting specific groups of individuals or households can support wider participation in web surveys.

The analysis of auxiliary variables in mixed-mode surveys has certainly become much more important (De Leeuw 2018; De Leeuw et al. 2019), both because of the digital divide in the population, with highly educated and younger people more often having an internet connection (Biffignandi and Bethlehem 2012; Couper 2017; Mohorko et al. 2013), and because the interaction between the characteristics of the respondents and the decisions of the survey designer influences the response rate and the success of the survey (Antoun et al. 2017; Biffignandi and Pratesi 2002; Durrant and Steele 2009; Scherpenzeel and Bethlehem 2011, 105).

In this article, we study the determinants that influence the cooperation of respondents with a sequential mixed-mode survey, such as the Italian Permanent Population and Housing Census (PPHC), in order to point out the specific characteristics of the population. This could enhance the efficacy of the actions of survey designers to improve participation in web surveys.

By estimating a multilevel logit model (Durrant and Steele 2009; Mohorko et al. 2013), we distinguish between the features of online respondents to a Computer Assisted Web Interview (CAWI) and the features of respondents in traditional modes (in the case of the PPHC, a Computer Assisted Personal Interview, CAPI), given that respondents with different characteristics tend to have different propensities for being interviewed in a given mode.

We consider the set of the socio-economic characteristics of a respondent as their "profile". The profiling of respondents, in the literature, has traditionally been achieved through a logistic model. Among others, Maslovskaya et al. (2019) studied the effect of different response tools in six social surveys in the UK. The results of their bivariate analysis suggest that, across the surveys, age, gender, marital status, employment, religion,

household size, children in household, household income, number of cars and frequency of internet use are significantly associated with the device used. Bianchi et al. (2017) studied the effects of mixed-mode design on participation rate, sample composition, and costs in a longitudinal survey in the UK. To investigate whether the mixed-mode design had different effects for different characteristics of the respondents (such as gender, age, race, working status, household type, urbanicity, web-user), the authors estimated a logit model employing individual characteristics and interactions, but this did not produce significant results. Pratesi et al. (2004) focused on the timeliness and quality of web surveys in Slovene households. Their findings suggest that nonresponse rates and quick reactions in web surveys are sensitive to individuals' characteristics, but that demographics are not significant in this. Methodological issues and the research perspectives of web surveys have been studied in depth and in a thorough and comprehensive way by Biffignandi and Bethlehem in 2012 and, more recently, in 2021.

The analysis presented here, then, explores the significant differences, in terms of individual and familial characteristics, between households who respond by completing a CAWI and those who respond by the face-to-face method. In this way, it is possible to define the profile of CAWI respondents (which we call the CAWI-H profile) to identify which household characteristics contribute the most to increasing the probability of responding via the web. As in previous research (Durrant and Steel 2009; Mohorko et al. 2013), this study employs multilevel modelling alongside logistic regression, and it provides also some evidence that a multilevel model is better than a logit model at explaining the probability of answering via the web.

Once the profiles of web respondents have been defined, it is possible to measure how far the features of CAPI respondents deviate from the CAWI-H profile, establishing a measure of distance that allows us to classify respondents according to their "resistance" to web completion. Compared to other studies employing propensity scores, the computation of distances from the profiles identified by the multilevel model has been preferred since it could be successfully employed for other surveys and used as a tool for planning specific survey interventions.

This study identifies the territorial features related to web responses providing specific tools for promoting web surveys and, in particular, addressing the logistical issues. Compared to previous studies concerned with Internet usage in general, this study provides insight on web survey.

The article is organised as follows: Section 2 presents the data collection design for the PPHC; Section 3 describes the variables employed to define the respondents' profiles; Section 4 presents the statistical methodology to obtain the respondents' profiles and to measure the distances between them; Section 5 presents the main outcomes of the study and the territorial distribution of the different types of respondents; and, finally, Section 6 presents some concluding remarks with reference to particular population groups that could be made the subject of actions aimed at improving the CAWI response.

2. Italian Population and Housing Census Data Collection Design

This article draws on data from the 2019 PPHC. In Italy, the PPHC, which was started in 2018 by the Italian National Institute of Statistics (Istat), currently provides a mixed-mode

survey, since respondents may choose to fill in the questionnaire via the web or by the traditional way, according to the data collection design described in Figure 1.

In Figure 1, the flowchart for the data collection process is presented. The current progression of the actions for data collection are represented in white, while our proposal to increase the share of web respondents is shown in grey.

The population census provides the official estimate of the resident population at the municipality level (Righi et al. 2021). In Italy, as in most European countries, an assessment of the coverage results of the last population census in 2011 and the high costs of data collection led to a change in the strategy for the population census from the traditional door-to-door enumeration (every ten years for the whole population) to a yearly register-based sample survey, combining the use of administrative sources with annual surveys (Chieppa et al. 2018; Citro 2014; Crescenzi 2015; Righi et al. 2021). A further advantage of these data collection methodologies is the greater containment of the census participation burden: as noted by UNECE (2018), the response burden on the population is lower with a combined census, both because the number of questions is reduced (since some information is available from the population register), and because the physical presence of a stranger in people's homes is avoided (which could be a reason for a refusal to participate, especially in a pandemic situation).

The PPHC foresees two different yearly sample surveys: an Areal sample and a List sample. The first is only used to update the population register and it is only conducted face-to-face; it is not of interest for a comparison of different data collection modes. Therefore, we analyse the List sample, which presents a most interesting survey design and is carried out in order to collect information on socio-economic issues that is not available in the population registers. The List sample is based on a yearly sample size of about 950,000 households (Table 1). Around 2,400 municipalities (of the 7,904 municipalities in Italy) are involved in the survey every year.

The List survey design is "sequential", and respondents may choose the most suitable mode to respond to the questionnaire. The first survey mode proposed by the survey researcher is the CAWI mode, since it is the cheapest and quickest. Helplines are provided to the respondents in case they need more information or specific support in answering to the survey. If the CAWI is not answered, respondents are contacted, in order to recall the interview and to offer help in the filling of the questionnaire. However, as a second option, respondents can go in person to the many offices that are present in each municipality and that are devoted to help respondents, and to conduct a "face to face" interview. Finally, as a last option, it is foreseen also the traditional interviewer administered survey at home, if required by the respondent for any reason (Istat 2018). The share of CAWI responses is therefore unknown "a priori".

Data are collected at the household level, as a unique questionnaire is provided to collect the data. Households included in the List sample receive an official letter from Istat, giving information about the importance of the survey and the fact that answering is mandatory by law, and providing the credentials for answering the census questionnaire online. The letter is addressed to the oldest member of the household, who is assumed to be the one who will answer the questionnaire, providing information on the whole household. The questionnaire is written in Italian, but a complete guide is provided in 14 foreign languages.

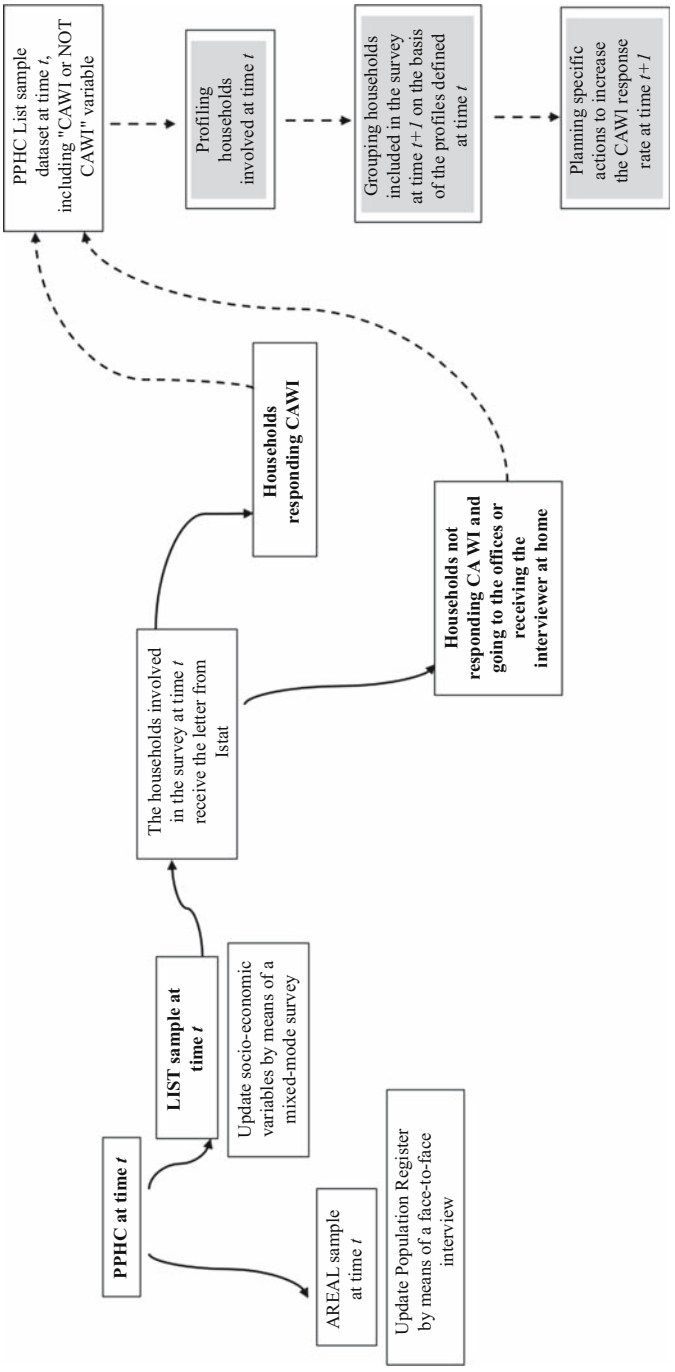


Fig. 1. Italian population and housing census data collection design.

Table 1. Sample size – municipalities and households by population size of municipalities – 2019.

Demographic size of Italian municipalities (number of inhabitants)	Municipalities	Households
Up to 5,000	1,028	190,458
From 5,001 to 20,000	815	288,360
From 20,001 to 50,000	376	184,593
From 50,001 to 100,000	98	120,149
From 100,001 to 250,000	34	94,226
More than 250,000	12	73,253
Total	**2,363**	**951,039**

Source: 2019 PPHC-Istat

It is only if the household cannot fill in the CAWI that they can answer in a face-to-face interview that must take place in a public office responsible for data collection.

Under the auspices of NSIs, this share should be close to 100%. In Italy, the use of the internet is quite widespread, with 76% of households having internet access and 75% having a broadband connection in 2019. However, there is still a large digital divide between households, mainly due to generational and cultural factors. Almost all households with at least one member younger than 18 years old have a broadband connection (95%), while among households composed exclusively of people over the age of 65, this share falls to 34% (Istat 2019). Since the share of CAWI respondents is stuck at 50% of the population, the diffusion of the internet would suggest the possibility of a much higher share of CAWI responses, provided that targeted actions such as those identified in this study are conducted on the profiles of respondents.

Although the PPHC is not a panel, since households are included in the sample only once, the profiling of web respondents could be of great support in the successive data collection occasions, because the survey is repeated over time with the same design.

3. Data

The PPHC List sample, for each household, provides information on the survey mode, the socio-demographic characteristics of the household, and the geographical features of the municipality where the household resides.

The socio-demographic information available on households is household size, household citizenship, age of youngest household member, and highest educational level in the household. The household size indicates the number of people residing in the same dwelling. The citizenship variable classifies the households based on the citizenship of the members: all foreigners, all Italians, or a mixed citizenship household. The youngest household-member age variable indicates the age of the youngest member of the household; this variable was considered in the analysis on the basis of the hypothesis that the presence of young individuals in a household facilitates the use of information technology and therefore has an effect on the decision to fill in the questionnaire online. Minors are not included in the definition of this variable since the census questionnaire can only be filled in for members who are over 18 years of age. The highest educational qualification held by at least one member of the household is classified according to the

primary, secondary and tertiary levels that group together the grades (0, 1, 2), (3, 4), and (5, 6, 7, 8) of the International Standard Classification of Education (ISCED).

All these features are coherent with the predictors accounted for in previous studies. Besides the literature cited in the Introduction, it is worth mentioning the work of Hargittai (2002), which is one of the first studies on the profiling of online skills, employing personal and familial features, such age and education. Similar predictors were chosen by Durrant and Steele (2009).

Furthermore, in order to take into account the strong territorial disparities, both in socio-economic issues and in digitalization (Benassi and Naccarato 2017; Cellini and Torrisi 2014; Cracolici et al. 2007; De Clercq et al. 2020; Santarelli and Cottone 2009), several characteristics of municipalities are considered in the analysis: local capital or metropolitan city, municipality population size, municipality degree of urbanisation, and altitudinal zone. Local capitals or metropolitan cities are those municipalities which are the seats of their boroughs. The variable 'municipality population size' classifies municipalities on the basis of the number of their inhabitants, from very small municipalities (up to 5,000 inhabitants) to very large ones (more than 250,000 inhabitants), while the degree of urbanisation is a classification of municipalities introduced by Eurostat and based on a criterion of geographical contiguity and on minimum population thresholds (European Union 2017). The territory is classified as one of three types of area: (1) a densely populated area (city or large urban area), defined as clusters of contiguous cells of 1 km^2, with a density of not less than 1,500 inhabitants per km^2 and a population of not less than 50,000 inhabitants; (2) an area with an intermediate density level (or small urban area), defined as clusters of contiguous cells with a density of not less than 300 inhabitants per km^2 and a population of not less than 5,000 inhabitants; and (3) a sparsely populated area (or rural area), defined as single cells (rural) not classified in the previous groups (Eurostat 2021). In order to attribute this classification to single municipalities, the areas identified according to the degree of urbanisation are compared with the municipal boundaries. The altitudinal zone defines homogeneous areas constituted by aggregating contiguous municipalities on the basis of threshold values for the altitude (elevation above sea level). According to this classification, mountain, hill, and plain areas can be identified. The mountain and hill areas are divided into inland mountain and inland hill areas and coastal mountain and coastal hill areas, respectively, in order to take account of the moderating action of the sea on climate. The altitudinal zone is taken into account in the analysis since it could affect the choice to answer online due to the different internet access capacity of mountainous areas, which presents some specificities regardless of the population density and the composition of population (Reynaud et al. 2020).

Municipalities can be grouped into regions, which are very important territorial units that are responsible for various economic and social policies, or into macro-regions: the North, the Centre, and the southern areas of Italy plus the two major islands, together traditionally referred to as the "Mezzogiorno".

Table 2 shows descriptive statistics on the frequency distribution of the households according to the CAWI or not CAWI survey mode and the explanatory variables included in the analysis (Section 5). The share of CAWI respondents varies according to the household citizenship, size, age of the youngest member, and highest education level in the household. The size and sign of the relationships are presented (when significant) in Subsection 5.1. At this point of the study, however, it is worth mentioning that for instance

Table 2. Households by survey mode, and familial and social characteristics[(a)].

Household characteristics		Survey mode: Not CAWI (%)	Survey mode: CAWI (%)	Composition of the CAWI sample (%)	Number of cases
Household size	1	54.01	45.99	27.4	245155
	2	50.98	49.02	28.6	239632
	3	45.67	54.33	21.6	163624
	4	43.93	56.07	17.6	128977
	5 or more	56.16	43.84	4.8	25294
	Pearson Chi2(4)=5.50E+03; Pr=0.000				
Household citizenship	All foreigners	75.27	24.73	1.7	28296
	All Italians	48.86	51.14	95.9	770700
	Mixed citizenship	57.29	42.71	2.4	23431
	Pearson Chi2(2)=8.10E+03; Pr=0.000				
Youngest household-member age	18-34	46.98	53.02	31.4	243631
	35-64	47.23	52.77	45.8	356701
	65+	57.78	42.22	22.8	222095
	Pearson Chi2(2) = 7.40 + E03; Pr = 0.000				
Household highest educational level	Primary	65.03	34.97	24.9	293213
	Secondary	46.59	53.41	44.1	339277
	Tertiary	32.90	67.10	31.0	189937
	Pearson Chi2(2)=5.00E+04; Pr=0.000				
Total		50.01	49.99	100.0	822427

Source: 2019 PPHC-Istat

[(a)]There are no missing cases in the data set for the variables used here.

the share of CAWI respondents increases with higher level of education. The Chi-square tests referring to these distributions preliminarily suggest that there are relationships between the type of response and the variables used to define profiles.

4. Methods

The first step of the research is to identify the profile of the online respondent households (CAWI-H) and, more specifically, to identify the variables that are associated with the households with the highest probability of responding via the web.

To compute the probability that a family responds in CAWI mode, we estimate a multilevel logit model that also takes into account territorial specifications (Goldstein 2010). The endogenous variable is the dichotomous variable that assumes the value 1 if the family responded in CAWI mode, and 0 otherwise (Durrant and Steele 2009; Keusch et al. 2019; Mohorko et al. 2013), the binary response Y_i defined as follows:

$$Y_i = \begin{cases} 0 & Not\ CAWI\ response \\ 1 & CAWI\ response \end{cases} \quad (1)$$

where i ($i = 1, \ldots, n$) denotes the household.

The auxiliary variables are the social and demographic characteristics of the households illustrated in Section 3.

The multilevel model can be written as:

$$log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(0)}}\right) = \boldsymbol{\beta}^{(s)}\boldsymbol{x}_{ij}^{(s)} + u_j^{(s)} \tag{2}$$

where j ($j = 1, \ldots k$) indicates the territorial level, $\pi_{ij}^{(s)} = \Pr(y_{ij} = s)$, $s = \{^{Not\ CAWI}_{CAWI}$, $\boldsymbol{x}_{ij}^{(s)}$ is a vector of household level covariates, $\boldsymbol{\beta}^{(s)}$ is a vector of coefficients, and $u_j^{(s)}$ is a random effect representing unobserved regional characteristics.

By means of the $u_j^{(s)}$ component, the multilevel model (2) takes into account the relationship between a CAWI response and the region where the household resides. In this way, the model acknowledges the unobserved regional influences on the different types of response, and the estimates of the coefficients referring to the explanatory variables are more accurate. The use of random effects is justified by the fact that the respondent living in a region is a feature that does not change over the period accounted for in the model.

As a further check on the robustness of the results and to illustrate their stability to various model specifications, we test whether a model including region fixed effects fits the data better than the multilevel model. The estimated logistic regression with binary response Y_i in Equation (1) (Wooldridge 2012) is:

$$logit(g_i) = \log\left(\frac{g_i}{1 - g_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots \tag{3}$$

in which the probabilities $g_i = \Pr(Y_i = 1|x_i)$ are related to a linear predictor $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots$ through the logit function.

In Appendix (Section 7), we present the results of the logit model with different specifications, including joint effects of the explanatory variables.

In this first phase of the research, the profile of the households with the highest probability of answering via the web is identified. All in all, the number of possible combinations of categories (i.e., profiles) is 135, since there are four exogenous variables used in the model (2): household size, with five categories; household citizenship, with three categories; youngest household-member age, with three categories; and household highest educational level, with three categories (Table 2). Among the 135 profiles, we give the name 'CAWI-H' to the profile that presents the highest probability of answering via CAWI, that is to say, the combination of the categories that presents the highest values of the multilevel model coefficients. Each household, then, presents a certain profile because it differs from the CAWI-H profile for one or more categories. Therefore, it is possible to establish a measure of the distance between the profile of a household and the CAWI-H profile, and to compute this distance for each household.

The maximum possible value for this distance is 4 (if all the explanatory variables lie in a category that is different from the CAWI-H profile), while the minimum value is equal to 0 if the household presents the CAWI-H profile. The higher the value of the distance, the lower the probability that the specific family type will respond to on the web.

Since the variables that define the profiles are all categorical, the measure employed is the Jaccard distance (Jaccard 1908; McCormick et al. 1992), defined as follows:

$$D_i = \frac{M_{c,i}}{M_1 + M_2 + M_3 + M_4} \tag{4}$$

where i corresponds to a single household, $M_k, k = 1, \ldots, 4$ is the number of categories for each categorical variable, and $M_{c,i}$ is the number of categories for which the CAWI-H profile and the profile of household i are different. For instance, if household i has the same size and age of the youngest member as CAWI-H, but different citizenship and level of education, then $M_{c,i}$ is equal to 2.

To obtain a measure that varies between 0 and 1, the relative distance D_{Ri} is calculated by dividing the value D_i in equation [4] by the maximum value.

Households with the same value of the distance D_{Ri} have similar characteristics and, thus, the same probability of answering by a CAWI.

One of the most common methods for controlling for differential nonresponse in mixed-mode surveys is to use the estimated propensity of a respondent to participate in each mode. These propensity scores are typically estimated from a generalised linear model (e.g., probit, logit), where a given mode is treated as a possible outcome, conditioned on available covariates information, often limited to participant demographics, as is done by Hox et al. (2015), and more recently by Maslovskaya et al. (2019) and Rivero et al. (2019). However, in this article, we propose a procedure that would allow survey design researchers to control the distance from the CAWI-H profile in the actual data collection and also before proceeding to a new phase of data collection, while propensity scores can only be estimated after collecting the data.

In order to plan interventions for the promotion of web surveys in different territories, it is useful to study where households present the strongest 'resistance' to CAWI and the characteristics of the areas of the country in which it is more important to intervene. The third and final step of the article is then to create and analyse a measure of 'Municipality Resistance' to CAWI data collection. The Municipality Resistance (MR) is computed as the average of the distances (D_{Ri}) for all the households residing in a given municipality. In this way, this average can be considered as the score assigned to the particular municipality. Using the MR as an endogenous variable in a regression model in which the covariates are the geographical variables (Section 3), it is possible to identify the characteristics of municipalities that are predictors of the MR. The characteristics of the territories employed in the analysis are all variables associated with the penetration of the internet (Ciapanna and Roma 2020; Rados 2021).

5. Results

The results provided by the multilevel model are presented in Subsection 5.1. In Subsection 5.2, the distribution of households, based on their distance from the CAWI-H profile, is illustrated. Finally, in Subsection 5.3, the estimation of a model that identifies the geographical predictors of municipalities' ability to collect responses via the web is presented.

5.1. Determinants of a CAWI Response

As illustrated in Section 4, a multilevel binomial model is estimated to explore the effects of household characteristics on the probability of CAWI or not CAWI responses. More specifically, the CAWI-H profile is a household whose members are all Italian, with at least one member with a tertiary level of education, in which the youngest member is between 35 and 64 years old, and, to a lesser extent, having four members (Table 3).

The highest coefficient for a CAWI response refers to households in which the level of education is highest. For people with a secondary or tertiary education, the coefficients referring to the probability of answering via a CAWI method are significantly higher than for those who have a primary level (respectively 0.734 and 1.350).

Moreover, households exclusively composed of foreigners present a lower probability of using the CAWI option than those with at least one Italian member. In particular, households whose members are all Italian present a coefficient equal to 1.379.

The empirical evidence also suggests that the CAWI response rate of households where the age of the youngest member is between 35 and 64 years old is significantly higher than the rate of households where the age of the youngest member is between 18 and 24 years old. Households composed only of elderly people present the lowest probability of a CAWI response.

Table 3. Estimate of parameters of multilevel logit model.

CAWI	Coefficient	Std. Err.	P > |z|
Household size (Base = 1)			
2	0.012	0.006	0.05
3	0.018*	0.008	0.02
4	0.112***	0.008	0.00
5 or more	-0.233***	0.012	0.00
Household citizenship (Base = All foreigners)			
All Italians	1.379***	0.015	0.00
Mixed	0.711***	0.020	0.00
Youngest household-member age (Base = 18–34 years old)			
35–64	0.050***	0.006	0.00
65+	-0.045***	0.008	0.00
Household highest educational level (Base = Primary)			
Secondary	0.734***	0.006	0.00
Tertiary	1.350***	0.007	0.00
Constant	-2.090***	0.016	0.00
Region			
var(_cons)	3.50E + 08	3548290	
LR test vs logistic model: Chibar2(01) = 40684.27, Prob > = Chibar2 = 0.0000			

Legend: * p < .05; ** p < .01; *** p < .001
Source: 2019 PPHC-Istat.

The analysis of the size of the household offers more equivocal results: the coefficient referring to a CAWI response is the same in households with up to three members, it increases for households with four members, while it significantly decreases in households with five or more members.

All in all, the survey designer should pay particular attention to households with foreigners, those without any members with at least a secondary-level education, and households whose members are all elderly.

Finally, the estimate indicates that the effect of education in choosing a CAWI prevails over the other determinants. The results are coherent with those of Keusch et al. (2019), Maslovskaya et al. (2019), and Mohorko et al. (2013). In fact, education in general been found to be among the determinants of response in surveys, with a higher level linked to a higher propensity to answer through the CAWI option (Hox et al. 2015).

It is worth noting that the results are coherent with those obtained by estimation of the logit model in Appendix.

We found that the likelihood-ratio test (LR), which evaluates the goodness of fit of two comparable models, justifies the use of a multilevel model (Table 3). Indeed, the result confirms the validity of the multilevel model, compared to the logistic regression, providing a significant coefficient (Maddala and Lahiri 2010).

5.2. *Households' Distance from the CAWI-H Profile*

Tables 4 and 5 show the distribution of households by survey mode and their relative distance D_{Ri}.

The large majority (69%) of households that present the CAWI-H profile ($D_{Ri} = 0$) do indeed answer through a web interview. By contrast, the households that are most distant from the identified CAWI-H profile ($D_{Ri} = 1$) typically answer via a traditional mode. The dependence between the distance from the CAWI-H profile and the survey mode is significant, according to the Chi-square test results (Table 4). Altogether, the probability of answering through a CAWI questionnaire decreases as the distance from the CAWI-H profile increases. Therefore, the results suggest that our models select useful categories that identify the characteristics of respondents with the highest probability of a CAWI response. In this way, we also recognise those profiles for which contact with the

Table 4. Household distribution by type of response and relative distance from CAWI-H profile.

Relative distance from CAWI-H	Survey mode		
	CAWI	NOT CAWI	Total
0	69.11	30.89	100
0.25	54.70	45.30	100
0.5	47.42	52.58	100
0.75	45.89	54.11	100
1	27.35	72.65	100
Total	49.99	50.01	100
Pearson Chi2(4)= 1.30E+04; Pr= 0.000			

Source: 2019 PPHC-Istat.

Table 5. Household distribution by survey mode, relative distance from CAWI-H, and macro-region.

		Survey mode		
Macro-region	Relative distance from CAWI-H	CAWI	NOT CAWI	Total
North	0	77.27	22.73	100
	0.25	63.25	36.75	100
	0.5	55.83	44.17	100
	0.75	58.61	41.39	100
	1	29.56	70.44	100
	Total	58.91	41.09	100
	Pearson Chi2(4) = 7.30E+03 Pr=0.000			
Centre	0	70.84	29.16	100
	0.25	58.46	41.54	100
	0.5	50.65	49.35	100
	0.75	51.40	48.60	100
	1	29.83	70.17	100
	Total	53.77	46.23	100
	Pearson Chi2(4) = 2.70E+03; Pr=0.000			
Mezzogiorno	0	56.83	43.17	100
	0.25	39.88	60.12	100
	0.5	35.06	64.94	100
	0.75	31.81	68.19	100
	1	16.86	83.14	100
	Total	36.55	63.45	100
	Pearson Chi2(4) = 3.90E+03; Pr=0.000			

Source: 2019 PPHC-Istat.

respondents (often under the control of the survey designer) like an e-mail solicitation plan, survey awareness, and so on, is most necessary and useful.

The study of the different profiles and their distance from the CAWI-H profile allows us to highlight two situations: those families for which a CAWI response is almost certain (0) or very unlikely (1), and those households that are on the border between being CAWI and being not CAWI. Considering only the households presenting the lowest relative distance from the CAWI-H profile (e.g., the first two groups in Table 4), the gain in the share of web respondents that it would be obtained if all these HHs would respond in a CAWI mode is 7.4 percentage points. Therefore, the share of CAWI responses would increase from 49.99 to 57.37%, if all the HHs with a distance from CAWI-H lower than 0.5 would actually respond via web.

The geographical effects that appear in the models to be significant for the probability of responding through a CAWI method (Table 3) become apparent in the analysis of CAWI-H distances (Table 5).

In the North and Centre of the country, the households that have a distance of 0.5 from the CAWI-H profile also have a 50% probability of answering through a web questionnaire. By contrast, in the Mezzogiorno the probability is much lower (35%). This means that even households with a CAWI-H profile have more difficulty in answering via the web. Moreover, in the Mezzogiorno area 43% of households with a CAWI-H profile do not choose to answer via the web. Therefore, the survey designer should pay more attention to the less developed areas.

This methodology can be applied to any country, since it is widely known that territorial differences, such as the urbanisation or geo-morphology of the territory, can be important predictors of digitalisation and other socio-economic features.

5.3. The Geographical Distribution of Municipalities with Web Respondents

The MR to CAWI mode increases when moving from the municipalities of the northern regions of Italy to the central areas and increases again when moving to the municipalities of the southern regions, as shown in Figure 2.

To verify the effects of the geographical characteristics of the municipalities on the MR variable, a generalised linear model (GLM) is estimated. Table 6 shows the results of the estimated model, where the endogenous variable is the MR, and the exogenous variables are the geographical variables (Section 4).

The demographic dimension of the municipality is important in explaining the MR to a CAWI response, and it is negatively related to the distance from the CAWI-H profile. Therefore, small centres with fewer than 20,000 inhabitants should be the focus of the survey designer's attention. In particular, local capitals or metropolitan cities present the highest probability of the CAWI-H profile. Indeed, internet users generally experience faster download and upload speeds in urban areas, although there is high variability in the internet coverage in particular territories (Rizzato 2020).

The results shown in Table 6 indicate that the geographical differences among the macro-regions are significant; thus, the model is also estimated separately for the North, the Centre, and the Mezzogiorno, in order to find out the specific details for each macro-region. In the Mezzogiorno, the MR is significantly higher (0. 0186).

The degree of urbanisation has no effect on the MR in the North and Centre of Italy, while in the Mezzogiorno small towns and rural areas are instead related to a smaller

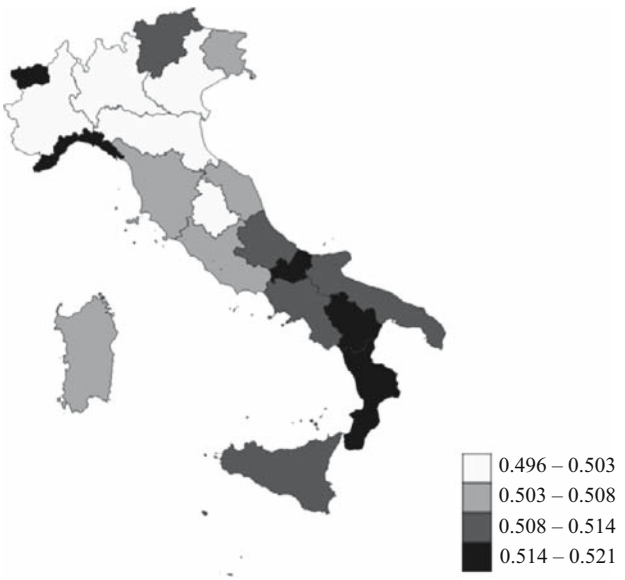


Fig. 2. Distribution of "Municipality Resistance" (MR – Regional Averages), Italy 2019.

Table 6. Estimate of parameters for generalised linear model and standard errors (in parentheses).

Municipality Resistance	Items	Italy	North	Centre	Mezzogiorno
Population size (Base = Less than 5,000 inhabitants)					
	From 5,001 to 20,000 inhabitants	0.0023	0.0013	-0.0017	.0073*
		(0.0015)	(0.0019)	(0.0038)	(0.0031)
	From 20,001 to 50,000 inhabitants	-.00437*	-.0104***	-0.0067	0.0048
		(0.0020)	(0.0028)	(0.0052)	(0.0038)
	From 50,001 to 100,000 inhabitants	-.00754*	-.0140*	-0.0128	-0.0020
		(0.0037)	(0.0064)	(0.0078)	(0.0058)
	From 100,001 to 250,000 inhabitants	-0.0115	-0.0166	0.0050	-0.0156
		(0.0064)	(0.0094)	(0.0160)	(0.0108)
	Over 250,000 inhabitants	-.01958*	-.0268*	-0.0400	0.0047
		(0.0088)	(0.0126)	(0.0211)	(0.0154)
Altimetric area (Base = Inland mountain)					
	Coastal mountain	-0.0038	-0.0186	0.0141	0.0029
		(0.0045)	(0.0128)	(0.0147)	(0.0055)
	Inland hill	.00374*	0.0029	0.0044	.0077**
		(0.0016)	(0.0022)	(0.0038)	(0.0029)
	Coastal hill	-.00422*	-.0218***	-0.0032	0.0004
		(0.0020)	(0.0043)	(0.0044)	(0.0031)
	Plain	.0114***	.0115***	0.0007	.0133***
		(0.0016)	(0.0019)	(0.0060)	(0.0033)
Municipality urbanisation degree (Base = Densely populated city or area)					
	Small town and suburb or intermediate population density zone	-.01047***	0.0015	0.0054	-.0274***
		(0.0025)	(0.0034)	(0.0125)	(0.0039)
	Rural area or sparsely populated area	-.00832**	0.0047	0.0029	-.0231***
		(0.0029)	(0.0037)	(0.0132)	(0.0050)
Local capital/Metropolitan city (Base = No)					
	Yes	-.0355***	-.0212***	-.0246**	-.0497***
		(0.0038)	(0.0061)	(0.0082)	(0.0061)
Macro-region (Base = North)					
	Centre	.0051**			
		(0.0017)			
	Mezzogiorno	.0186***			
		(0.0013)			
Constant		.4601***	.4504***	.4537***	.4869***
		(0.000)	(0.039)	(0.0138)	(0.054)
R^2		0.1114	0.168	0.1437	0.1983
Chi2		542.99	236.65	56.539	201.31
AIC		-10554	-5372.6	-1612.5	-3646

Legend: * $p < .05$; ** $p < .01$; *** $p < .001$
Source: 2019 PPHC-Istat.

distance from the CAWI-H profile. In the Mezzogiorno, the effect of being a local capital is maximal: living in one of the principal towns makes a real difference with respect to the probability of a CAWI response.

6. Final Remarks

For any statistical survey, web interviews imply considerable advantages: first of all, a reduction of costs, and then the containment of the interviewer effect (Bethlehem et al. 2011; Scherpenzeel 2011), and the timeliness of data collection. These advantages become even more significant when reference is made to official statistical surveys because of the large size of the samples and the number of variables surveyed, as well as the high standards that must be guaranteed.

The interest in identifying the profile of CAWI respondents is that web surveys still face some resistance from the population. In order to foster a positive attitude towards web interviewing, an awareness campaign is necessary, and this would be more effective if it is targeted at a specific population.

It is also necessary to adapt the data collection strategy to geographical imbalances. For instance, small centres, rural areas, and some remote and mountainous zones show the highest resistance to filling in census questionnaires online and should be made target of specific actions by the survey designer.

All over the world, NSIs employ population lists to design surveys, and to define samples. In these population or firms' registers, some characteristics are included, and the amount of information available in advance is further increasing because the huge development of registers and archives. The variables employed in this study are those available in the lists of population that are used before the survey. Therefore, these variables are known in advance and they can be employed by the survey researchers to design specific actions to enhance web responses before the actual data collection. This study thus proposes a procedure to better exploit this information that is available before the survey.

This contribution has shown how some structural characteristics of households allow us to classify them on the basis of an attitude: a preference for filling in the population census questionnaire online, and that this attitude refers to household that generally are classified as living in less fragile conditions (Benassi and Naccarato 2017; D'Ambrosio 2003).

In conclusion, families with a higher level of education, composed of Italian citizens, with at least one member younger than 65 years old, and who live in urban areas are those for whom the probability of answering via the web is highest; they are also those who historically live in better economic and social conditions. We can therefore conclude that, regardless of the advantages of a statistical survey, the probability of answering online can be considered as a further sign of the conditions of greater or lesser social disadvantage in which Italian families live.

Other than providing some new insights on the profiling of web respondents in mixed-mode surveys, this research offers a procedure for calculating distances between profiles that could be successfully employed for any other survey and used as a tool for planning specific survey interventions to enhance web survey participation.

The classification of households in profiles identifies more homogeneous population groups, and it could be useful in the design of successive survey occasions, with the aim of

reducing the variability of the estimates and, eventually, the sample size. Also, the correction of partial or total nonresponse data would benefit from the profiling of households, through the analysis of more homogeneous groups.

Nevertheless, further research could be devoted to an analysis of the stability over time of the profiles and their geographical distribution. Having identified territorial differences in the profiles, also identifying regional profiles could be an insight to further develop.

7. Appendix

Tables 7 and 8 show the results of a base model (M0) that takes into account only the socio-economic characteristics of the households as described in Section 4; a model (M1) that also considers the geographical macro-region variable; and models from model (M2) to (M6) that include the variable indicating the region of residence and the joint effects of the explanatory variables.

Table 7. Estimate of parameters of logit model (odds ratios and standard errors).

	M0		M1		M2		M3	
Variable	Odds ratios	SE	Odds ratios	SE	Odds ratios	SE	Odds ratios	SE
2	0.013*	.006	0.014*	.006	0.011	.006	-0.030	.019
3	-0.022**	.007	0.023**	.008	0.017*	.008	-0.047**	.017
4	0.009	.008	0.110***	.008	0.109***	.008	-0.050**	.018
5 or more	-0.349***	.012	-0.240***	.012	-0.237***	.012	-0.382***	.020
All Italians	1.158***	.014	1.380***	.015	1.383***	.015	1.384***	.015
Mixed citizenship	0.716***	.020	0.718***	.020	0.714***	.020	0.713***	.020
35–64	0.105***	.105	0.059***	.	0.051***	.006	-0.089***	.017
65+	0.030***	.030	-0.040***		-0.044***	.008	-0.120***	.017
Secondary	0.789***	.006	0.729***		0.736***	.006	0.735***	.006
Tertiary	1.363***	.007	1.341***		1.355***	.007	1.351***	.007
Centre			-0.262***					
Mezzogiorno			-0.995***					
Regional Fixed Effects								
Trentino Alto Adige	-		-		0.024	.018	0.022	.018
Lombardia	-		-		0.281***	.010	0.281***	.010
Piemonte	-		-		0.062***	.012	0.061***	.012
Friuli Venezia Giulia	-		-		0.038*	.015	0.037*	.015
Veneto	-		-		0.116***	.012	0.115***	.012
Liguria	-		-		-0.014	.015	-0.014	.015
Valle d'Aosta	-		-		-0.167***	.031	-0.168***	.031
Toscana	-		-		-0.060***	.011	-0.060***	.011
Marche	-		-		-0.224***	.014	-0.224***	.014

Table 7. Continued

	M0		M1		M2		M3	
Variable	Odds ratios	SE	Odds ratios	SE	Odds ratios	SE	Odds ratios	SE
Lazio	-		-		-0.221***	.013	-0.221***	.013
Umbria	-		-		-0.209***	.019	-0.209***	.019
Abruzzo	-		-		-0.527***	.015	-0.527***	.015
Campania	-		-		-1.049***	.012	-1.049***	.012
Sardinia	-		-		-0.482***	.014	-0.482***	.014
Molise	-		-		-0.924***	.024	-0.924***	.0235
Puglia	-		-		-0.660***	.012	-0.660***	.0122
Basilicata	-		-		-0.951***	.020	-0.951***	.0205
Sicilia	-		-		-1.046***	.012	-1.046***	.0115
Calabria	-		-		-1.337***	.015	-1.337***	.0151
2#35–64							0.089***	.021
2#65+							-0.005	.021
3#35–64							0.080***	.020
3#65+							-0.123*	.058
4#35–64							0.300***	.021
4#65+							-0.662**	.238
5 or more#35–64							0.294***	.028
5 or more#65+							-0.407	1.162
_cons	-1.785***	.016	-1.556***	.016	-1.667***	.018	-1.569***	.022
Pseudo R^2	0.054		0.086		0.091		0.091	
Chi2	61841		97970		1.00E + 05		1.00E + 05	
AIC	1.10E + 06		1.00E + 06		1.00E + 06		1.00E + 06	

Legend: * $p < .05$; ** $p < .01$; *** $p < .001$

Source: 2019 PPHC-Istat.

Table 8. Estimate of parameters of logit model (odds ratios and standard errors).

	M4		M5		M6	
Variable	Odds ratios	Standard Errors	Odds ratios	SE	Odds ratios	SE
2	0.031***	.006	0.032***	.006	-0.024	.020
3	0.014	.008	0.014	.008	-0.046	.018
4	0.097***	.008	0.096***	.008	-0.056	.018
5 or more	-0.240***	.012	-0.241***	.012	-0.378	.020
All Italians	1.361***	.015	0.935***	.027	0.940	.027
Mixed citizenship	0.695***	.020	0.438***	.042	0.439	.042
35–64	0.105***	.015	0.137***	.015	0.022	.022
65+	0.266***	.015	0.315***	.015	0.237	.021
Secondary	0.945***	.014	0.400***	.034	0.410	.034
Tertiary	1.610***	.015	1.138***	.041	1.140	.041
Regional Fixed Effects						
Trentino Alto Adige	0.022	.018	0.022	.018	0.020	.0181917
Lombardia	0.283***	.010	0.283***	.010	0.282	.0103586
Piemonte	0.065***	.012	0.065***	.012	0.064	.0118956
Friuli Venezia Giulia	0.039*	.015	0.039*	.015	0.038	.0151733
Veneto	0.113***	.012	0.113***	.0117817	0.112	.0117847
Liguria	-0.005	.0152751	-0.005	.0152751	-0.005	.0152787
Valle d'Aosta	-0.158***	.0316612	-0.158***	.0316612	-0.158	.0316687
Toscana	-0.057***	.011311	-0.057***	.011311	-0.050	.0113138
Marche	-0.229***	.0140247	-0.229***	.0140247	-0.220	.0140282
Lazio	-0.219***	.012665	-0.219***	.012665	-0.219	.0126678
Umbria	-0.211***	.0190954	-0.211***	.0190954	-0.211	.0191004
Abruzzo	-0.530***	.0154521	-0.532***	.0154521	-0.531	.0154557
Campania	-1.047***	.0124042	-1.048***	.0124042	-1.040	.012407
Sardinia	-0.470***	.0143121	-0.468***	.0143121	-0.467	.014

Table 8. Continued

	M4		M5		M6	
Variable	Odds ratios	Standard Errors	Odds ratios	SE	Odds ratios	SE
Molise	-0.929***	.0235753	-0.931***	.0235753	-0.931	.023
Puglia	-0.656***	.0122558	-0.656***	.0122558	-0,655	.012
Basilicata	-0.957***	.0205521	-0.959***	.0205521	-0.959	.020
Sicilia	-1.040***	.0115886	-1.040***	.0115886	-1.040	.011
Calabria	-1.336***	.0151669	-1.338***	.0151669	-1.330	.015
2#35–64					0.089	.0217572
2#65+					0.030	.0216346
3#35–64					0.060	.0203841
3#65+					-0.046	.0571932
4#35–64					0.274	.0212905
4#65+					-0.612	.2352247
5 or more#35–64					0.260	.0279282
5 or more#65+					-0.317	1.167.796
Secondary#35–64	-0.002	.017	-0.036*	.0168228	-0.050	.0168708
Secondary #65+	-0.584***	.018	-0.643***	.0183936	-0.644	.0184273
Tertiary#35–64	-0.084***	.018	-0.116***	.0184085	-0.130	.0184731
Tertiary #65+	-.72548***	.022	-.77502***	.0225984	-0,776	.0226243
Secondary#Mixed citizenship			0.607***	.0340457	0.600	.0340634
Secondary#All Italians			0.356***	.05087	0.354	.0508955
Tertiary#Mixed citizenship			0.523***	.0401673	0.519	.0401839
Tertiary#All Italians			0.400***	.0573441	0.399	.0573705
_cons	-1.845***	.021	-1.469***	.029	-1.380	.031
Pseudo R^2	0.093		0.094		0.093	
Chi2	1.10E+05		1.10E+05		1.00E+05	
AIC	1.00E+06		1.00E+06		1.00E+06	

Legend: * $p < .05$; ** $p < .01$; *** $p < .001$
Source: 2019 PPHC-Istat.

8. References

Antoun, C., M.P. Couper, and F G. Conrad. 2017. "Effects of Mobile versus PC Web on Survey Response Quality: A Crossover Experiment in a Probability Web Panel." *Public Opinion Quarterly* 81: 280–306. DOI: https://doi.org/10.1093/poq/nfw088.

Bandilla, W., M.P. Couper, and L. Kaczmirek. 2014. "The Effectiveness of Mailed Invitations for Web Surveys and the Representativeness of Mixed-Mode versus Internet-only." *Survey Practice* 7(4). DOI: https://doi.org/10.29115/SP-2014-0020.

Benassi, F., and A. Naccarato. 2017. "Households in Potential Economic Distress. A Geographically Weighted Regression Model for Italy, 2001–2011." *Spatial Statistics* 21: 362–376. DOI: https://doi.org/10.1016/j.spasta.2017.03.002.

Benzeval, M., J. Burton, T.F. Crossley, P. Fisher, C. Gardiner, A. Jäckle, and J. Moore. 2021. "High Frequency Online Data Collection in an Annual Household Panel Study: Some Evidence on Bias Prevention and Bias Adjustment." *Understanding Society Working Paper Series* 03: 1–26. Available at: https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2021-03.pdf (accessed September 2021).

Bethlehem, I., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. New York: Wiley.

Bianchi, A., S. Biffignandi, and P. Lynn. 2017. "Web-face-to-face Mixed-mode Design in a Longitudinal Survey: Effects on Participation Rate, Sample Composition, and Costs." *Journal of Official Statistics* 33(2): 385–408. DOI: http://dx.doi.org/10.1515/JOS-2017-0019.

Biemer P.P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74: 817–848. DOI: https://doi.org/10.1093/poq/nfq058.

Biffignandi, S. and J.G. Bethlehem. 2012. "Web Surveys: Methodological Problems and Research Perspectives." In *Advanced Statistical Methods for the Analysis of Large Data sets,* edited by A. Di Ciaccio, M. Coli, and J.M.A. Ibanez: 363–374. Berlin: Springer.

Biffignandi S., and J. Bethlehem. 2021 *Handbook on Web Surveys.* Second edition. New York: Wiley.

Biffignandi, S. and M. Pratesi. 2002. "Modelling the Respondents' Profile in a Web Survey on Firms in Italy." In *Developments in Social Science Methodology,* edited by A. Ferligoj and A. Mrvar: 171–185. Ljubljana: FDV.

Brick, J.M., and R. Tourangeau. 2017. "Responsive Survey Designs for Reducing Nonresponse Bias." *Journal of Official Statistics* 33(3): 735–752. DOI: https://doi.org/10.1515/jos-2017-0034.

Cellini R., and G. Torrisi. 2014. "Regional Resilience in Italy: A Very Long-run Analysis." *Regional Studies* 48(11): 1779–1796. DOI: https://doi.org/10.1080/00343404.2013.861058.

Chieppa, A., G. Gallo, V. Tomeo, F. Borrelli, and S. di Domenico. 2018. "Knowledge Discovery for Inferring the Usually Resident Population from Administrative Registers." *Mathematical Population Studies* 26(2): 96–102. DOI: https://doi.org/10.1080/08898480.2017.1418114.

Ciapanna, E. and G. Roma. 2020. *Connected Italy*. Rome: Bank of Italy. (Questioni di Economia e Finanza. Occasional Papers No. 573). Available at: https://www.banca-

ditalia.it/pubblicazioni/qef/2020-0573/QEF_573_20.pdf?language_id = 1 (accessed September 2021).

Citro, C.F. 2014. "From Multiple Modes for Surveys to Multiple Data Sources for Estimates". *Survey Methodology* 40(2): 137–161. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf?st = x9gGl05C (accessed September 2021).

Cobben, F., and J.G. Bethlehem. 2013. *Web Panels for Official Statistics.* The Hague: Statistics Netherland. (Discussion Paper 201307). Available at: https://www.cbs.nl › 2013 › 2013-07-x10-pub (accessed September 2021).

Couper M.P. 2017. "New Developments in Survey Data Collection." *Annual Review of Sociology* 43: 121–145. DOI: https://doi.org/10.1146/annurev-soc-060116-053613.

Cracolici M.F., M. Cuffaro and P. Nijkamp. 2007. "Geographical Distribution of Unemployment: An Analysis of Provincial Differences in Italy." *Growth and Change* 38(4): 649–670. DOI: https://doi.org/10.1111/j.1468-2257.2007.00391.x.

Crescenzi, F. 2015. "Changing census by combining administrative sources and sample surveys". In *Proceedings of the 60th ISI World Statistics Congress*, 26–31 July 2015, Rio de Janeiro, Brazil. Available at: https://2015.isiproceedings.org/Files/IPS022-P4-S.pdf (accessed September 2021).

D'Ambrosio C. 2003. "Household Characteristics and the Distribution of Income in Italy: An Application of Social Distance Measures." *The Review of Income and Wealth* 47(1): 43–64. DOI: https://doi.org/10.1111/1475-4991.00003.

De Clercq M., J. Buysse, and M. D'Haese. 2020. The state of Rural Digitalisation in Europe. Desira. Available at: https://desira2020.eu/wp-content/uploads/2020/11/Briefing_Digitalisation-in-Europe.pdf (accessed September 2021).

De Leeuw, E.D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys". *Journal of Official Statistics* 21(2): 233–255.

De Leeuw E.D. 2018. "Mixed Mode: Past, Present, and Future." *Survey Research Methods* 12(2): 75–89. DOI: doi:10.18148/srm/2018.v12i2.7402.

De Leeuw, E.D., and N. Berzelak. 2016. "Survey Mode or Survey Modes?" In *The Sage Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T.W. Smith, and Y. Fu, 142–156. London: Sage Publications.

De Leeuw, E.D., J. Hox, and D. Dillman. 2008. *International Handbook of Survey Methodology.* New York: Routledge.

De Leeuw E.D., Z.T. Suzer-Gurtekin, and J.J. Hox. 2019. "The Design and Implementation of Mixed-mode Surveys." In *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural contexts (3MC)*, edited by P. Johnson, B.E. Pennell, I.A.L. Stoop, and B. Dorer: 387–408. New York: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Fourth edition. New York: Wiley.

Durrant, G.B., and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-contact in Household Survey Evidence from Six UK Government Surveys." *Journal of the Royal Statistical Society: Series A* 172(2): 361–381. DOI: https://doi.org/10.1111/j.1467-985X.2008.00565.x.

European Union 2017. *Regulation of the European Parliament as Regards the Territorial Typologies 2017/2391.* Official Journal of the European Union. Available at: https://eur-

lex.europa.eu/legal-content/EN/TXT/?uri = celex%3A32017R2391 (accessed September 2021).

Eurostat 2020. Digital Economy and Society Statistics – Households and Individuals. Luxemburg. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php/-Digital_economy_and_society_statistics_-_households_and_individuals (accessed September 2021).

Eurostat 2021. *Applying the Degree of Urbanisation. A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons*. 2021 edition. Eurostat: Luxemburg.

Goldstein, H. 2010. *Multilevel Statistical Models*. Fourth Edition. New York: Wiley.

Hargittai, E. 2002."Second-Level Digital Divide: Differences in People's online skills." *First Monday* 7(4). DOI: https://doi.org/10.5210/fm.v7i4.942.

Hox J.J., E.D. de Leeuw, and E.A.O. Zijlmans. 2015. Measurement Equivalence in Mixed Mode Surveys. *Frontiers in Psychology* 6: 87. DOI: https://doi.org/10.3389/fpsyg.2015.00087.

Istat 2018. "Preliminary Experimental Results on the Italian Population and Housing Census Estimation Methods." In *Proceedings of the Twentieth Conference of European Statisticians Group of Experts on Population and Housing Censuses*, 26–28 September, 2018. Geneva: United Nations Economic Commission for Europe. Available at: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2018/Meeting-Geneva-Sept/ECE_CES_GE.41_2018_20-1811953E.pdf (accessed September 2021).

Istat 2019. *Cittadini e ICT*. Statistiche report. Rome: Istat. Available at: https://www.istat.it/it/archivio/236920 (accessed September 2021).

Jaccard, p. 1908. "Nouvelles recherches sur la distribution florale." *Bulletin de la Société Vaudoise des Sciences Naturelles* 44(163): 223–270.

Kappelhof, J.W.S. 2015. "Face-to-Face or Sequential Mixed-Mode Survey Among Non-Western Minorities in the Netherlands: The effect of Differential Survey Designs on the Possibility of Nonresponse Bias." *Journal of Official Statistics*, 31(1): 1–3. DOI: https://doi.org/10.1515/jos-2015-0001.

Keusch F., B. Struminskaya, C. Antoun, M.P. Couper, F. Kreuter. 2019. "Willingness to Participate in Passive Mobile Data Collection." *Public Opinion Quarterly* 83: 210–235. DOI: https://doi.org/10.1093/poq/nfz007.

Luiten A., J. Hox and E.D. de Leeuw. 2020. Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics* 36(3): 469–487. DOI: https://doi.org/10.2478/-jos-2020-0025.

Maddala, G.S. and K. Lahiri. 2010. Introduction to Econometrics. Fourth Edition. New York: Wiley.

Maslovskaya, O., G.B. Durrant, P.W.F. Smith and A. Villar. 2019. "What are the Characteristics of Respondents using Different Devices in Mixed-device Online Surveys? Evidence from Six UK Surveys." *International Statistical Review* 87(2): 326–346. DOI: https://doi.org/10.1111/insr.12311.

McCormick, W.P., N.I. Lyons, and K. Hutcheson. 1992. "Distributional Properties of Jaccard's Index of Similarity." *Communications in Statistics – Theory and Methods* 21(1): 51–68. DOI: https://doi.org/10.1080/03610929208830764.

Messer, B.L., and D.A. Dillman. 2011. "Surveying the general public over the internet using address-based sampling and mail contact procedures." *Public Opinion Quarterly*, 75(3), 429–457. DOI: https://doi.org/10.1093/poq/nfr021.

Mohorko A., E. de Leeuw, and J. Hox. 2013. "Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time." *Journal of Official Statistics* 29(4): 609–622. DOI: https://doi.org/10.2478/jos-2013-0042.

Pratesi, M., K. Lozar Manfreda, S. Biffignandi, and V. Vehovar. 2004. "List-based Web Surveys: Quality, Timeliness and Nonresponse in the Steps of the Participation Flow." *Journal of Official Statistics* 20(3): 451–465. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/list-based-web-surveys-quality-timeliness-and-nonresponse-in-the-steps-of-the-participation-flow.pdf.

Rados, J. 2021. "Broadband Access in Rural and Mountainous Areas." The European Parliament Intergroup on Rural, Mountainous and Remote Areas and Smart Villages. Available at: https://www.smart-rural-intergroup.eu/broadband-access-in-rural-and-mountainous-areas-in-the-eu/ (accessed September 2021).

Reynaud C., S. Miccoli, F. Benassi, A. Naccarato, and L. Salvati. 2020. "Unravelling a Demographic 'Mosaic': Spatial Patterns and Contextual Factors of Depopulation in Italian Municipalities, 1981–2011." *Ecological Indicators* 115: 106356. DOI: https://doi.org/10.1016/j.ecolind.2020.106356.

Righi P., P.D. Falorsi, S. Daddi, E. Fiorello, P. Massoli, and M.D. Terribili. 2021. "Optimal Sampling for the Population Coverage Survey of the New Italian Register Based Census." *Journal of Official Statistics* 37(3): 655–671. DOI: https://doi.org/10.2478/jos-2021-0029.

Rivero, M.S., M.C.R. Rangel, and J.M.S. Martín. 2019. "Geotourist Profile Identification using Binary Logit Modeling: Application to the Villuercas-Ibores-Jara Geopark (Spain)." *Geoheritage* 11: 1399–1412. DOI: https://doi.org/10.1007/s12371-019-00384-z.

Rizzato, F. 2020. "Broadband Snapshots: Italy". Opensignal. Available at https://www.opensignal.com/2020/09/04/broadband-snapshots-italy (accessed September 2021).

Santarelli, E., and F. Cottone. 2009. "Leaving Home, Family Support and Intergenerational Ties in Italy: Some Regional Differences." *Demographic Research* 21(1): 1–22. DOI: https://doi.org/10.4054/DemRes.2009.21.1.

Scherpenzeel, A.C. 2011. "Data Collection in a Probability-based Internet Panel: How the LISS Panel Was Built and How It Can Be Used." *Bulletin of Sociological Methodology* 109(1): 56–61. DOI: https://doi.org/10.1177/0759106310387713.

Scherpenzeel, A.C. and J.G. Bethlehem. 2011. "How Representative are Online Panels? Problems of Coverage and Selection and Possible Solutions." In *Social and Behavioural Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by J.W.M. Das, P. Ester, and L. Kaczmirek: 105-132. New York: Routledge.

Tourangeau R. 2017. Presidential Address: Paradoxes of Nonresponse. *Public Opinion Quarterly* 81: 803–817. DOI: https://doi.org/10.1093/poq/nfx031.

UNECE (United Nations Economic Commission for Europe). 2018. *Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses.* ECE/CES/STAT/2018/4. New York: United Nations. Available at: https://unece.org/-guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0 (accessed September 2021).

Wooldridge J.M. 2012. *Introductory Econometrics: A Modern Approach*. Fifth Edition. Boston: Cengage Learning.

Received March 2021
Revised October 2021
Accepted July 2022

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 27–44, http://dx.doi.org/10.2478/JOS-2023-0002

A Multivariate Regression Estimator of Levels and Change for Surveys Over Time

Anne Konrad[1] *and Yves Berger*[2]

Rotations are often used for panel surveys, where the observations remain in the sample for a predefined number of periods and then rotate out. The information of previous waves can be exploited to improve current estimates. We propose a multivariate regression estimator which captures all information available from both waves. By adding additional auxiliary variables describing the information of the rotational design, the proposed estimator captures the sample correlation between waves. It can be used for the estimation of levels and changes.

Key words: Generalized regression estimation; composite estimator; rotating samples.

1. Introduction

Repeated socioeconomic surveys are often the basis for evaluating changes and levels over time (e.g., Smith et al. 2003). Estimates are usually based on repeated or rotational surveys, which involve rotations, that is, units remain in a survey for a predefined number of waves and then are replaced by new sampled units (e.g., Gambino and Silva 2009; Kalton 2009; Eurostat 2012). There are different rotation schemes. In an in-for-x rotational design the units remain in the sample for x consecutive waves and then are replaced by new sampled units. In an x-(y)-z rotational design, the units remain in the sample for x consecutive waves, leave the sample for y waves and then return for z consecutive waves. Then they are dropped from the sample completely and replaced by new sampled units (e.g., Bonnéry et al. 2020, 170). We shall consider two waves, but the proposed approach can be extended to more than two waves (see Subsection 3.1).

Rotational designs give partially overlapping samples between waves. Thus, between two consecutive waves, we have units sampled at both waves (the overlapping units), units sampled only at the first wave (units that rotate out) and units sampled only at the second wave (units that rotate in). The sample information from the previous wave can be used to improve the current wave estimates. We expect to have more efficient estimates when the variables are correlated over time (Steel and McLaren 2008).

We propose a "multivariate generalised regression" (GREG) estimator that exploits the sample overlap between two waves, as well as the non-overlapping samples containing the units observed in only one of the waves. The proposed estimator includes "extended design variables" as additional auxiliary variables, which capture the sample correlation between the variables and the sample rotation. Thereby, it borrows strength from all available sample information on the variables of interest and the auxiliary variables from both waves. This may provide efficient change and levels estimates. Furthermore, the extended design

[1] Leibniz Institute for Educational Trajectories, Survey Statistical Mathods, 96047 Bamberg, Germany, Email: anne.konrad@lifbi.de
[2] University of Southampton, Social Statistics, SO17 1BJ, Southampton, UK, Email: Y.G.Berger@soton.ac.uk

variables capture the sample design information, such as stratification and unequal probabilities. The proposed estimator can be applied for rotational samples of any rotation scheme or for the simultaneous estimation of two or more consecutive waves; for example, impact evaluation surveys with a baseline and a post-intervention data collection.

The idea of including the sample information on variables of interest from previous waves is not new. Hansen et al. (1953) and Gurney and Daly (1965) introduced a class of composite estimators that exploit the sample overlap between two consecutive waves. The "modified regression estimator" of Singh et al. (1997) includes additional auxiliary variables based on the variables of interest from previous waves. However, for the new units that rotate in, the values of these additional variables are unknown and usually imputed. The control totals of the additional variables are also unknown and have to be estimated, which leads to a variance inflation of the current wave estimate. In contrast, the proposed estimator neither relies on imputation nor the estimation of unknown control totals.

The article is organised as follows. Section 2 introduces the basic framework on rotational surveys and GREG estimators. In Section 3, we derive the proposed multivariate GREG estimator and its properties. Asymptotic optimality and variance estimation is investigated in Section 4. Alternative estimators considered in the literature such as the modified regression estimator are discussed in Section 5. In Section 6, a Monte Carlo simulation study compares the proposed multivariate GREG estimator with the modified regression estimator. Section 7 summarises our results.

2. Rotation Design and Generalised Regression Estimator

Let $U = \{1, \ldots, i, \ldots, N\}$ be a population of N units. Without loss of generality, we consider two waves ($t = 1$ and $t = 2$). The proposed estimator introduced in Section 3, will be extended to more than two waves in Subsection 3.1. We assume that the population units are the same in both waves. In practice, a change in a population can be handled by adjusting the weights and the sampling frame in the cases of birth, death and emigration.

Let s_1 be the first wave sample of size n_1 selected without-replacement from U. The first-order inclusion probability of unit i for wave 1 and 2 are denoted respectively by $\pi_{i1} = Pr(i \in s_1)$ and $\pi_{i2} = Pr(i \in s_2)$, where $Pr(\cdot)$ denotes the probability with respect to the design. We assume that both sample sizes n_1 and n_2 are fixed. The common sample is $s_{12} = s_1 \cap s_2$, with a sample size $n_{12} = \#s_{12}$, where $0 \leq n_{12} \leq n_1$. We assume that n_{12} is fixed, because this is a common feature of rotational designs. It is common practice to assume that the units that rotate out cannot rotate in; that is, $Pr\{i \in (s_2 \setminus s_1) | i \in s_1\} = 0$.

Stratification is often used in practice. We suppose that the population U is stratified into H strata U_h, such that $\cup_{h=1}^{H} U_h = U$. We assume that stratification is the same at both waves. Let $s_{t,h}$ be the t-th wave sample of size $n_{t,h}$ selected without-replacement from the population U_h, where $t = 1$ or 2. At wave t, the overall sample is $s_t = \cup_{h=1}^{H} s_{t,h}$ with a total sample size $n_t = \sum_{h=1}^{H} n_{t,h}$. We assume that we have a rotation within strata, that is, the common sample within U_h is denoted by $s_{12,h} = s_{1,h} \cap s_{2,h}$, with a sample size $n_{12,h} = \#s_{12,h}$, fixed by design. The ratio $\theta_h = n_{12,h}/n_{1,h}$ is the fraction of the overlap within U_h. The quantities θ_h are allowed to vary between strata.

The objective is to estimate unknown population totals of a variable of interest y, for different waves. The total of wave t is

$$\tau_{y_t} := \sum_{i \in U} y_{it},$$

where y_{it} is the value of y for a unit $i \in U$ at wave t. The Horvitz and Thompson (1952) estimator

$$\hat{\tau}_{y_t} := \sum_{i \in s_t} \frac{y_{it}}{\pi_{it}}$$

is a design-unbiased estimator of τ_{y_t}. For estimation of a domain of interest, we impose $y_{it} = 0$ for the units i outside the domain.

The efficiency can be improved by incorporating auxiliary information in the estimation process. A widely used model-assisted estimator based on auxiliary information, is the generalised regression (GREG) estimator (Hansen et al 1953; Cassel et al. 1977; Särndal, 1980; Isaki and Fuller 1982; Wright 1983). Let $\mathbf{x}_{it}$ be the Q_t-vector of auxiliary variables for a unit i at wave t. Suppose that the vector of population totals $\boldsymbol{\tau}_{xt} = \sum_{i \in U} \mathbf{x}_{it}$ at wave t, is known from census, registers, or other reliable sources. The customary GREG estimator is defined by

$$\hat{\tau}^{g}_{y_t} := \hat{\tau}_{y_t} + \hat{\mathbf{B}}_t^{\top} (\tau_{x_t} - \hat{\tau}_{x_t}) \tag{1}$$

where

$$\hat{\boldsymbol{\tau}}_{x_t} := \sum_{i \in s_t} \frac{\mathbf{x}_{it}}{\pi_{it}}, \tag{2}$$

$$\hat{\mathbf{B}}_t := \left(\sum_{i \in s_t} \frac{\mathbf{x}_{it} \mathbf{x}_{it}^{\top}}{\pi_{it}} \right)^{-1} \sum_{i \in s_t} \frac{\mathbf{x}_{it} y_{it}}{\pi_{it}}. \tag{3}$$

The estimator (1) is motivated by the linear regression model

$$y_{it} = \mathbf{x}_{it}^{\top} \boldsymbol{\beta}_t + \varepsilon_{it}, \tag{4}$$

specifying the relationship between y_{it} and $\mathbf{x}_{it}$, where $E(\varepsilon_{it}) = 0$, $V(\varepsilon_{it}) = \sigma^2$ and $\mathrm{E}(\varepsilon_{it}\varepsilon_{jt}) = 0$ for all $i \neq j$. If $V(\varepsilon_{it}) = v_{it}\sigma^2$, a weighted least-squares estimator can be used instead of Equation (3) to reflect heteroscedasticity. In order to simplify the notation, we shall assume $v_{it} = 1$. Nevertheless, when $v_{it} \neq 1$, they can be easily added to the regression coefficient (13) of the proposed estimator. The use of v_{it} is more relevant for business surveys. Homoscedasticity ($v_{it} = 1$) is often assumed in household surveys (Steel and Clark 2007, 52).

The asymptotic design-unbiased estimator (1) does not depend on whether the model (4) holds or not. Its efficiency is driven by the predictive power of the model (cf. Särndal et al. 1992, 227, 239). Hereafter, we shall use a design-based approach, that is, the model (4) shall not be used for inference.

3. Proposed Multivariate Regression Estimator

Let us consider the "*combined sample*" defined by the set $s_b = s_1 \cup s_2$ comprising all units from both waves. The corresponding sample size of s_b is denoted by $n_b = \#s_b = n_1 + n_2 - n_{12}$. Let the "*extended weighted variable of interest*" be defined by

$$\breve{y}_{it} := \frac{y_{it}}{\pi_{it}} \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2, \tag{5}$$

where $\delta\{i \in s_t\} = 1$ if $i \in s_t$, and $\delta\{i \in s_t\} = 0$ otherwise. Note that $\breve{y}_{i2} = 0$ for all units $i \in s_b \setminus s_2$ that rotates out. We also have $\breve{y}_{i1} = 0$ for all units $i \in s_b \setminus s_1$ that rotates in. Figure 1 is a visual representation of two waves, with units on the horizontal axis and the two waves on the vertical axis.

The "*extended weighted auxiliary variables*" are defined by

$$\breve{\mathbf{x}}_{it} := \frac{\mathbf{x}_{it}}{\pi_{it}} \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2. \tag{6}$$

The set of auxiliary variables used at $t = 1$ can be different from the one used at $t = 2$. The set of auxiliary variables can also be the same. This is usually the case for panel surveys.

Note that Equation (2) can also be re-written as $\hat{\tau}_{x_t} = \sum_{i \in s_b} \breve{\mathbf{x}}_{it}$. We also consider "*extended design variables*" given by

$$\mathbf{z}_{it} := (z_{it,1}, \ldots, z_{it,h}, \ldots, z_{it,H})^\top \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2$$

with $z_{it,h} = 1$ if the unit i belongs to stratum h in wave t, and $z_{it,h} = 0$ otherwise. The vector $\mathbf{z}_{it}$ represents the sampling design information given by the stratification. The Hadamard product $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ will play a key role. It reveals the information induced by the rotation, because it identifies the units within the common sample. Indeed, the h-th component of $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ is equal to one if and only if the unit i belongs to the common sample of strata h. This component equals zero if and only if the unit i rotates in or out. Thus, $\mathbf{z}_{it}$ can be used to describe the sample information given by the rotation and the stratification.

It can be verified that

$$\sum_{i \in s_t} \mathbf{z}_{it} = \mathbf{n}_t \quad \text{and} \quad \sum_{i \in s_b} \mathbf{z}_{i1} \circ \mathbf{z}_{i2} = \mathbf{n}_{12}, \tag{7}$$

where

$$\mathbf{n}_t := (n_{t,1}, \ldots, n_{t,h}, \ldots, n_{t,H})^\top,$$
$$\mathbf{n}_{12} := (n_{12,1}, \ldots, n_{12,h}, \ldots, n_{12,H})^\top.$$

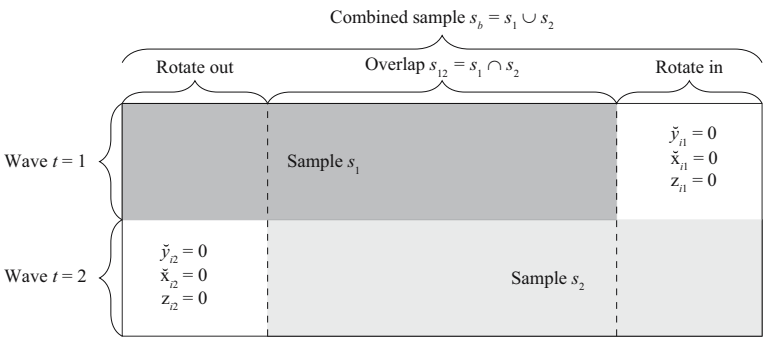


Fig. 1. Visual representation of two waves. The vertical axis represents the two waves: $t = 1$ and $t = 2$. The horizontal axis represents the units of the combined sample $s_b = s_1 \cup s_2$. The sample s_1 and s_2 are given in two different gray scales: for the sample s_1 and for the sample s_2.

Equations (7) hold, because we have stratified design and we have a rotation within strata.

Let $\breve{\mathbf{y}}_i = (\breve{y}_{i1}, \breve{y}_{i2})^\top$ be the "*combined extended variable of interest*" of wave 1 and wave 2. We also pool together the extended weighted auxiliary variables and the extended design variables into a single vector $\boldsymbol{\gamma}_i$ of dimension $(Q_1 + Q_2 + 3H)$, given by

$$\boldsymbol{\gamma}_i := \{\breve{\mathbf{x}}_{i1}^\top, \breve{\mathbf{x}}_{i2}^\top, \mathbf{z}_{i1}^\top, \mathbf{z}_{i2}^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^\top\}^\top. \tag{8}$$

This new auxiliary variable $\boldsymbol{\gamma}_i$ contains the original auxiliary variables $\mathbf{x}_{it}$, the stratification variables $\mathbf{z}_{it}$ and the variables $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ which specify the rotation within strata.

Berger et al. (2003) proposed using the stratification variables as auxiliaries within a GREG estimator, when we have a single-stage stratified sampling designs. This has the merit of achieving asymptotic optimality. The resulting estimator is easy to implement and does not rely on joint-inclusion probabilities. The proposed multivariate GREG estimator (9) is based on a similar idea, except that we use the additional variables $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ to capture the rotation.

The proposed multivariate GREG estimator for the unknown vector $\boldsymbol{\tau}_y = (\boldsymbol{\tau}_{y_1}, \boldsymbol{\tau}_{y_2})^\top$ of totals, is defined by

$$\hat{\boldsymbol{\tau}}_y^{\text{greg}} := \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_\gamma^\top (\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma), \tag{9}$$

where

$$\hat{\boldsymbol{\tau}}_y := (\hat{\boldsymbol{\tau}}_{y_1}, \hat{\boldsymbol{\tau}}_{y_2})^\top, \tag{10}$$

$$\boldsymbol{\tau}_\gamma := (\boldsymbol{\tau}_{x_1}^\top, \boldsymbol{\tau}_{x_2}^\top, \mathbf{n}^\top)^\top, \tag{11}$$

$$\hat{\boldsymbol{\tau}}_\gamma := (\hat{\boldsymbol{\tau}}_{x_1}^\top, \hat{\boldsymbol{\tau}}_{x_2}^\top, \mathbf{n}^\top)^\top, \tag{12}$$

$$\hat{\mathbf{B}}_\gamma := \left(\sum_{i \in s_b} c_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top \right)^{-1} \sum_{i \in s_b} c_i \boldsymbol{\gamma}_i \breve{\mathbf{y}}_i^\top, \tag{13}$$

$$\mathbf{n} := (\mathbf{n}_1^\top, \mathbf{n}_2^\top, \mathbf{n}_{12}^\top)^\top, \tag{14}$$

$$c_i := 1 - Pr(i \in s_b). \tag{15}$$

The matrix (13) is a regression coefficient matrix of dimension $(Q_1 + Q_2 + 3H) \times 2$. We introduce the c_i to achieve asymptotic optimality (see Section 4). Since $s_b = s_1 \cup s_2$, we have $Pr(i \in s_b) = \pi_{i1} + \pi_{i2} - Pr(i \in s_{12})$. Now, since $s_{12} = s_{12} \cap s_1$, $Pr(i \in s_{12}) = Pr(i \in s_1) Pr(i \in s_{12} | i \in s_1)$. Thus,

$$Pr(i \in s_b) = \pi_{i1} + \pi_{i2} - \pi_{i1} Pr(i \in s_{12} | i \in s_1). \tag{16}$$

The conditional probability $Pr(i \in s_{12} \mid i \in s_1)$ depends on the design and can be approximated by $\theta_h = n_{12,h}/n_{1,h}$ where $U_h \ni i$. Therefore, hereafter we shall use

$$c_i = 1 - \pi_{i1} - \pi_{i2} + \pi_{i1} \theta_h, \quad \text{where } h: \ U_h \ni i. \tag{17}$$

Exact computation of $Pr(i \in s_{12} \mid i \in s_1)$ is of little use. With large sampling fractions, the c_i are less than 1 and can be interpreted as finite population corrections within Equation (13). They should not affect the consistency of Equation (9), because they are weights used

only within Equation (13). Note that with negligible sampling fractions $c_i \approx 1$. The c_i will be also used for variance estimation (see Equation (27)).

Because of nonresponse, we could have units within the overlapping sample, which are not available at both occasions. Re-weighting should be used to compensate for the missing observations. In this case, within Equations (5) and (6), the basic weights π_{it}^{-1} should be replaced by weights that takes the missingness into account.

Theorem 1 gives an alternative expression for the proposed estimator which will be used to show its asymptotic optimality in Section 4.

Theorem 1. *An alternative expression for* $\hat{\boldsymbol{\tau}}_y^{greg}$ *is*

$$\hat{\boldsymbol{\tau}}_y^{greg} = \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_x^{\top}(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x), \tag{18}$$

where

$$\hat{\mathbf{B}}_x := (\breve{\mathbf{X}}^{\top}\mathbf{C}\mathbf{M}_z\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}^{\top}\mathbf{C}\mathbf{M}_z\,\breve{\mathbf{y}}, \tag{19}$$

$$\mathbf{M}_z := \mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\top}\mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}^{\top}\mathbf{C}, \tag{20}$$

$$\begin{aligned}
\breve{\mathbf{X}} &:= (\breve{\mathbf{x}}_1^{\top}, \ldots, \breve{\mathbf{x}}_{n_b}^{\top})^{\top},\\
\breve{\mathbf{y}} &:= (\breve{\mathbf{y}}_1^{\top}, \ldots, \breve{\mathbf{y}}_{n_b}^{\top})^{\top},\\
\mathbf{Z} &:= (\mathbf{z}_1^{\top}, \ldots, \mathbf{z}_{n_b}^{\top})^{\top},\\
\mathbf{C} &:= \mathrm{diag}\{c_1, \ldots, c_{n_b}\},
\end{aligned} \tag{21}$$

$$\begin{aligned}
\breve{\mathbf{y}}_i &:= (\breve{y}_{i1}, \breve{y}_{i2})^{\top},\\
\breve{\mathbf{x}}_i &:= (\breve{\mathbf{x}}_{i1}^{\top}, \breve{\mathbf{x}}_{i2}^{\top})^{\top},\\
\mathbf{z}_i &:= \{\mathbf{z}_{i1}^{\top}, \mathbf{z}_{i2}^{\top}, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^{\top}\}^{\top},
\end{aligned} \tag{22}$$

$$\begin{aligned}
\boldsymbol{\tau}_x &:= (\boldsymbol{\tau}_{x_1}^{\top}, \boldsymbol{\tau}_{x_2}^{\top})^{\top},\\
\hat{\boldsymbol{\tau}}_x &:= (\hat{\boldsymbol{\tau}}_{x_1}^{\top}, \hat{\boldsymbol{\tau}}_{x_2}^{\top})^{\top}
\end{aligned} \tag{23}$$

and $\boldsymbol{I}$ *is the* $n_b \times n_b$ *identity matrix.*

The proof can be found in the Appendix (Section 8) and is based on the fact that the Horvitz-Thompson estimators of the totals of the design variables are equal to their population totals, that is, $\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma = \{(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x)^{\top}, \mathbf{0}^{\top}\}^{\top}$.

The underlying multivariate model that leads to Equation (18) is

$$\mathbf{y}_i = \mathbf{x}_i^{\top}\boldsymbol{\beta}_x + \varepsilon_i,$$

where $\mathbf{y}_i := (\pi_{i1}\breve{y}_{i1},\ \pi_{i2}\breve{y}_{i2})^{\top}$ and $\mathbf{x}_i := (\pi_{i1}\breve{\mathbf{x}}_{i1}^{\top}, \pi_{i2}\breve{\mathbf{x}}_{i2}^{\top})^{\top}$. This model takes the correlation between waves into account, because variables of both waves are included within $\mathbf{y}_i$ and $\mathbf{x}_i$. Furthermore, $\hat{\boldsymbol{\tau}}_x$ contains the totals of both waves.

The proposed estimator borrows strength from both waves, by using both waves auxiliary variables. Furthermore, it takes the stratification into account, because of the extended design variables $\mathbf{z}_{i1}$ and $\mathbf{z}_{i2}$. In addition, the variable $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ exploits the rotation between s_1 and s_2 induced by the sample overlap. In contrast, the regression coefficient of the wave specific GREG estimator, given by Equation (3), does not involve design variables or information about the rotation. It does not take into account the correlation between the waves for the auxiliary variables and the variable of interest.

3.1. Extension to More than Two Waves

The proposed estimator can be easily extended for more than two waves. Suppose we have three consecutive waves. At wave 2, the multivariate GREG estimator produces two estimates: $\hat{\tau}_{y_1}^{\text{greg}}$ for wave 1 and $\hat{\tau}_{y_2}^{\text{greg}}$ for wave 2, where $\hat{\tau}_{y_2}^{\text{greg}}$ borrows strength from the information of wave 1. At wave 3, we obtain a new estimate $\hat{\tau}_{y_2}^{\text{greg(2)}}$ for wave 2 and an estimate $\hat{\tau}_{y_3}^{\text{greg}}$ for wave 3. Therefore, we have two estimates for wave 2: $\hat{\tau}_{y_2}^{\text{greg}}$ and $\hat{\tau}_{y_2}^{\text{greg(2)}}$. In official statistics, due to the need for up-to-date information, the estimate $\hat{\tau}_{y_2}^{\text{greg}}$ is immediately published at wave 2. The second estimate $\hat{\tau}_{y_2}^{\text{greg(2)}}$ is not published and should not be viewed as a revised estimate for the second wave total. It is only used to produce $\hat{\tau}_{y_3}^{\text{greg}}$. Furthermore, there is no reason for $\hat{\tau}_{y_2}^{\text{greg(2)}}$ to be more precise than $\hat{\tau}_{y_2}^{\text{greg}}$, since both are based on the same controls and correlations. The estimates $\hat{\tau}_{y_2}^{\text{greg}}$ and $\hat{\tau}_{y_2}^{\text{greg(2)}}$ are not used as controls to produce $\hat{\tau}_{y_3}^{\text{greg}}$, as with the modified regression estimator (see Section 5).

The proposed estimator is flexible, because it can be also use to borrow strength over more than two waves. In this case, the dimension of the vectors $\hat{\boldsymbol{\tau}}_y^{\text{greg}}$ and $\breve{\mathbf{y}}_i$ is the number of waves. The vectors $\breve{\mathbf{y}}_i$ and $\breve{\mathbf{x}}_i$ contain the variables of the waves considered. In this case, the vector (22) may need to include additional components depending on the design. For simplicity, we recommend to use $c_i = 1$ in this case.

For example, suppose we have three waves, the sample sizes of the overlapping sets between the three samples from the same stratum can be fixed by design; i.e $n_{12,h}, n_{23,h}, n_{13,h}$ and $n_{123,h}$ may be fixed, where $n_{tl,h}$ denotes the sample size of $s_{t,h} \cap s_{l,h}$ within stratum U_h. Here, $n_{123,h}$ is the sample size of $s_{1,h} \cap s_{2,h} \cap s_{3,h}$. This situation occurs when we use the customary rotation group method. In this case, we need to include within $\mathbf{z}_i$: (i) $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ for the fixed sample size of $s_{1,h} \cap s_{2,h}$, (ii) $\mathbf{z}_{i2} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{2,h} \cap s_{3,h}$, (iii) $\mathbf{z}_{i1} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{1,h} \cap s_{3,h}$, (iv) $\mathbf{z}_{i1} \circ \mathbf{z}_{i2} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{1,h} \cap s_{2,h} \cap s_{3,h}$; that is, the vectors (8) and (14) should be replaced respectively by

$$\gamma_i = \{\breve{\mathbf{x}}_{i1}^\top, \breve{\mathbf{x}}_{i2}^\top, \breve{\mathbf{z}}_{i1}^\top, \breve{\mathbf{z}}_{i2}^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^\top, (\mathbf{z}_{i2} \circ \mathbf{z}_{i3})^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i3})^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2} \circ \mathbf{z}_{i3})^\top\}^\top,$$
$$\mathbf{n} = (\mathbf{n}_1^\top, \mathbf{n}_2^\top, \mathbf{n}_{12}^\top, \mathbf{n}_{23}^\top, \mathbf{n}_{13}^\top, \mathbf{n}_{123}^\top)^\top,$$

with $\mathbf{n}_{23} := (n_{23,1}, \ldots, n_{23,H})^\top$, $\mathbf{n}_{13} := (n_{13,1}, \ldots, n_{13,H})^\top$ and $\mathbf{n}_{123} := (n_{123,1}, \ldots, n_{123,H})^\top$.

4. Asymptotic Optimality and Variance Estimation

In this section, we show the asymptotic optimality when we have two waves; The asymptotic optimal GREG estimator (Montanari 1987) of the vector of totals $\boldsymbol{\tau}_y = (\tau_{y_1}, \tau_{y_2})^\top$ is

$$\hat{\boldsymbol{\tau}}_y^{\text{opt}} := \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_{\text{opt}}^\top (\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x), \tag{24}$$

where

$$\hat{\mathbf{B}}_{\text{opt}}^\top := \hat{\mathbf{V}}(\hat{\boldsymbol{\tau}}_x)^{-1} \widehat{\mathbf{Cov}}(\hat{\boldsymbol{\tau}}_x, \hat{\boldsymbol{\tau}}_y). \tag{25}$$

See Guandalini and Tillé (2017, 3) for more details. By using the Horvitz and Thompson (1952) variance and covariance estimators, the expression (25) reduces to

$$\hat{\mathbf{B}}_{\text{opt}} = (\check{\mathbf{X}}^{\top}\boldsymbol{\Delta}\check{\mathbf{X}})^{-1}\check{\mathbf{X}}^{\top}\boldsymbol{\Delta}\check{\mathbf{y}}, \tag{26}$$

where

$$\boldsymbol{\Delta} := \{(\pi_{ij} - \pi_i\pi_j)\pi_{ij}^{-1}; i,j \in s_b\}.$$

Here, $\pi_{ij} = Pr(i, j \in s_b)$ denotes the joint-inclusion probability of units i and j for the sample s_b. These are different from the joint probabilities of s_1 and s_2, because π_{ij} takes the rotation into account. Since the probabilities π_{ij} are unknown, we propose to use the asymptotic approximation of Hájek (1964), based on the assumption that the rotation design is asymptotically rejective according to the design constraints (7). This approximation is given by $\boldsymbol{\Delta} \approx \mathbf{C}\mathbf{M}_z$, where $\mathbf{C}$ and $\mathbf{M}_z$ are defined respectively by Equations (20) and (21) (Hájek 1981 chap. 14; Berger et al. 2003; and Deville and Tillé 2005). Now, by replacing this approximation of $\boldsymbol{\Delta}$ within Equation (26), we obtain Equation (19). Thus, the proposed estimator $\hat{\boldsymbol{\tau}}_y^{\text{greg}}$ is indeed optimal asymptotically.

A variance of the estimator (9) can be derived, based on principle that the variance under Poisson sampling of the regression estimator (9) based on the auxiliary and design variables, is asymptotically the same as the variance of the regression estimator (18) under a rejective design (Hájek 1964; Berger 2004) with the design constraints (7). Thus, the variance estimator of (9), assuming that s_b is a Poisson sample with inclusion probabilities (16), is given by the variance-covariance matrix

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\tau}}_{\mathbf{y}}^{\text{greg}}) := (\mathbf{M}_{\boldsymbol{\Gamma}}\check{\mathbf{y}})^{\top}\mathbf{C}\,\mathbf{M}_{\boldsymbol{\Gamma}}\check{\mathbf{y}}, \tag{27}$$

where

$$\mathbf{M}_{\Gamma} := \mathbf{I} - \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^{\top}\mathbf{C}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\top}\mathbf{C},$$
$$\boldsymbol{\Gamma} := (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)^{\top}.$$

Note that (27) is a residual variance as in Särndal et al. (1992, 235), because $\mathbf{M}_{\boldsymbol{\Gamma}}\check{\mathbf{y}}$ are residuals. Note that the variance estimator takes the stratification into account, because the information about the strata is included within $\mathbf{M}_{\boldsymbol{\Gamma}}$. However, if within Equations (5) and (6), the basic weights π_{it}^{-1} are substituted by weights which take the missingness into account, the variance estimator (27) may be biased, because nonresponse is not accounted for.

5. Alternative Approaches

Composite estimators also use the information from previous waves. Hansen et al. (1953) introduced the K-composite estimator for levels and change between two waves. The AK-composite estimator (Gurney and Daly 1965) takes the difference between the common sample s_{12} and the unmatched sample s_2 into account. The optimal choice of the weighting factors A and K, within the AK-composite estimator, depends on the variables of interest (Kumar et al. 1983). This dependency may result in an inconsistency, in the sense that the sub-group total estimates may not add up to the overall total (Gambino et al. 2001, 66).

Singh (1996) and Singh et al. (1997) introduced the modified regression estimator, abbreviated MR hereafter. The idea is to extend the auxiliary variables in the current wave by an additional artificial auxiliary variable, which contains the information on the

variable of interest from the previous wave. The definition of this variable depends on whether the primary interest lies on levels or change. If the main focus lies on levels, the artificial variable refers to the variable of interest y_{i1} from the previous wave. However, due to the rotation, y_{i1} is only known for $i \in s_{12}$. Singh (1996) suggested to use mean imputation for the unknown values for the units $i \in s_2 \setminus s_{12}$. Thus, in this case, the artificial variable is

$$\tilde{x}_{i2}^{\text{MR1}} := \begin{cases} y_{i1} & \text{for } \; i \in s_{12} \\ \hat{\mu}_{y1} & \text{for } \; i \in s_2 \backslash s_{12}, \end{cases} \tag{28}$$

where $\hat{\mu}_{y_1}$ is an estimator for the mean of y_1. The control total of the variable (28) is unknown and can be estimated by $N\hat{\mu}_{y_1}$ (Fuller and Rao 2001, 47). Hence, the modified regression estimator for $\tau_{y_2} = \sum_{i \in U_2} y_{i2}$ is given by

$$\hat{\tau}_{y_2}^{\text{MR1}} := \hat{\tau}_{y2} + \hat{\mathbf{B}}_{x\tilde{x}}^{\top}(\tilde{\tau}_{x\tilde{x}} - \hat{\tau}_{x\tilde{x}}), \tag{29}$$

with

$$\hat{\mathbf{B}}_{x\tilde{x}} := (\mathbf{B}_{x_2}^{\top}, \hat{B}_{\tilde{x}_2})^{\top},$$

$$\tilde{\boldsymbol{\tau}}_{x\tilde{x}} := (\boldsymbol{\tau}_{x_2}^{\top}, N\hat{\mu}_{y_1})^{\top},$$
$$\hat{\boldsymbol{\tau}}_{x\tilde{x}} := (\hat{\boldsymbol{\tau}}_{x_2}^{\top}, \hat{\tau}_{\tilde{x}_2})^{\top}.$$

If the primary interest is to estimate a change, the artificial variable refers to the variable of interest y_{i2} from the current wave. The variable recommended by Singh (1996) and Singh et al. (1997) is

$$\tilde{x}_{i2}^{\text{MR2}} := \begin{cases} y_{i2} + \dfrac{n_2}{n_{12}}(y_{i1} - y_{i2}) & \text{for } i \in s_{12}, \\ y_{i2} & \text{for } i \in s_2 \backslash s_{12}. \end{cases} \tag{30}$$

The MR2 estimator may suffer from a drift in levels estimates over a long period (Gambino et al. 2001, 65; Fuller and Rao 2001, 50). In order to overcome this problem, Fuller and Rao (2001) introduced the regression composite estimator (RC) given by

$$\tilde{x}_{i2}^{\text{RC}} := (1 - \alpha)\tilde{x}_{i2}^{\text{MR1}} + \alpha\tilde{x}_{i2}^{\text{MR2}}, \tag{31}$$

where $\alpha \in [0, 1]$ is a tune-in parameter which reflects the importance given to levels or change estimates. The advantage of the regression composite estimator compared with MR1 and MR2 is the fact that it is a compromise between levels and change estimation. An alternative estimator could be based on Definitions (28) and (30). However, the increased number of auxiliaries and control totals may lead to a distortion in the final weights (Gambino et al. 2001, 65).

Singh et al. (2001) suggested a jackknife variance estimator that takes the estimation of the control totals into account. Indeed, ignoring the additional source of randomness would lead to an underestimation of the true variance. Berger et al. (2009) proposed a linearised variance estimator that takes the estimation of the controls into account.

The optimal BLUE estimator is based on a time series of the variable of interest (Yansaneh and Fuller 1998; Bell 2001; Australian Bureau of Statistics 2007). This estimator

requires that the variances and covariances of the rotation group estimates are known (Bell 2001, 56). If they were substituted by their estimates, it is no longer guaranteed that the BLUE estimator is optimal. Bonnéry et al. (2020) showed that the BLUE with an estimated variance-covariance matrix is less efficient than the composite estimators. Some disadvantages are discussed in Fuller (1990) and Steel and McLaren (2009). Since the BLUE estimator is based on a time series, it is less comparable with the proposed estimator and the modified estimators, which are both based on regression estimation.

6. Simulation Study

We consider three waves ($t = 0, 1, 2$), because the estimators (9) and (29) at wave $t = 1$, require the sample information from wave $t = 0$. The results are reported for levels at waves $t = 1$ and $t = 2$, and changes between waves $t = 1$ and $t = 2$.

Consider N population values of y_{it} and x_{it} ($t = 1, 2, 3$) generated from a multivariate normal distribution; that is,

$$(y_{i0}, y_{i1}, y_{i2}, x_{i0}, x_{i1}, x_{i2})^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Here, $\boldsymbol{\Sigma}$ denotes a covariance matrix with an heterogeneous exchangeable structure, that is,

$$\boldsymbol{\Sigma} := \text{diag}(\sigma)\{\rho \mathbf{J}_6 + (1-\rho)\mathbf{I}_6\}\text{diag}(\sigma).$$

where diag(σ) is the diagonal matrix with $\sigma = (\sigma_{y0}, \sigma_{y1}, \sigma_{y2}, \sigma_{x0}, \sigma_{x1}, \sigma_{x2})^\top$ as its diagonal. The matrices $\mathbf{J}_6$ is 6×6 matrix of ones and $\mathbf{I}_6$ is the 6×6 identity matrix. Thus, the correlations $cor(y_{it}, y_{it'}) = cor(x_{it}, x_{it'}) = cor(y_{it}, x_{it'}) = \rho$, with $t \neq t'$. Let $\sigma_{y_0} = 10$, $\sigma_{y_1} = 15$, $\sigma_{y_2} = 20$, $\sigma_{x_0} = 30$, $\sigma_{x_1} = 40$ and $\sigma_{x_2} = 50$. The correlations considered are $\rho = 0.1, 0.5$ and 0.9. Two values for the vector $\boldsymbol{\mu} = (\mu_{y0}, \mu_{y1}, \mu_{y2}, \boldsymbol{\mu}_{\mathbf{x}0}, \mu_{x1}, \mu_{x2})^\top$ are used:

$$\boldsymbol{\mu}_I := (59, 60, 61, 99, 100, 101)^\top,$$
$$\boldsymbol{\mu}_{II} := (40, 60, 80, 100, 150, 200)^\top,$$

that is, we have a small change with $\boldsymbol{\mu}_I$ and a large change with $\boldsymbol{\mu}_{II}$.

For each wave t, we have stratified samples of size $n_t = 1{,}000$. We consider four strata formed by the quantile classes of the population distribution of $y_{i1} + y_{i2}$. The same fraction of common samples between waves is used within strata, that is, $\theta_h = \theta = n_{12}/n_1 = n_{01}/n_0$, where $\theta = 0.25$, $\theta = 0.5$ or $\theta = 0.75$. Rotation groups sampling is implemented. Within each strata U_h, q units are randomly allocated into P rotation groups of same size $p = \lfloor q/P \rfloor$. The sample $s_{0,h}$ contains the units of the first $\lfloor n_h\, p^{-1} \rfloor$ groups. At wave $t = 1$, we obtain the sample $s_{1,h}$ by rotating out the first group and replacing it by the $(\lfloor n_h\, p^{-1} \rfloor + 1)$-th group. At wave $t = 2$, the second group rotates out and $(\lfloor n_h\, p^{-1} \rfloor + 2)$-th group rotates in. For $\theta = 0.25$, we use $q = 625$ and $P = 10$. With $\theta = 0.5$, we use $q = 400$ and $P = 4$ and with $\theta = 0.75$, we set $q = 300$ and $P = 6$. We consider 1,000 iterations.

In the first simulation setup, we consider equal allocation for all strata with $n_{t,h} = 250$ and $N = 100{,}000$. Thus, the inclusion probabilities are the same across the strata and the sampling fractions are small. In the second simulation setup, unequal probabilities with large sampling fractions are used. Consider $n_{1,h} = 50$, $n_{2,h} = 200$, $n_{3,h} = 350$, $n_{4,h} = 400$ and $N = 4{,}000$. The resulting within strata inclusion probabilities are 0.05, 0.2, 0.35 and

0.4. In the second simulation setup, the population size is $N = 4{,}000$, to allow for large sampling fractions.

The estimators considered are the proposed multivariate regression estimator (9) (PROP), the customary regression estimator (1) (GREG) and the modified estimator (29) with (28) as auxiliaries (MR1) and with (30) as auxiliaries (MR2). For MR1 and MR2, we use $\hat{\bar{\tau}}_{y_0}^{\text{greg}}$ as the estimated control total of the previous wave $t = 1$.

In order to explore the efficiency of point estimates, we compare the empirical '*relative root mean squared errors*' (RRMSE). Let $\hat{\tau}_r$ be an estimate for the r-th iteration with $r = 1, \ldots, 1{,}000$. The RRMSE is defined as

$$\text{RRMSE}(\hat{\tau}) := \frac{1}{|\tau|}\left\{\frac{1}{1{,}000}\sum_{r=1}^{1{,}000}(\hat{\tau}_r - \tau)^2\right\}^{1/2},$$

where τ denotes the population total.

The RRMSE $\times$ 100% for different values of ρ, g and θ, are reported in Tables 1 and 2. For Table 1, we have equal strata sizes with the same inclusion probabilities across strata and small sampling fractions. For Table 2, the inclusion probabilities differs between strata and some sampling fractions are large. The proposed PROP estimator outperforms GREG and MR1 under all scenarios. For all estimators under consideration, the RRMSE decreases with the correlation ρ between the variables.

We observe slightly smaller RRMSE for MR1 with $\theta = 0.75$, when $\rho = 0.1$, because the higher the overlap, the less values have to be imputed. The amount of overlap θ has little

Table 1. Equal strata sizes. equal probabilities and small sampling fractions. RRMSE X 100% of levels estimates under different scenarios for 1,000 iterations.

ρ	$\boldsymbol{\mu}$	θ	GREG		PROP		MR1	
			$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$
0.1	$\boldsymbol{\mu}_I$	0.25	1.21	1.49	0.62	0.69	1.00	1.14
		0.50	1.26	1.44	0.63	0.68	0.99	1.12
		0.75	1.24	1.40	0.62	0.67	0.92	1.06
	$\boldsymbol{\mu}_{II}$	0.25	1.21	1.12	0.62	0.53	1.37	1.27
		0.50	1.26	1.08	0.63	0.52	1.29	1.15
		0.75	1.24	1.06	0.62	0.51	1.16	1.04
0.5	$\boldsymbol{\mu}_I$	0.50	1.03	1.26	0.48	0.54	0.78	0.84
		0.50	0.99	1.27	0.49	0.58	0.80	0.90
		0.75	0.99	1.25	0.50	0.56	0.80	0.90
	$\boldsymbol{\mu}_{II}$	0.50	1.03	0.92	0.48	0.41	1.06	0.83
		0.50	0.99	0.93	0.49	0.44	1.09	0.89
		0.75	0.99	0.91	0.50	0.43	1.09	0.91
0.9	$\boldsymbol{\mu}_I$	0.25	0.49	0.60	0.28	0.35	0.39	0.46
		0.50	0.49	0.61	0.27	0.35	0.40	0.45
		0.75	0.48	0.60	0.28	0.33	0.37	0.43
	$\boldsymbol{\mu}_{II}$	0.25	0.49	0.43	0.28	0.27	0.51	0.37
		0.50	0.49	0.43	0.27	0.27	0.53	0.38
		0.75	0.48	0.42	0.28	0.25	0.53	0.41

Table 2. Unequal strata sizes. unequal probabilities and some large sampling fractions. RRMSE X 100% of levels estimates under different scenarios for 1,000 iterations.

ρ	$\boldsymbol{\mu}$	θ	GREG		PROP		MR1	
			$t=1$	$t=2$	$t=1$	$t=2$	$t=1$	$t=2$
0.1	$\boldsymbol{\mu}_I$	0.25	1.51	1.61	0.84	0.89	1.26	1.21
		0.50	1.51	1.58	0.87	0.88	1.24	1.26
		0.75	1.53	1.58	0.87	0.90	1.23	1.20
	$\boldsymbol{\mu}_{II}$	0.25	1.51	1.26	0.84	0.68	1.67	1.29
		0.50	1.51	1.24	0.87	0.67	1.60	1.27
		0.75	1.53	1.24	0.87	0.68	1.53	1.14
0.5	$\boldsymbol{\mu}_I$	0.25	1.29	1.51	0.70	0.77	0.98	0.98
		0.50	1.32	1.42	0.71	0.74	1.03	0.99
		0.75	1.26	1.48	0.69	0.74	1.04	1.04
	$\boldsymbol{\mu}_{II}$	0.25	1.29	1.14	0.70	0.59	1.30	0.93
		0.50	1.32	1.07	0.71	0.56	1.38	0.98
		0.75	1.26	1.11	0.69	0.56	1.42	1.07
0.9	$\boldsymbol{\mu}_I$	0.25	0.72	0.83	0.38	0.50	0.53	0.61
		0.50	0.70	0.85	0.40	0.49	0.53	0.60
		0.75	0.71	0.84	0.40	0.48	0.53	0.59
	$\boldsymbol{\mu}_{II}$	0.25	0.72	0.59	0.38	0.38	0.69	0.50
		0.50	0.70	0.60	0.40	0.37	0.73	0.51
		0.75	0.71	0.59	0.39	0.37	0.78	0.58

impact on the precision of the proposed estimator. However, for MR1, we observe some slight differences in the RRMSE between different values for θ. For small correlation ($\rho = 0.1$), we indeed have a larger RRMSE for MR1 with $\theta = 0.25$. For larger correlation, the differences are negligible for MR1. With MR1, we notice differences between the RRMSE of $\hat{\tau}_{y_1}$ for small ($\boldsymbol{\mu}_I$) and large changes ($\boldsymbol{\mu}_{II}$). There are hardly any differences for the proposed estimator. These observations are the same for Tables 1 and 2, except that the RRMSE are higher for all estimators in case of unequal strata sizes.

The RRMSE of the proposed estimator does not seem to be affected by the amount of overlap θ, because we can see from the expression (18) that the precision is driven by the correlations between the variables of interest and the auxiliary information for both waves, which is not affected by θ. This can also be seen from the variance (27), where the residuals $\mathbf{M}_{\boldsymbol{\Gamma}}\check{\mathbf{y}}$ do not depend on θ. The information about the rotation is implicitly included within the vector $\mathbf{z}_i$ given by Equation (22), and used for the weights within the regression coefficient (19) (see Equation (20)). These weights ensure efficiency (see Section 4). On the other hand, the precision of MR1 is related to θ, because θ has an impact on the precision of the control totals with MR1. With the proposed method, we use different control totals unaffected by θ.

Let $\triangle = \tau_{y_2} - \tau_{y_1}$ be the change between waves $t=1$ and $t=2$. We propose estimating $\triangle$ by $\hat{\triangle} = \hat{\tau}_{y_2} - \hat{\tau}_{y_1}$, where $\hat{\tau}_{y_1}$ and $\hat{\tau}_{y_2}$ are the corresponding cross-sectional estimators. Tables 3 and 4 give the RRMSE $\times$ 100% of the estimators of changes, for equal and unequal strata sizes. As expected, the RRMSE decreases with ρ. The proposed estimator PROP significantly outperforms GREG and MR2. The efficiency gain compared with MR2

Table 3. Equal strata sizes. equal probabilities and small sampling fractions. RRMSE × 100% of change estimates under different scenarios for 1,000 iterations.

ρ	$\boldsymbol{\mu}$	θ	GREG	PROP	MR2
0.1	$\boldsymbol{\mu}_I$	0.25	120.28	60.60	100.83
		0.50	125.23	67.39	90.90
		0.75	118.08	69.03	82.45
	$\boldsymbol{\mu}_{II}$	0.25	5.86	2.98	5.40
		0.50	6.10	3.31	5.05
		0.75	5.76	3.39	4.37
0.5	$\boldsymbol{\mu}_I$	0.25	98.04	46.96	79.61
		0.50	99.73	51.67	71.26
		0.75	94.29	52.56	64.99
	$\boldsymbol{\mu}_{II}$	0.25	4.76	2.34	4.42
		0.50	4.85	2.57	4.12
		0.75	4.59	2.61	3.60
0.9	$\boldsymbol{\mu}_I$	0.25	46.14	27.00	37.37
		0.50	45.38	26.22	32.06
		0.75	43.62	26.43	28.95
	$\boldsymbol{\mu}_{II}$	0.25	2.23	1.36	2.15
		0.50	2.19	1.32	2.01
		0.75	2.11	1.32	1.74

Table 4. Unequal strata sizes. unequal probabilities and some large sampling fractions. RRMSE × 100% of change estimates under different scenarios for 1,000 iterations.

ρ	$\boldsymbol{\mu}$	θ	GREG	PROP	MR2
0.1	$\boldsymbol{\mu}_I$	0.25	140.96	79.27	123.09
		0.50	138.27	86.79	109.39
		0.75	144.61	93.57	110.40
	$\boldsymbol{\mu}_{II}$	0.25	6.90	3.83	6.43
		0.50	6.76	4.19	5.79
		0.75	7.06	4.51	5.75
0.5	$\boldsymbol{\mu}_I$	0.25	114.65	61.44	95.13
		0.50	113.51	61.60	81.51
		0.75	109.77	62.43	74.85
	$\boldsymbol{\mu}_{II}$	0.25	5.95	3.20	5.48
		0.50	5.89	3.21	4.92
		0.75	5.69	3.25	4.37
0.9	$\boldsymbol{\mu}_I$	0.25	64.40	37.30	48.98
		0.50	61.67	36.67	41.70
		0.75	60.04	35.11	36.51
	$\boldsymbol{\mu}_{II}$	0.25	3.16	1.91	2.96
		0.50	3.03	1.88	2.61
		0.75	2.95	1.80	2.20

ranges from 5% to 53%. Since the relative RMSE is considered, it is not surprising to observe larger RRMSE for a small change ($\boldsymbol{\mu}_I$). The RRMSE of MR2 decreases with θ. In contrast, the RRMSE of PROP increases slightly with θ except for large vales of ρ.

Table 5 shows the relative bias (RB) of the variance estimator (27) for PROP. The RB is defined by

$$\mathrm{RB}\{\hat{V}(\hat{\tau})\} := V(\hat{\tau})^{-1}\left\{\frac{1}{1{,}000}\sum_{r=1}^{1{,}000}\hat{V}(\hat{\tau}_r) - V(\hat{\tau})\right\},$$

where

$$V(\hat{\tau}) := \frac{1}{1{,}000}\sum_{r=1}^{1{,}000}(\hat{\tau}_r - \tau)^2.$$

Here, $\hat{\tau}_r$ and $\hat{V}(\hat{\tau}_r)$ are respectively the point and variance estimate for the r-th iteration. The RB are within an acceptable range. We observe larger RB for $\hat{\tau}_{y_2}^{\mathrm{greg}}$ when $\rho = 0.9$ and $\theta = 0.75$, because the variance is small in this case.

The biases of the variance estimates in the case of unequal strata sizes is larger than the biases of equal strata sizes. The reason is the small sample size for two strata in the unequal strata size scenario. The residuals of the smallest strata vary much more and thus have a larger contribution towards the variance than the residuals of the large strata. The negative bias can also be caused by small sample sizes, because the Taylor linearization method has a tendency to underestimate the true variance in this case (Särndal et al. 1992, 176).

Table 5. RB%100 of variance estimates for the proposed estimator under different scenarios for 1,000 iterations.

ρ	$\boldsymbol{\mu}$	θ	Equal strata sizes		Unequal strata sizes	
			$t=1$	$t=2$	$t=1$	$t=2$
0.1	$\boldsymbol{\mu}_I$	0.25	1.5	−7.4	−16.3	−15.3
		0.50	−3.4	−5.1	−18.1	−12.1
		0.75	1.5	0.0	−15.4	−11.9
	$\boldsymbol{\mu}_{II}$	0.25	1.5	−7.4	−16.4	−15.3
		0.50	−3.4	−5.1	−18.1	−12.1
		0.75	1.5	0.0	−15.2	−11.9
0.5	$\boldsymbol{\mu}_I$	0.25	5.7	4.9	−21.0	−19.3
		0.50	1.8	−8.8	−18.3	−12.9
		0.75	−0.0	−1.4	−12.2	−7.6
	$\boldsymbol{\mu}_{II}$	0.25	5.6	4.9	−20.8	−19.3
		0.50	1.8	−8.8	−18.3	−12.9
		0.75	−0.0	−1.4	−11.5	−7.6
0.9	$\boldsymbol{\mu}_I$	0.25	−2.7	0.3	−16.1	−22.0
		0.50	4.0	−1.4	−18.5	−18.7
		0.75	−2.0	10.1	−15.1	−13.7
	$\boldsymbol{\mu}_{II}$	0.25	−2.7	0.3	−15.7	−22.0
		0.50	4.0	−1.4	−18.5	−18.7
		0.75	−1.9	10.1	−14.3	−13.7

7. Conclusion

We propose a multivariate GREG estimator for estimation of levels and changes. It has the advantage of involving the information from both waves, and takes into account the correlations between the variables of interest and the auxiliaries within and between the waves. Additionally, it also takes the sampling design into account, in terms of stratification, rotation and sampling fractions.

The simulation study shows that the proposed estimator may outperform its competitors, in particular with respect to change estimates. Nevertheless, the advantages of the proposed estimator over the modified estimator are manifold. It does not require any imputation and does not suffer from a drift, unlike the composite estimator. It can be easily implemented using existing statistical software. The variance estimator is simpler than the variance estimator of composite estimators, because neither estimated totals nor imputation is required. It also takes the auxiliary variables and the variables of interest from both waves into account.

Nonresponse and panel attrition are important issues with repeated surveys. It is beyond the scope to tackle these problems fully. Previous wave imputation can be used for the auxiliary variables $\check{\mathbf{x}}_{it}$ which suffer from attrition. Re-weighting could be used to compensate for nonresponse and panel attrition for the variable of interest. In this case the new weight should replace the basic weights $1/\pi_{it}$ within Equations (5) and (6). In this case, s_t would be the sample of respondents at wave t. The proposed estimator (9) can be directly used in this case. It is approximately unbiased, as long as a proper re-weighting technique has been used. However, in this case, the vectors $\mathbf{n}_t$ and $\mathbf{n}_{12}$ are random. Consequently, we may lose the asymptotic optimality, because the asymptotic approximation of Hájek (1964) for the joint-inclusion probabilities are based on fixed $\mathbf{n}_t$ and $\mathbf{n}_{12}$. The variance estimator (27) should be used cautiously, because it does not incorporate nonresponse adjustments. A possible solution would be to incorporate the re-weighting variables within $\check{\mathbf{x}}_{it}$ and use $\check{\mathbf{x}}_{it}$ within Equations (9) and (27). It would be useful to investigate this idea further.

8. Appendix

Proof of Theorem 1:

Since $\boldsymbol{\gamma}_i = (\check{\mathbf{x}}_i^\top, \mathbf{z}_i^\top)^\top$, we have

$$\left(\sum_{i\in s_b} c_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top\right)^{-1} = \left\{\begin{pmatrix}\check{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}^\top \mathbf{C}\begin{pmatrix}\check{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}\right\}^{-1} = \begin{pmatrix}\boldsymbol{\Gamma}_{xx} & \boldsymbol{\Gamma}_{xz}\\ \boldsymbol{\Gamma}_{xz}^\top & \boldsymbol{\Gamma}_{zz}\end{pmatrix},$$

$$\sum_{i\in s_b} c_i \boldsymbol{\gamma}_i \check{\mathbf{y}}_i^\top = \begin{pmatrix}\check{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}^\top \mathbf{C}\check{\mathbf{y}},$$

where

$$\boldsymbol{\Gamma}_{xx} = (\check{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\check{\mathbf{X}})^{-1},$$

$$\boldsymbol{\Gamma}_{zz} = (\mathbf{Z}^\top \mathbf{C}\mathbf{M}_x\mathbf{Z})^{-1},$$

$$\boldsymbol{\Gamma}_{xz} = -\boldsymbol{\Gamma}_{xx}\check{\mathbf{X}}^\top \mathbf{C}\mathbf{Z}(\mathbf{Z}^\top \mathbf{C}\mathbf{Z})^{-1}$$

and $\mathbf{M}_x$ is defined by

$$\mathbf{M}_x = \mathbf{I} - \breve{\mathbf{X}}(\breve{\mathbf{X}}^\top \mathbf{C}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}^\top \mathbf{C}. \tag{32}$$

Now, we have

$$\hat{\mathbf{B}}_\gamma = \begin{pmatrix} \hat{\mathbf{B}}_x \\ \boldsymbol{\Gamma}_{xz}^\top \breve{\mathbf{X}}^\top \mathbf{C}\breve{\mathbf{y}} + \boldsymbol{\Gamma}_{zz}\mathbf{Z}^\top \mathbf{C}\breve{\mathbf{y}} \end{pmatrix}, \tag{33}$$

because

$$\begin{aligned} \boldsymbol{\Gamma}_{xx}\breve{\mathbf{X}}^\top \mathbf{C}\breve{\mathbf{y}} + \boldsymbol{\Gamma}_{xz}\mathbf{Z}^\top \mathbf{C}\breve{\mathbf{y}} &= \boldsymbol{\Gamma}_{xx}\breve{\mathbf{X}}^\top \mathbf{C}\{\mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{C}\}\breve{\mathbf{y}} \\ &= \boldsymbol{\Gamma}_{xx}\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\breve{\mathbf{y}} \\ &= (\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\breve{\mathbf{y}} \\ &= \hat{\mathbf{B}}_x. \end{aligned}$$

Finally, Equations (11) and (12) imply that $\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma = \{(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x)^\top, 0\}^\top$ Thus, by using Equation (33), we obtain Equation (18).

9. References

Australian Bureau of Statistics. 2007. *Forthcoming changes to labour force statistics:* 6292.0, Australian Bureau of Statistics, Canberra, Australia. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/6292.0 (accessed July 2022).

Bell, P. 2001. "Comparison of alternative labour force survey estimators." *Survey Methodology* 27: 53–63. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5854-eng.pdf.

Berger, Y.G. 2004. "Variance estimation for measures of change in probability sampling." *Canadian Journal of Statistics* 32: 451–467. DOI: https://doi.org/10.2307/3316027.

Berger, Y.G., J.F. Muñoz, and E. Rancourt. 2009. "Variance estimation of survey estimates calibrated on estimated control totals – an application to the extended regression estimator and the regression composite estimator." *Computational Statistics and Data Analysis* 53: 2596–2604. DOI: https://doi.org/10.1016/j.csda.2008.12.011.

Berger, Y.G., M.E.H. Tirari, and Y. Tillé. 2003. "Towards optimal regression estimation in sample surveys." *Australian & New Zealand Journal of Statistics* 45: 319–329. DOI: https://doi.org/10.1111/1467-842X.00286.

Bonnéry, D., Y. Cheng, and P. Lahiri. 2020. "An evaluation of design-based properties of different composite estimators." *Statistics in Transition New Series* 21: 166–190. DOI: https://doi.org/10.21307/stattrans-2020-037.

Cassel, C.-M., C. Särndal, and J. Wretman. 1977. *Foundations of inference in survey sampling*. New York: Wiley.

Deville, J.C., and Y. Tillé. 2005. "Variance approximation under balanced sampling." *Journal of Statistical Planning and Inference* 128: 569–591. DOI: https://doi.org/10.1016/j.jspi.2003.11.011.

Eurostat. 2012. *European union statistics on income and living conditions* (EU-SILC). Available at: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions (accessed July 2022).

Fuller, W. 1990. "Analysis of repeated surveys." *Survey Methodology* 16: 167–180. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199000214537.

Fuller, W.A., and J. Rao. 2001. "A regression composite estimator with application to the canadian labour force survey." *Survey Methodology* 27: 45–52. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5853-eng.pdf.

Gambino, J., B. Kennedy, and M.P. Singh. 2001. "Regression composite estimation for the canadian labour force survey: Evaluation and implementation." *Survey Methodology* 27: 65–74. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5855-eng.pdf.

Gambino, J.G., and P.L.N. Silva. 2009. "Sampling and estimation in house-hold surveys." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao. 29A of Handbook of Statistics: 407–439. Amsterdam: Elsevier.

Guandalini, A., and Y. Tillé. 2017. "Design-based estimators calibrated on estimated totals from multiple surveys." *International Statistical Review* 85: 250–269. DOI: https://doi.org/10.1111/insr.12160.

Gurney, M.A., and J.F. Daly. 1965. "A multivariate approach to estimation in periodic sample surveys." In *Proceedings of the Section of Survey and Research Methods*, American Statistical Association, September 1965: 242–257, Philadelphia USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Hájek, J. 1964. "Asymptotic theory of rejective sampling with varying probabilities from a finite population." *The Annals of Mathematical Statistics* 35: 1491–1523. DOI: https://doi.org/10.1214/aoms/1177700375.

Hájek, J. 1981. *Sampling from a Finite Population*. New York: Marcel Dekker.

Hansen, M., Hurwitz, W., and W. Madow. 1953. *Sample survey methods and theory*, vol. I and II. New York: Wiley.

Horvitz, D.G., and D.J. Thompson. 1952. "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association* 47: 663–685. DOI: https://doi.org/10.1080/01621459.1952.10483446.

Isaki, C.T., and W.A. Fuller. 1982. "Survey design under the regression superpopulation model." *Journal of the American Statistical Association* 77: 89–96. DOI: https://doi.org/10.1080/01621459.1982.10477770.

Kalton, G. 2009. "Design for surveys over time." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 29A of *Handbook of Statistics*: 89–108. Amsterdam: Elsevier.

Kumar, S., and H. Lee. 1983. "Evaluation of composite estimation for the canadian labor force survey." *Survey Methodology* 9: 403–408. Availabe at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198300214342.

Montanari, G. 1987. "Post sampling efficient qr-prediction in large sample survey." *International Statistical Review* 55: 191–202. DOI: https://doi.org/10.2307/1403195.

Särndal, C.E. 1980. "On π-inverse weighting versus best linear unbiased weighting in probability sampling." *Biometrika* 67: 639–650. DOI: https://doi.org/10.1093/biomet/67.3.639.

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

Singh, A.C. 1996. "Combining information in survey sampling by modified regression." . In Proceedings of the Section on Survey Research Methods, American Statistical Association. August 1996: 120–129. Chicago, USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Singh, A.C., B. Kennedy, and S. Wu. 2001. "Regression composite estimation for the canadian labour force survey with a rotating panel design." *Survey Methodology* 27: 33–44. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5852-eng.pdf.

Singh, A.C., B. Kennedy, S. Wu, and F. Brisebois. 1997. "Composite estimation for the canadian labour force survey." In *Proceedings of the Survey Research Methods Section*, American Statistical Association, August 1997: 300–3005, Anaheim, USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Smith, P., M. Pont, and T. Jones. 2003. "Developments in business survey methodology in the office for national statistics, 1994–2000." *Journal of the Royal Statistical Society.* 52: 257–295. DOI: https://doi.org/10.1111/1467-9884.03571.

Steel, D., and R. Clark. 2007. "Person-level and household-level regression estimation in household surveys." *Surveys Methodology* 33: 55–60. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20070019852.

Steel, D., and C. McLaren. 2008. *Design and analysis of repeated surveys*. Centre for Statistical and Survey Methodology, University of Wollongong: 11-08. Available at: http://ro.uow.edu.au/cgi/viewcontent.cgi?article = 1009&context = cssmwp (accessed July 2022).

Steel, D., and C. McLaren. 2009. "Design and analysis of surveys repeated over time." In *Handbook of statistics*, edited by C.R. Rao. 289–313 Elsevier.

Wright, R.L. 1983. "Finite population sampling with multivariate auxiliary information." *Journal of the American Statistical Association* 78: 879–884. DOI: https://doi.org/10.1080/01621459.1983.10477035.

Yansaneh, I.S., and W.A. Fuller. 1998. "Optimal recursive estimation for repeated surveys." *Survey Methodology* 24: 31–40. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1998001/article/3907-eng.pdf.

Received December 2021
Revised April 2021
Accepted August 2022

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 45–78, http://dx.doi.org/10.2478/JOS-2023-0003

Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration

Zhan Liu[1] *and Richard Valliant*[2]

Matching a nonprobability sample to a probability sample is one strategy both for selecting the nonprobability units and for weighting them. This approach has been employed in the past to select subsamples of persons from a large panel of volunteers. One method of weighting, introduced here, is to assign a unit in the nonprobability sample the weight from its matched case in the probability sample. The properties of resulting estimators depend on whether the probability sample weights are inverses of selection probabilities or are calibrated. In addition, imperfect matching can cause estimates from the matched sample to be biased so that its weights need to be adjusted, especially when the size of the volunteer panel is small. Calibration weighting combined with matching is one approach to correct bias and reduce variances. We explore the theoretical properties of the matched and matched, calibrated estimators with respect to a quasirandomization distribution that is assumed to describe how units in the nonprobability sample are observed, a superpopulation model for analysis variables collected in the nonprobability sample, and the randomization distribution for the probability sample. Numerical studies using simulated and real data from the 2015 US Behavioral Risk Factor Surveillance Survey are conducted to examine the performance of the alternative estimators.

Key words: Calibration adjustment; doubly robust estimation; nearest neighbour matching; sample matching; target sample; volunteer panels.

1. Introduction

Probability samples have been the standard for finite population estimation for many decades. However, probability samples can have many nonsampling problems like low contact and response rates or missing data for units that do respond. Response rates in US surveys, in particular, have been declining for at least two decades (Brick and Williams 2013). Since nonprobability samples can be faster and cheaper to administer and collect, some organizations are gravitating toward their use (Terhanian and Bremer 2012). Baker et al. (2013) review the reasons that nonprobability samples, like volunteer internet panels, may be used rather than a probability sample. Among them are lower costs and compressed data collection periods. Quick turnaround can be especially important to gauge public well-being in health crises like the COVID-19 pandemic of 2020.

There are a variety of problems with nonprobability samples, especially among persons in a panel that have been recruited to participate in future surveys (e.g., see Baker et al. 2013; Valliant and Dever 2011; Valliant et al. 2018). These include selection bias, coverage error, panel nonresponse, attrition, and measurement error. We concentrate on the use of matching and calibration to adjust for the first two of these−selection bias and

[1] School of Mathematics and Statistics, Hubei University, China. Email: eleen_20040109@163.com
[2] Universitiy of Michigan, Institute for Social Research, 4620 North Park Avenue Apt 1406W, Chevy Chase, Maryland, 20815, U.S.A. Email: valliant@umich.edu

coverage error. Selection bias occurs if the sample differs from the nonsample in such a way that the sample cannot be projected to the full population without some type of statistical adjustment. Coverage error can occur if, for example, a volunteer panel consists of only persons with access to the Internet, assuming that the entire population of a country is the target of the survey. Other, more subtle forms of coverage error can occur if certain demographic groups would rarely or never participate in the particular type of nonprobability survey being conducted.

Because the selection of a nonprobability sample is not controlled by a survey designer, estimation methods other than standard design-based approaches are needed. At least six alternatives can be considered for weighting and estimation with nonprobability samples:

(1) Naïve method where all units are assigned the same weight,
(2) Quasi-randomization where a pseudo-inclusion probability is estimated for each nonprobability unit,
(3) Superpopulation modeling of analysis variables (Y's),
(4) Doubly robust estimation where quasi-randomization and superpopulation modeling are combined,
(5) Mass imputation of Y's into a probability sample using values from a nonprobability sample to form an imputation model,
(6) Matching of a nonprobability sample to a probability sample whose units are used as donors of weights to the nonprobability sample.

The naïve method of equal weighting is rarely, if ever, appropriate because nonprobability samples are not generally distributed proportionally across demographic or other important groups in the population. Alternatives (2)–(4) were reviewed by Elliott and Valliant (2017) and Valliant (2020) and further studied by Chen et al. (2020). Wang et al. (2020) refined alternative (2) by kernel-smoothing the propensity weights. Alternative (5) was proposed by Kim et al. (2021) and involves fitting an imputation model using data from the nonprobability sample and imputing Y values to the units in the probability sample using that model. The probability sample with imputed values is provided to analysts but not the nonprobability sample. Mass imputation solves the weighting problem by using the weights associated with the probability sample. The dissertation of Wang (2020) studied a version of (6) in which a kernel-smoothing method was used to proportionally distribute the probability sample weights to units in the nonprobability sample. We study another, somewhat simpler version of alternative (6), and particularly address some problems with the method.

1.1. Notation and Models Used for Analysis

Both a probability sample, denoted by S_p and a nonprobability sample, denoted by S_{np} will be used in subsequent sections. The target population for which estimates are made is U and has N units. To examine properties of estimators, three distributions will be used. Expectations taken with respect to the sample design used to select the probability sample S_p will be denoted by a π subscript. The probability of selection of unit i in S_p is π_i. To analyze the nonprobability sample S_{np}, we assume that its units are selected by an unknown quasi-randomization distribution; expectations taken with respect to that

distribution will be specified by a subscript R. The probability that unit j is included in S_{np} is $R(\mathrm{x}_j)$ where x_j is a C-vector of covariates or auxiliaries associated with unit j. To simplify notation in later sections, we set $R(\mathrm{x}_j) \equiv R_j$. The analysis variable Y will also be assumed to be generated by a superpopulation model, ξ. Consider the linear model for Y_i defined by

$$Y_i = \mathbf{x}_i\beta + \varepsilon_i (i \in U), \tag{1}$$

where β is a $C \times 1$ parameter vector, and the ε_i are independent, random errors with mean zero and variance σ_i^2. Theory for nonlinear models can also be worked out, as in Chen et al. (2020), but a linear model is convenient for purposes here. Under model (1), the expected value of the population total, $Y_U = \sum_{i \in U} Y_i$, is $E_\xi(Y_U) = \mathbf{X}_U\beta$ where $\mathbf{X}_U = \sum_{i \in U} \mathrm{x}_i$.

The remainder of the article is organized as follows. Section 2 describes how matching can be applied to obtain basic weights for units in the nonprobability sample and reviews the methods of matching. Section 3 presents the theory for bias and variance in different situations. Section 4 investigates properties of matched estimators when the nonprobability sample is calibrated to population totals of covariates. In Section 5, the sample matching and the calibration adjustment are applied in a simulation study using artificial data. In Section 6, an application to a real population is conducted to evaluate the performance of the proposed estimates. The last section summarizes our findings.

2. Applications of Matching

Sample matching has been an option for estimating treatment effects in causal inference for some time (e.g., see Cochran 1953; Rubin 1973; Rosenbaum and Rubin 1983). Moreover, it has been widely applied in evaluation research, observational studies and epidemiological studies (Rothman et al. 2008). More recently, it also has been applied as a way of identifying a sample in market research, public opinion surveys (e.g., Vavreck and Rivers 2008; Terhanian and Bremer 2012), and other nonprobability sampling surveys, especially using volunteer panel surveys. Baker et al. (2013) review some of the applications of matching in survey sampling. Its purpose in non-probability sampling surveys is to reduce selection bias and to estimate population characteristics. Another application of statistical matching is to overcome the problem of missing data created when some persons do not consent to having their survey responses linked to administrative databases (Gessendorfer et al. 2018).

The basic idea of sample matching in survey sampling is that first a random, probability sample, S_p, is selected from the sampling frame of the target population. This probability sample is matched to a pool of nonprobability cases, for example, a volunteer panel of persons. The resulting matched sample from the nonprobability pool is denoted by S_{np}. The probability sample should have none of the coverage problems of the nonprobability sample. This probability sample is also called a reference sample (Lee 2006) and can be an existing survey (or subsample of one) rather than one specially conducted to serve as the reference. For example, in the US the ACS, American Community Survey (U.S. Census Bureau 2023) is one possibility for a large, well-conducted household, reference survey. The probability sample should be representative of the target population in the sense that it can be used to make unbiased and/or consistent estimates of population quantities.

We assume that S_p does not include the Y variables for which estimates are to be made; these are collected from the nonprobability sample.

The application of matching described by Rivers (2007) is one in which S_p is a simple random sample (*srs*). The nonprobability sample S_{np} is obtained by a one-to-one match of S_p to a much larger pool of nonprobability cases, yielding a set S_{np} of the same size as S_p. Since S_p was treated as an *srs*, every unit in S_{np} was given the same weight. When S_p is an *srs*, the distribution across various characteristics of S_{np} is expected to be the same as that of the population. However, in an evaluation of the nonprobability samples offered by nine commercial vendors, Kennedy et al. (2016) found that a nonprobability sample may still produce biased estimators even though it had the same demographic distribution as the population. In other words, matching to an *srs* S_p to obtain S_{np} can be inadequate without further weighting. Rivers and Bailey (2009) describe an election polling application where the sample was obtained by matching, as described above, but inverses of estimated propensities of being in the nonprobability sample were used as weights.

Sample matching in alternative (6), as applied in this article, fits into the quasi-randomization approach. Each unit in a probability sample is matched to a unit in the nonprobability sample based on a set of covariates. The logical extension of Rivers (2007) is for the probability sample unit to "donate" its weight to the matched, nonprobability sample unit. The intuitive argument to justify this is that if the nonprobability units match the probability units on an extensive list of covariates, then the S_{np} units are exchangeable for the S_p units, S_{np} constitutes the same sort of sample as S_p, and the units in S_{np} can be weighted in the same way. This approach has the advantage of straightforward retention of all analytic data collected in the nonprobability sample unlike alternative (5) which could require a separate imputation model for every Y variable.

The probability sample used for matching can be larger, smaller, or equal in size to the non-probability sample, although the method in which S_{np} is selected to have the same size as S_p has advantages. If a pool of nonprobability units is used that is much larger than the probability sample, finding a close match for each unit in the probability sample may be more feasible. This would be the case when a large panel of volunteers has been accumulated. For example, the method used by many panel vendors now is to stockpile a large set of persons who have agreed to participate in online surveys when requested. For example, (see https://www.ipsos.com/en-au/ipsos-online-access-panels). Toluna, the parent company of Harris Interactive, does something similar. Their promotional material says that it has a panel of over 40 million consumers in 70 markets (see Toluna 2023).

If the nonprobability sample is smaller than S_p, a unit in S_{np} may be matched to more than one unit in S_p, making it unclear how to weight the S_{np} cases. In this article, we assume that the resulting sample size of the matched, nonprobability sample, S_{np}, equals the sample size of the probability sample, S_p, that it is matched against. Denote this sample size by n.

2.1. *Methods of Matching*

Which matching algorithm to use is a question. There are various algorithms, including nearest neighbour matching, caliper and radius matching, stratification and interval matching, as well as kernel and local linear matching (Caliendo and Kopeinig 2008).

Among these matching algorithms, nearest neighbour matching based on Euclidean distance is most straightforward. It contains, as special cases, single nearest neighbour matching without replacement, single nearest neighbour matching with replacement and multiple nearest neighbour matching. In single nearest neighbour matching, for a unit in S_p only one unit from the nonprobability pool can be chosen as its matching unit based on the covariates present in both data sets. If single nearest neighbour matching is done without replacement, a unit in the nonprobability pool can be chosen only once as a match. Single nearest neighbour matching without replacement may, however, have poor performance when the target sample and the volunteer panel have very different covariate distributions (Dehejia and Wahba 2002). To overcome this problem, single nearest neighbour matching with replacement and multiple nearest neighbour matching were proposed to increase the average quality of matching and reduce the bias (Smith and Todd 2005). Other matching methods have been suggested that use more than one unit in the nonprobability pool as the matching unit for an individual in the probability sample. Caliper and radius matching use this approach.

Another issue in sample matching is that the matching process will become relatively more difficult as the number of relevant covariates increases. This is the curse of dimensionality noted by Rosenbaum and Rubin (1983). In order to solve this problem, they propose the propensity score, which is the conditional probability of receiving a treatment given the covariates X, denoted by $p(X) = P(D = 1|X)$, where D is the binary indicator taking either the value 1 (receiving treatment, e.g., participation in a volunteer panel) or 0 (not receiving treatment). Rosenbaum and Rubin (1983) have proved that matching on the propensity score $p(X)$ is also valid when it is valid to match on the covariates X. Compared with direct matching based on all covariates, propensity score matching can reduce multiple dimensions (many covariates) to a single dimension, greatly simplifying the matching process. Consequently, it has been widely used in medical and epidemiological studies, economics, market research and a host of other fields (Schonlau et al. 2009; Baker et al. 2013).

3. Estimation from Matched Sample

In this section, we introduce estimators of means and totals based on the matched sample S_{np} and derive their properties. An estimator of a population total is

$$\hat{Y}_M = \sum_{j \in S_{np}} \tilde{w}_j y_j, \tag{2}$$

where $\tilde{w}_j$ is the weight from the probability sample unit that is matched to unit j in S_{np} and y_j is the Y value observed for that unit. We assume that these weights are appropriately scaled for estimating population totals. In particular, $\hat{N}_M = \sum_{j \in S_{np}} \tilde{w}_j$ is an estimator of N, the size of the target population. The mean of Y is estimated by $\hat{\bar{Y}}_M = \hat{Y}_M / \hat{N}_M$. We will consider two cases of weighting of the probability sample S_p:

(1) The weight used for each unit in S_p is the inverse of the selection probability of that unit, that is, $\tilde{w}_j = \pi_j^{-1}$; the estimator of the total with this weight is denoted by $\hat{Y}_{M1}$ subsequently;

(2) The S_p weights are those for a general regression (GREG) estimator; the estimator with this weight is denoted by $\hat{Y}_{M2}$.

Case (1), in which the S_{np} weight is the inverse of the selection probability of its matched unit in S_p is the simplest baseline case for S_{np} weighting. If $\hat{Y}_{M1}$ is biased in this case, then calibration of S_{np} must correct that flaw. Note that the GREG in case (2) includes the commonly used poststratification estimator and requires that the population total $\mathbf{X}_U$ of the covariate vector be known. Whether the weights above are related to the pseudo-inclusion probabilities of the units in S_{np} largely determines whether $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$ are biased or not as shown below. The arguments given are largely heuristic, although they can be formalized using technical conditions like those in Chen et al. (2020).

Properties of estimators can be calculated in several ways: with respect to the ξ-model only, with respect to the R-distribution only, with respect to the π-distribution, or with respect to a combination of the distributions. In subsequent sections, we compute biases and variances with respect to the combined $R\pi$-distribution. The $R\pi$ calculation is analogous to the design-based calculations used in much of sampling theory. In addition, bias and variance calculations are made with respect to the ξ-model and combined $R\pi\xi$-distributions. The calculations made with respect to the ξ-distribution are conditional on the S_{np} and S_p samples. In principle, ξ calculations are more reflective of the statistical properties for the particular sets of units in S_{np} and S_p.

3.1. Bias of the Matched Estimator for Case (1)

Taking the expectation of $\hat{Y}_{M1} - Y_U$ under case (1) with respect to the pseudo-randomization distribution only gives

$$E_R\left(\hat{Y}_{M1} - Y_U\right) = E_R\left(\sum_{j \in S_{np}} \pi_j^{-1} y_j\right) - Y_U$$
$$= \sum_{j \in U} \frac{R_j}{\pi_j} y_j - Y_U.$$

If $R_j = \pi_j$, then $\hat{Y}_M$ will be R-unbiased. However, this does not have to be true generally. For example, if $R_j = Pr(j \in S_{np} | \mathbf{x}_j)$ is a complicated logistic function of a set of covariates that were not used in determining $\tilde{w}_j$, the estimator is R-biased. In this case, some approaches can be employed to reduce bias. For example, some important covariates can first be selected by establishing the relationship models between the S_p weights and covariates, then sample matching can be conducted based on these important covariates. Another situation leading to R-bias would be when the pseudo-inclusion mechanism is *nonignorable*, that is, $Pr(j \in S_{np} | \mathbf{x}_j, y_j) \neq Pr(j \in S_{np} | \mathbf{x}_j)$. Since in a probability sample, the selection mechanism is always ignorable ($\tilde{w}_j \neq 1/Pr(j \in S_{np} | \mathbf{x}_j, y_j)$) when inclusion in the nonprobability sample depends on Y.

If the expectation is taken over the Y-model, the result is

$$E_\xi\left(\hat{Y}_{M1} - Y_U | S_{np}, S_p\right) = \left(\hat{\mathbf{X}}_{np}(\pi) - \mathbf{X}_U\right)\beta,$$

where $\hat{X}_{np}(\pi) = \sum_{S_{np}} \mathbf{x}_j / \pi_j$. The ξ-bias is non-zero unless $\hat{\mathbf{X}}_{np}(\pi) = \mathbf{X}_U$. If $\hat{\mathbf{X}}_{np}(\pi)$ is an unbiased estimator of $\mathbf{X}_U$ under the quasi-randomization R-distribution, $\hat{Y}_{M1}$ will be

unbiased when averaged over both the $R-$ and ξ-distributions (and, equivalently, over the R, π, and ξ distributions). But, as for $\hat{Y}_{M1}$, $\hat{\mathbf{X}}_{np}(\pi)$ will be biased if the correct R-model is not linked to the S_p weights, that is, if $\tilde{w}_j = \pi_j^{-1} \neq 1/R_j$.

3.2. Bias of the Matched Estimator for Case (2)

If the weights from the probability sample have been calibrated to population totals of some covariates $\mathbf{x}$, the bias calculation changes some what. Take the case of the general regression (GREG) estimator being used for S_p. That is, $\tilde{w}_j = g_j/\pi_j$ where

$$g_j = 1 + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \mathbf{x}_j / \tilde{\sigma}_j^2 \tag{3}$$

with $\hat{\mathbf{X}}_p = \sum_{S_p} \mathbf{x}_j/\pi_j$ and $\tilde{\mathbf{A}}_p = \sum_{S_p} \mathbf{x}_j \mathbf{x}_j^T / \left(\pi_j \tilde{\sigma}_j^2\right)$. The values of $\tilde{\sigma}_j^2$ are often set to a constant in practice, but for completeness, we include $\tilde{\sigma}_j^2$ in subsequent formulas. If $\tilde{\sigma}_j^2$'s are used in estimators of totals, they will be generally assumed values of the model variances in Equation (1); but, we do not require that $\tilde{\sigma}_j^2 = \sigma_j^2$. Note also that the π_j's must be available separately for each unit in the probability sample in order to recover $\mathbf{A}_p$ separately from the $\tilde{w}_j$. In some public-use files, users may only be presented with the $\tilde{w}_j = g_j/\pi_j$ and not π_j.

The estimator of the total is then

$$\hat{Y}_{M2} = \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \sum_{j \in S_{np}} \mathbf{x}_j y_j / \left(\pi_j \tilde{\sigma}_j^2\right), \tag{4}$$

where $\hat{Y}_{np}(\pi) = \sum_{j \in S_{np}} y_j/\pi_j$. This ξ-bias is

$$E_\xi\left(\hat{Y}_{M2} - Y_U | S_{np}, S_p\right) = \hat{\mathbf{X}}_{np}(\pi)\beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \tilde{\mathbf{A}}_{np}(\pi)\beta - \mathbf{X}_U\beta \tag{5}$$

where $\tilde{\mathbf{A}}_{np}(\pi) = \sum_{S_{np}} \mathbf{x}_j \mathbf{x}_j^T / \left(\pi_j \tilde{\sigma}_j^2\right)$. Thus $\hat{Y}_{M2}$ is ξ-model biased even though the weights in S_p are calibrated on the x's. The R-expectation (which is also the $R\pi$-expectation) is

$$E_R\left(\hat{Y}_{M2}\right) = \sum_U \frac{R_j}{\pi_j} y_j + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \sum_U \frac{R_j}{\pi_j} \frac{\mathbf{x}_j y_j}{\tilde{\sigma}_j^2}$$

which is also generally not equal to Y_U.

If $R_j = \pi_j$ and sampling for S_{np} and S_p is ignorable, reasonable assumptions are that $N\tilde{\mathbf{A}}_p^{-1}$ and $N^{-1}\tilde{\mathbf{A}}_{np}(\pi)$ converge in probability to $N^{-1}\tilde{\mathbf{A}}_U = N^{-1}\sum_U \mathbf{x}_j \mathbf{x}_j^T/\tilde{\sigma}_j^2$. (See assumption (v) in the Appendix). In that case, $\tilde{\mathbf{A}}_p^{-1}\tilde{\mathbf{A}}_{np}(\pi) \xrightarrow{p} \mathrm{I}_c$ with $\mathbf{I}_C$ being the $C \times C$ identity matrix, and $E_\xi\left(\hat{Y}_{M2} - Y_U | S_{np}, S_p\right) \to \left(\hat{\mathbf{X}}_{np}(\pi) - \hat{\mathbf{X}}_p\right)\beta$.

Taking the expectation of Equation (5) with respect to the R- and π-distributions shows that $\hat{Y}_{M2}$ is approximately $R\pi\xi$-unbiased, but this depends on $R_j = \pi_j$ for all units in S_{np}. Under the same conditions (i.e., $R_j = \pi_j$ and $N\tilde{\mathbf{A}}_p^{-1}$ and $N^{-1}\tilde{\mathbf{A}}_{np}(\pi)$ converging),

$$E_R\left(\hat{Y}_{M2}\right) \doteq Y_U + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)\tilde{\mathbf{B}}_U$$

where $\tilde{\mathbf{B}}_U = \tilde{\mathbf{A}}_U^{-1} \sum_U \mathbf{x}_j y_j / \tilde{\sigma}_j^2$. Consequently, $E_R E_\pi\left(\hat{Y}_{M2}\right) \doteq Y_U$, assuming that $\hat{\mathbf{X}}_p$ is π-unbiased. Similarly, $E_R E_\pi E_\xi\left(\hat{Y}_{M2} - Y_U\right) = 0$.

The results in Subsections 3.1 and 3.2 can be summarized as follows:

- Case (1), $\tilde{w}_j$ is the inverse of the selection probability for its matched unit in the probability sample, $S_{p,}$ that is, $\tilde{w}_j = \pi_j^{-1}$
  - $\hat{Y}_{M1}$ is ξ-biased when the linear model (1) holds;
  - $\hat{Y}_{M1}$ is $R\pi$-unbiased if $R_j = \pi_j$, that is, the probability of a unit's being in the nonprobability sample, S_{np}, equals its probability of being in the probability sample, S_p;
  - $\hat{Y}_{M1}$ is $R\pi\xi$-unbiased when the linear model (1) holds and $R_j = \pi_j$;
- Case (2), $\tilde{w}_j$ is the GREG weight for its matched unit in S_p
  - $\hat{Y}_{M2}$ is ξ-biased under Equation (1) even though S_p is calibrated on the x's in the model;
  - $\hat{Y}_{M2}$ is R-biased in general;
  - $\hat{Y}_{M2}$ is approximately $R\pi$-unbiased and $R\pi\xi$-unbiased if $R_j = \pi_j$;

The key requirement (in addition to ignorability) for unbiasedness of any type is that the observation probability of a unit in the nonprobability sample should be equal to the selection probability of its matched unit from the probability sample. This seems unlikely to be exactly true in most applications.

3.3. Variance of the Matched Estimator in Case (1)

Since a variance estimator is mainly useful in a situation where a point estimator is unbiased or consistent, we concentrate on the case where $R_j = \pi_j$ and $\hat{Y}_{M1}$ is R-unbiased. Calculation of the variance of $\hat{Y}_{M1}$ with respect to the pseudo-inclusion probability distribution depends on the joint distribution of the indicators, $\{\delta_j\}_{j\in U}$ where $\delta_j = 1$ if $j \in S_{np}$ and 0 if not. If the δ_j have the same joint distribution as that of the indicators for being in the probability sample, then S_{np} can be treated as having the same sample design as S_p. If so, $V_R\big(\hat{Y}_{M1}\big) = V_\pi\big(\hat{Y}_{M1}\big)$, and the variance estimator for $\hat{Y}_{M1}$ would be determined by the sample design for S_p. For example, if the probability sample was a stratified, cluster sample, then the variance estimator appropriate to that design would be used.

When $\tilde{w}_j = \pi_j^{-1}$, a more realistic assumption, given the way that nonprobability samples are often acquired, is to treat the $\{\delta_j\}_{j\in U}$ as being independent. With that assumption, the R-variance can be estimated with a formula appropriate for a Poisson sample. Another option is the formula for a sample selected with replacement and with probabilities equal to $R_j = \pi_j$:

$$v_{R\pi}\big(\hat{Y}_{M1}\big) = \frac{n}{n-1}\sum_{j\in S_{np}}\left(\tilde{w}_j y_j - \frac{1}{n}\sum_{j'\in S_{np}}\tilde{w}_{j'}y_{j'}\right)^2. \tag{6}$$

Because $\hat{Y}_{M1}$ does not depend on S_p, $V_{R\pi}(\hat{Y}_{M1}) = V_R(\hat{Y}_{M1})$ and Equation (6) can be interpreted as an estimator of either. Estimator (6) is convenient because it is the default in survey software packages like R survey, Stata, and SAS. However, as shown in Appendix (Subsection 8.1), $v_{R\pi}$ is a biased estimator of the model variance given below.

The ξ-variance under Equation (1) in case (1) is $V_\xi\big(\hat{Y}_{M1}\big) = \left(\sum_{S_{np}}\sigma_j^2/\pi_j^2\right)$ which can be estimated by

$$v_\xi\big(\hat{Y}_{M1}\big) = \sum_{S_{np}}\frac{e_j^2}{\pi_j^2}, \tag{7}$$

where $e_j^2 = (y_j - \mathbf{x}_j^T \hat{\tilde{\mathbf{B}}}_{np}(\pi))^2$ is an approximately ξ-unbiased estimator of σ_j^2 with

$$\hat{\tilde{\mathbf{B}}}_{np}(\pi) = \left(\sum_{S_{np}} \mathbf{x}_j \mathbf{x}_j^T / \left(\pi_j \tilde{\sigma}_j^2 \right) \right)^{-1} \sum_{S_{np}} \mathbf{x}_j y_j / \left(\pi_j \tilde{\sigma}_j^2 \right)$$

(MacKinnon and White 1985). Note that, because the Y's are not available in the probability sample, we must estimate β from the nonprobability sample.

The $R\pi\xi$-variance, in general, is equal to

$$\begin{aligned} V_{R\pi\xi}\left(\hat{Y}_{M1} | S_p, S_{np}\right) &= E_R\left\{ V_{\pi\xi}\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\} + V_R\left\{ E_{\pi\xi}\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\} \\ &= E_R\left\{ E_\pi V_\xi\left(\hat{Y}_{M1} | S_p, S_{np}\right) + V_\pi E_\xi\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\} + \\ &\quad V_R\left\{ E_\pi E_\xi\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\}. \end{aligned} \tag{8}$$

For case (1) $\hat{Y}_{M1}$ does not depend on S_p and Equation (8) reduces to $V_{R\pi\xi}\left(\hat{Y}_{M1} | S_p, S_{np}\right) = V_{R\xi}\left(\hat{Y}_{M1} | S_p, S_{np}\right) = E_R\left\{ V_\xi\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\} + V_R\left\{ E_\xi\left(\hat{Y}_{M1} | S_p, S_{np}\right) \right\}$.

As shown in Appendix (Subsection 8.1) for case (1) with $R_j = \pi_j$, this is

$$V_{R\pi\xi}\left(\hat{Y}_{M1}\right) = \sum_U \frac{\sigma_j^2}{\pi_j} + \beta^T V_R\left(\hat{\mathbf{X}}_{np}\right)\beta. \tag{9}$$

Notice that, even though $\hat{Y}_{M1}$ does not directly depend on x, the $R\pi\xi$-variance does after accounting for the ξ-model structure. Expression (9) can be estimated by

$$\upsilon_{R\pi\xi}\left(\hat{Y}_{M1}\right) = \sum_{S_{np}} \frac{e_j^2}{\pi_j^2} + \hat{\tilde{\mathbf{B}}}_{np}(\pi)^T \upsilon_R\left(\hat{\mathbf{X}}_{np}\right) \hat{\tilde{\mathbf{B}}}_{np}(\pi), \tag{10}$$

where $\upsilon_R\left(\hat{\mathbf{X}}_{np}\right)$ is, for example, a version of Equation (6) adapted to estimate a covariance matrix.

3.4. Variance of the Matched Estimator in Case (2)

The ξ-model variance is $V_\xi\left(\hat{Y}_{M2}\right) = \sum_{S_{np}} \left(g_j/\pi_j\right)^2 \sigma_j^2$, which can be estimated by

$$\upsilon_\xi\left(\hat{Y}_{M2}\right) = \sum_{S_{np}} \left(g_j/\pi_j\right)^2 e_j^2. \tag{11}$$

As noted in Appendix (Subsection 8.2), the estimator of total can be approximated by

$$\hat{Y}_{M2} \doteq \bar{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)\tilde{\mathbf{B}}_U \tag{12}$$

and the approximate $R\pi$ variance is

$$V_{R\pi}\left(\hat{Y}_{M2}\right) \doteq V_R\left(\hat{Y}_{np}(\pi)\right) + \tilde{\mathbf{B}}_U^T V_\pi\left(\hat{\mathbf{X}}_p\right)\tilde{\mathbf{B}}_U, \tag{13}$$

which can be estimated as

$$\upsilon_{R\pi}\left(\hat{Y}_{M2}\right) = \upsilon_R\left(\hat{Y}_{np}(\pi)\right) + \hat{\tilde{\mathbf{B}}}_{np}(\pi)^T \upsilon_\pi\left(\hat{\mathbf{X}}_p\right) \hat{\tilde{\mathbf{B}}}_{np}(\pi). \tag{14}$$

Note that, in the situation studied here, both terms of the variance in Equation (13) have the same order of magnitude, $O(N^2/n)$, since they are based on samples of the same size. Thus,

the $R\pi$-variance is the variance in the nonprobability sample of the estimator with inverse pseudo-inclusion probability weights plus a term reflecting the variance of the estimator of the **x**-totals in the probability sample.

Since $\hat{Y}_{np}(\pi) = \hat{Y}_{M1}$, Equation (13) also implies that the $R\pi$-variance of $\hat{Y}_{M2}$ with calibrated S_p weights is larger than that of the uncalibrated $\hat{Y}_{M1}$. This disagrees with the usual expectation that calibration on an effective predictor of Y reduces variances. To better understand this, note that if the matched x's in S_p and S_{np} were identical, then $\hat{\mathbf{X}}_p = \hat{\mathbf{X}}_{np}$ and the variable part in Equation (12) could be written as a weighted sum over S_{np} of residuals, which can then be used to show that $\hat{Y}_{M2}$ can have a smaller variance than $\hat{Y}_{M1}$. However, with imperfect matching the relationship in Equation (12) becomes more realistic.

As shown in Appendix (Subsection 8.2) the approximate $R\pi\xi$-variance when $R_j = \pi_j$ is

$$V_{R\pi\xi}\left(\hat{Y}_{M2}\right) \doteq \sum_{U} \frac{\sigma_j^2}{\pi_j} + \beta^T V_R\left(\hat{\mathbf{X}}_{np}\right)\beta + \beta^T V_\pi\left(\hat{\mathbf{X}}_p\right)\beta. \tag{15}$$

A natural estimator of Equation (15) is then

$$\upsilon_{R\pi\xi}\left(\hat{Y}_{M2}\right) = \sum_{S_{np}} \left(e_j/\pi_j\right)^2 + \hat{\mathbf{B}}_{np}(\pi)^T \upsilon_R\left(\hat{\mathbf{X}}_{np}\right)\hat{\mathbf{B}}_{np}(\pi) + \hat{\mathbf{B}}_{np}(\pi)^{\mathrm{T}} \upsilon_\pi\left(\hat{\mathbf{X}}_p\right)\hat{\mathbf{B}}_{np}(\pi). \tag{16}$$

Consequently, there are several options for variance estimation for $\hat{Y}_{M2}$ for cases (1) and (2). They can be summarized as:

- Case (1), $\tilde{w}_j = \pi_j^{-1}$
  - Estimate the ξ-variance in Equation (7)
  - Estimate the quasi-randomization ($R\pi$) variance with the with-replacement estimator in Equation (6);
  - Estimate the $R\pi\xi$-model variance with $v_{R\pi\xi}$ in Equation (10);
- Case (2), $\tilde{w}_j =$ GREG weight from S_p
  - Estimate the ξ-variance with Equation (11)
  - Estimate the $R\pi$-variance with Equation (14);
  - Estimate the $R\pi\xi$-variance with Equation (16);

4. Calibration Adjustment After Matching

The $R-$, $R\pi-$, or $R\pi\xi$-bias of the matched estimators, $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$, in Section 3 depend critically on whether $Pr\,(j \in S_{np}|\mathbf{x}_j) = Pr\,(i \in S_p)$ for matched units i and j. Matching on covariates attempts to ensure this; however, there is no guarantee that the condition is satisfied regardless of how extensive the set of covariates is.

Consequently, one might hope that calibrating the weights for the nonprobability sample will provide some bias protection. Suppose that the $\{\tilde{w}_j\}_{j\in S_{np}}$ weights are calibrated to the $\mathbf{X}_U$ population totals using the chi-square distance function associated with a GREG. Using the standard formula from Särndal et al. (1992, eq. 6.5.10), the resulting weight for unit j is

$$w_j^* = \tilde{w}_j\left[1 + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w})\right)^T\left[\mathbf{A}_{np}^*(\tilde{w})\right]^{-1}\mathbf{x}_j/\sigma_j^{*2}\right]$$
$$\equiv \tilde{w}_j g_j^*, \tag{17}$$

where $\hat{\mathbf{X}}_{np}(\tilde{w}) = \sum_{S_{np}} \tilde{w}_j\mathbf{x}_j$ and $\mathbf{A}_{np}^*(\tilde{w}) = \sum_{S_{np}} \tilde{w}_j\mathbf{x}_j\mathbf{x}_j^T/\sigma_j^{*2}$. (Note that σ_j^{*2} does not have to be the same as $\tilde{\sigma}_j^2$ used in constructing the GREG weight in S_p.) As in Section 3, σ_j^{*2} is often set to a constant in which case it drops out of the formula for w_j^*. The matched, calibrated estimator is then

$$\hat{Y}_{MC} = \sum_{S_{np}} w_j^* y_j$$
$$= \sum_{S_{np}} \tilde{w}_j y_j + \left(\mathbf{X}_U - \hat{\mathbf{X}}_M\right)^T\left[\mathbf{A}_{np}^*(\tilde{w})\right]^{-1}\sum_{S_{np}} \tilde{w}_j\mathbf{x}_j y_j/\sigma_j^{*2}$$
$$= \hat{Y}_M + \left(\mathbf{X}_U - \hat{\mathbf{X}}_M\right)^T\hat{\mathbf{B}}_{np}^*(\tilde{w}), \tag{18}$$

where $\hat{\mathbf{B}}_{np}^*(\tilde{w}) = \left[\mathbf{A}_{np}^*(\tilde{w})\right]^{-1}\sum_{S_{np}} \tilde{w}_j\mathbf{x}_j y_j/\sigma_j^{*2}$. As in Section 3, calculations depend on cases (1) and (2) of the $\tilde{w}_j$ weights. When case (1) weights are used from S_p, the calibrated estimator will be denoted by $\hat{Y}_{MC1}$; when case (2) weights are used, $\hat{Y}_{MC2}$ denotes the calibrated estimator in subsequent sections.

4.1. Biases in Case (1)

When $\tilde{w}_j = \pi_j^{-1}$, $\hat{\mathbf{X}}_M = \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)$ and, after calibration, the estimator of the total can be written as

$$\hat{Y}_{MC1} = \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)^T\hat{\mathbf{B}}_{np}^*(\pi),$$

where $\hat{\mathbf{B}}_{np}^*(\pi)$ is the special case of $\hat{\mathbf{B}}_{np}^*(\tilde{w})$ with $\tilde{w}_j = \pi_j^{-1}$. Since $E_\xi\left(\hat{Y}_{np}(\pi)\right) = \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\beta$ under model (1) and $E_\xi\left(\hat{\mathbf{B}}_{np}^*(\pi)\right) = \beta$, $E_\xi\left(\hat{Y}_{MC1} - Y_U\right) = 0$, that is, $\hat{Y}_{MC1}$ is ξ-unbiased. Thus, calibrating on the x's in the ξ-model yields an ξ-unbiased estimator even if $R_j \neq \pi_j$.

To calculate the $R\pi$-expectation, define $\mathbf{B}_U^* = \mathbf{A}_U^{*-1}\left(\sum_U \frac{R_j}{\pi_j}\mathbf{x}_j y_j/\sigma_j^{*2}\right)$ with $\mathbf{A}_U^* = \sum_U \frac{R_j}{\pi_j}\mathbf{x}_j\mathbf{x}_j^T/\sigma_j^{*2}$. By the same type of Taylor series argument as in Särndal et al. (1992, sec. 6.5),

$$\hat{Y}_{MC1} \doteq \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)^T\mathbf{B}_U^*. \tag{19}$$

It follows that $E_R E_\pi\left(\hat{Y}_{MC1}\right) = E_R\left(\hat{Y}_{MC1}\right) \doteq \sum_U R_j y_j/\pi_j + \left(\mathbf{X}_U - \sum_U R_j\mathbf{x}_j/\pi_j\right)^T\mathbf{B}_U^*$. If $R_j = \pi_j$, then $\hat{Y}_{MC1}$ is approximately $R\pi$-unbiased. Another consequence is that, when S_{np} is calibrated with the x's in model if (1) and S_p has case (1) weights, $\hat{Y}_{MC1}$ is $R\pi\xi$-unbiased if $R_j = \pi_j$.

4.2. Biases in Case (2)

In case (2) with $\tilde{w}_j = g_j/\pi_j$ and g_j defined in Equation (3), the matched estimator after calibration equals

$$\hat{Y}_{MC2} = \sum_{S_{np}} g_j^* g_j y_j / \pi_j,$$

where

$$g_j^* = 1 + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\tilde{w})\right)^T \left[\tilde{\mathbf{A}}_{np}^*(\tilde{w})\right]^{-1} \mathbf{x}_j / \sigma_j^{*2}.$$

As show in Appendix (Subsection 8.3), the calibrated estimator of the total is approximately

$$\hat{Y}_{MC2} \doteq \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{B}}_U + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\tilde{w})\right)^T \mathrm{B}_U^*. \tag{20}$$

Using Equation (19), the ξ-expectation is

$$E_\xi\left(\hat{Y}_{MC2}\right) \doteq \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w})\right)^T \beta,$$

which is not $E_\xi(Y_U) = \mathbf{X}_U^T \beta$. That is, $\hat{Y}_{MC2}$ is ξ-biased. This bias occurs even though the nonprobability sample is calibrated on the x's in the model for Y.

If $R_j = \pi_j$, then $E_R E_\pi\left(\hat{Y}_{MC2}\right) \doteq Y_U$ and $E_R E_\pi E_\xi\left(\hat{Y}_{MC2} - Y_U\right)$ is approximately zero.

The bias results for the matched, calibrated estimators $\hat{Y}_{MC1}$ and $\hat{Y}_{MC2}$ can be summarized as follows:

- Case (1) $\tilde{w}_j = \pi_j^{-1}$ and the $\tilde{w}_j$ are then calibrated to population x-totals
  - When the linear model (1) holds, $\hat{Y}_{MC1}$ is ξ-unbiased regardless of whether $R_j = \pi_j$;
  - $\hat{Y}_{MC1}$ is approximately R-, $R\pi$-, and $R\pi\xi$-unbiased in large samples if $R_j = \pi_j$;
- Case (2), $\tilde{w}_j$ is the GREG weight for its matched unit in S_p and the $\tilde{w}_j$ are then calibrated to population x-totals
  - $\hat{Y}_{MC2}$ is ξ-biased even if Equation (1) holds and the nonprobability sample S_{np} is calibrated on the x's in the model;
  - $\hat{Y}_{MC2}$ is approximately $R\pi$-unbiased in large samples if $R_j = \pi_j$;
  - $\hat{Y}_{MC2}$ is approximately $R\pi\xi$-unbiased in large samples if $R_j = \pi_j$;

If case (1) holds where the weights assigned to matched units are inverses of selection probabilities from S_p, the situation is more straightforward than case (2). R-unbiasedness in case (1) requires that the pseudo-inclusion probabilities can be taken from the probability sample, that is, $R_j = \pi_j$. Nonetheless, in case (1) calibrating the nonprobability sample does produce an ξ-unbiased estimator even if $R_j \neq \pi_j$, as one would hope. However, in case (2) when the weights from the probability sample are calibrated and the nonprobability sample is further calibrated on the same x's, the resulting estimator is not ξ-unbiased.

4.3. Variance of the Matched, Calibrated Estimator in Case (1)

To compute the ξ-model variance, note that the estimator of total can also be written as $\hat{Y}_{MC1} = \sum_{S_{np}} g_j^* y_j / \pi_j$ with g_j^* defined in Equation (17) with $\tilde{w}_j = 1/\pi_j$. The ξ-variance is then

$$V_\xi\left(\hat{Y}_{MC1}\right) = \sum_{S_{np}} \left(\frac{g_j^*}{\pi_j}\right)^2 \sigma_j^2.$$

It follows that the $R\pi\xi$-variance is $V_{R\pi\xi}(\hat{Y}_{MC1}) = \sum_U \left(g_j^{*2}/\pi_j\right)\sigma_j^2$. The ξ-variance can be estimated with

$$v_\xi(\hat{Y}_{MC1}) = \sum_{S_{np}} \left(\frac{g_j^*}{\pi_j}\right)^2 \hat{e}_j^{*2}, \tag{21}$$

where $\hat{e}_j^{*2} = y_j - \mathbf{x}_j^T \hat{\mathbf{B}}_{np}^*(\pi)$ with $\hat{\mathbf{B}}_{np}^*(\pi) = \left(\sum_{S_{np}} \mathbf{x}_j \mathbf{x}_j^T / \left(\pi_j \sigma_j^{*2}\right)\right)^{-1} \sum_{S_{np}} \mathbf{x}_j y_j / \left(\pi_j \sigma_j^{*2}\right)$.

To compute the $R-$ and $R\pi$-variance, we use the approximation in Equation (19). Assume that $R_j = \pi_j$ so that $\hat{Y}_{MC1}$ is R-unbiased. Based on results in Subsection 4.1, the estimator can be approximated as

$$\begin{aligned}
\hat{Y}_{MC1} &= \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)^T \hat{\mathbf{B}}_{np}^*(\pi) \\
&\doteq \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)^T \mathbf{B}_U^* \\
&= \sum_{S_{np}} \pi_j^{-1} e_j^* + \mathbf{X}_U^T \mathbf{B}_U^*,
\end{aligned} \tag{22}$$

where $e_j^* = y_j - \mathbf{x}_j^T \mathbf{B}_U^*$. The R- (and $R\pi$-) variance is, thus, equal to the variance of the first term in the last line of Equation (22). If the sample S_{np} is treated as being selected with replacement, then a variance estimator is

$$v_{R\pi}(\hat{Y}_{MC1}) = \frac{n}{n-1} \sum_{j \in S_{np}} \left(\tilde{w}_j \hat{e}_j^* - \frac{1}{n} \sum_{j' \in S_{np}} \tilde{w}_{j'} \hat{e}_{j'}* \right)^2. \tag{23}$$

4.4. Variance of the Matched, Calibrated Estimator in Case (2)

As shown in Appendix (Subsection 8.4) approximation (20) can be rewritten as

$$\hat{Y}_{MC2} \doteq \sum_{S_{np}} y_j \left(\frac{1}{\pi_j} + F_j\right) + \sum_{U - S_{np}} y_j F_j,$$

where F_j is a term that is $O_p\left(n^{-1/2}\right)$. As a result, $V_\xi(\hat{Y}_{MC2}) \doteq \sum_{S_{np}} \left(\sigma_j/\pi_j\right)^2$, which can be estimated with

$$v_\xi(\hat{Y}_{MC2}) = \sum_{S_{np}} \left(\frac{\hat{e}_j^*}{\pi_j}\right)^2. \tag{24}$$

Rewriting Equation (20), the calibrated estimator of the total is also

$$\hat{Y}_{MC2} \doteq \sum_{S_{np}} \frac{e_j^*}{\pi_j} + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{B}}_U + \mathbf{X}_U^T \mathbf{B}_U^*, \tag{25}$$

where e_j^* was defined above. Using the total variance formula, the $R\pi$-variance can be derived as

$$V_{R\pi}(\hat{Y}_{MC2}) = V_R E_\pi(\hat{Y}_{MC2} | S_{np}) + E_R V_\pi(\hat{Y}_{MC2} | S_{np})$$

$$= V_R\left(\sum_{S_{np}} \frac{e_j^*}{\pi_j}\right) + E_R V_\pi\left[\left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{B}}_U\right]$$

$$= V_R\left(\sum_{S_{np}} \frac{e_j^*}{\pi_j}\right) + \tilde{\mathbf{B}}_U^T V_\pi\left(\hat{\mathbf{X}}_p\right)\tilde{\mathbf{B}}_U.$$

An estimator of this variance is

$$\upsilon_{R\pi}\left(\hat{Y}_{MC2}\right) = \upsilon_R\left(\sum_{S_{np}} \frac{e_j^*}{\pi_j}\right) + \hat{\tilde{\mathbf{B}}}_{np}(\pi)^T \upsilon_\pi\left(\hat{\mathbf{X}}_p\right)\hat{\tilde{\mathbf{B}}}_{np}(\pi) \tag{26}$$

with $\upsilon_R\left(\sum_{S_{np}} \frac{e_j^*}{\pi_j}\right)$ being a variance estimator of an estimated total appropriate to how the nonprobability sample is handled. We use $\hat{\tilde{\mathbf{B}}}_{np}(\pi)$ in Equation (26) rather than an estimator with $\tilde{w}$ weights since the former is expected to be somewhat more stable. If S_{np} is treated as being with-replacement, the first component in Equation (26) can be computed with Equation (23).

Details of calculating $V_{R\pi\xi}\left(\hat{Y}_{MC2}\right)$ are in Appendix (Subsection 8.4). This variance can be estimated with

$$\upsilon_{R\pi\xi}\left(\hat{Y}_{MC2}\right) = \sum_{S_{np}}\left(\frac{\hat{e}_j^*}{\pi_j}\right)^2 + \hat{\tilde{\mathbf{B}}}_{np}(\pi)^T \upsilon_\pi\left(\hat{\mathbf{X}}_p\right)\hat{\tilde{\mathbf{B}}}_{np}(\pi). \tag{27}$$

For each of the variance estimators above for the matched, calibrated estimator in cases (1) and (2), it is important to remember that unless $R_j = \pi_j$ the estimator of total itself will be biased. If so, the mean square error will have a bias-squared component that none of the variance estimators will reflect.

In the combination above, both the weights in S_p and those in S_{np} are calibrated to a given set of x's. This is similar to the situation studied by Rao et al. (2002, 368), who noted that in a regression with calibration weights, GREG residuals are based on the regression of model residuals on $\mathbf{X}$. If the model fits well, there will be very little association between those residuals and $\mathbf{X}$ leading to no gain compared to an estimator not using calibration weights. In our situation, when the estimators of totals are unbiased, we can expect $\hat{Y}_{M2}$ with calibration in S_p, $\hat{Y}_{MC1}$ with no calibration in S_p and calibration in S_{np}, and $\hat{Y}_{MC2}$ with calibration in both S_p and S_{np} to be about equally precise—a point borne out by the simulation in Section 5.

5. Simulation Studies

To study the performance of the proposed estimators described above, we performed two simulation studies with an artificial population. In the first, conditions are created where close matches can be found between units in the probability sample and the nonprobability sample. In the second simulation, close matches are much less likely.

5.1. Simulation Study I

In the simulation, a finite population of size $N = 100{,}000$ was based on the following model:

$$E_\xi(Y) = \alpha + \beta X, V_\xi(Y) = \sigma^2 X^{3/2},$$

where $\alpha = 0.4$, $\beta = 0.25$, $\sigma^2 = 0.0625$, and X follows a gamma distribution with density function $f(x) = 0.04x \exp(-x/5)$. This is the same model as used by Hansen et al. (1983); the function `HMT` in the `R PracTools` package (Valliant et al. 2020) was used to generate the population. Conditional on X, Y follows a gamma distribution with density function $g(y;\ x) = (1/b^c\Gamma(c))y^{c-1} \exp(-y/b)$, where $b = 1.25x^{3/2}(8 + 5x)^{-1}$, $c = 0.04x^{-3/2}(8 + 5x)^2$ and $\Gamma(\cdot)$ is the gamma function. The finite population is stratified into five strata by ranges of the variable X, such that each stratum has approximately the same total of X. A stratified, probability sample S_p of size $n = 250$ is then selected from the population using stratified, simple random sampling (*stsrs*) without replacement, in which the sample stratum sizes are given by (50, 50, 50, 50, 50). Further, a stratified, volunteer panel of size $M = 1{,}250$ is selected from the population with stratum sample sizes (250, 250, 250, 250, 250) using stratified, simple random sampling. Although the volunteer panel is a probability sample, their weights are treated as unknown for the simulation. Note that the sampling fractions of both S_p and S_{np} are small and, thus, will not affect the empirical variances of estimates.

For each unit of the probability sample of $n = 250$, we find the closest matching unit of the volunteer panel to obtain the matched, non-probability sample S_{np} of size $n = 250$, using single nearest neighbor matching without replacement based on the single auxiliary variable X. That is, it matches each unit i in the probability sample with a unit j in the volunteer panel with the closest distance $d(i, j) = \|X_i - X_j\|$ and the unit j in the volunteer panel can be used only once, where $\|\cdot\|$ is a norm, such as 1-norm, 2-norm and ∞-norm. The units in the volunteer panel are then assigned the weight of their nearest neighbor match from the probability sample using the `R` package `Matching` (Sekhon 2011). In this example, finding close matches is fairly easy, and we should have $R_j = \pi_j$, $j \in S_{np}$, in almost all cases because both S_p and S_{np} are *stsrs*. The parameter of interest is the population total of Y. Finally, the matched estimator and the matched, calibrated estimator under cases (1) and (2) in Section 3 are computed, denoted by

- $\hat{Y}_{M1}$, estimator (2) with $1/\pi$ weights from the matched units in S_p,
- $\hat{Y}_{M2}$, estimator (4) with GREG weights from the matched units in S_p,
- $\hat{Y}_{MC1}$, estimator (18) with $1/\pi$ weights from the matched units in S_p followed by calibration in S_{np}, and
- $\hat{Y}_{MC2}$, estimator (18) with GREG weights from the matched units in S_p followed by calibration in S_{np}.

The above process is repeated 5,000 times. The percentage relative biases (relbiases), the variances and the mean squared errors of the matched estimator and the matched, calibrated estimator under cases (1) and (2), are presented in Table 1. The empirical percent relative bias is defined as $100 \times bias(\hat{Y})/Y$.

Table 1. Simulation Study I: Monte Carlo percent relative biases, variances and mean squared errors of the point estimators

Estimators	Relative bias (%)	Variance ($\div 10^7$)	MSE ($\div 10^7$)	Ratio to min MSE
$\hat{Y}_{M1}$	-0.0318	9.02	9.02	1.17
$\hat{Y}_{M2}$	-0.0078	7.68	7.68	1.00
$\hat{Y}_{MC1}$	-0.0274	7.68	7.68	1.00
$\hat{Y}_{MC2}$	-0.0274	7.68	7.68	1.00
$\hat{Y}_{DR}$	-0.1062	9.23	9.24	1.20

For comparison we included a doubly robust estimator, denoted by $\hat{Y}_{DR}$, that was computed without matching. This estimator was computed in two steps as described in Elliott and Valliant (2017). First, an equal probability subsample of $n = 250$ was selected from the volunteer panel of $m = 1{,}250$. Then, S_p and S_{np} are combined. Units in S_{np} are given a weight of 1 while units in S_p were assigned their sampling weight of $1/\pi_i$. A logistic regression with X as the covariate was run to predict the probability of being in S_{np}. The weight for unit j in S_{np} was then calculated as $w_j = \left(1 - \hat{R}_j\right)/\hat{R}_j$ where $\hat{R}_j$ is the predicted probability of being in S_{np} (see Wang et al. 2021). Without the odds transformation, the estimator would be somewhat biased (Chen et al. 2020), but in this case the bias was negligible since S_{np} is a small fraction of the population (Wang et al. 2021). Finally, the estimator was calibrated with a model having an intercept and X. The general formulas for $\hat{Y}_{DR}$ and its variance estimator are

$$\hat{Y}_{DR} = \sum_{S_{np}} g_{DRj} w_j y_j; \quad v_{DR} = \sum_{S_{np}} (g_{DRj} w_j)^2 e_{DRj}^2 \tag{28}$$

where w_j is the estimated weight defined above, $e_{DRj} = y_j - \mathbf{x}_j^T \hat{\mathbf{B}}_{DR}$ with $\hat{\mathbf{B}}_{DR}$ being the weighted least squares estimator of slope with weights equal to w_j, and g_{DRj} is the g-weight defined using the w_j weights. The estimator $\hat{Y}_{DR}$ will be approximately unbiased if either the model for the propensities of being in S_{np} or the model for Y is correct. Two other things to note are that (a) $\hat{Y}_{DR}$ is not the same as the DR estimator considered by Chen et al. (2020) and (b) v_{DR} is theoretically incorrect because it treats the weights w_j as constants rather than estimates. Although Chen et al. (2020) present a variance estimator for their DR estimator, it is not appropriate here since our $\hat{Y}_{DR}$ differs from theirs.

Simulation results in Table 1 show that the absolute relative biases of the matched estimators under the two cases of weights from S_p are small and close to those of the corresponding matched, calibrated estimators under the two cases. The t-statistics, computed across the 5,000 samples, for testing whether the biases of the estimated totals are zero are 0.68 in absolute value or less with the exception of $\hat{Y}_{DR}$ which is -2.26. Since the relbias of the DR estimator is only -0.1062%, a t-statistic of this size is of no practical importance. Thus, both the matched estimators and the matched, calibrated estimators are unbiased when $R_j \cong \pi_j$, $j \in S_{np}$ in both cases (1) and (2) as predicted by the theory in Subsections 3.1 and 3.2. The variances and MSEs of $\hat{Y}_{M2}$, $\hat{Y}_{MC1}$, and $\hat{Y}_{MC2}$ are all equal as anticipated in the comment at the end of Subsection 4.4 and are about 17% smaller than those of $\hat{Y}_{M1}$. Consequently, while all estimates are approximately unbiased, the calibration adjustment after matching produces more efficient estimators compared to only

matching under case (1). Also noteworthy is the fact that the doubly robust estimator, $\hat{Y}_{DR}$, has a 20% larger MSE than the best of the matching estimates. This is a consequence of the logistic model used to estimate $Pr\ (j \in S_{np})$ being a misspecification.

In addition to the point estimators, the variance estimators of the matched estimator, the matched, calibrated estimator under cases (1) and (2), and the doubly robust estimator are also computed according to Equations (6), (7), (10), (11), (14), (16), (21), (23), (24), (26), (27), and (28). In all cases S_{np} is treated as an unstratified, with replacement sample. Percent relative biases (RB) are computed for the variance estimators with respect to the empirical variances (Empvar) and MSEs of the point estimators across the 5,000 simulations:

$$RB.Empvar = \frac{100 \times \left(\sum_{b=1}^{B} v^{(b)}(\hat{Y})/B - V(\hat{Y})\right)}{V(\hat{Y})},$$

$$RB.MSE = \frac{100 \times \left(\sum_{b=1}^{B} v^{(b)}(\hat{Y})/B - MSE(\hat{Y})\right)}{MSE(\hat{Y})},$$

where $V(\hat{Y})$ is the empirical or monte carlo variance of a point estimator $\hat{Y}$, $MSE(\hat{Y})$ is MSE of the point estimator $\hat{Y}$, $v^{(b)}(\hat{Y})$ is a variance estimator of $\hat{Y}$ computed from the b^{th} simulated sample, and $B = 5{,}000$ is the total number of simulation runs. The percent relative biases (RB) and 95% confidence interval (CI) coverages using the normal approximation and the different variance estimates, are presented in Table 2.

With three exceptions, the relbiases in Table 2 are small, ranging from -1.1% to 3.1%. An exception is $v_{\xi}(\hat{Y}_{M1})$ which is a 15.7% underestimate due to the fact that it does not account for the variability of $\hat{X}_{np}$ as shown in Subsection 3.3. The $R\pi$ and $R\pi\xi$ estimators for $\hat{Y}_{M2}$ are about 24% overestimates. As explained in Appendix (Subsection 8.2), these estimators will not fully account for precision gains due to calibration of weights in S_p when the x-matches are extremely close. Confidence interval coverage ranges from 94.8% to 97.0% except for $v_{\xi}(\hat{Y}_{M1})$ which covers in 92.7% of samples due to its underestimation.

Table 2. Simulation Study I: Percent relative biases and 95% confidence interval coverages of the variance estimators

Estimators	RB.Empvar (%)	RB.MSE (%)	CI coverage (%)
$v_{\xi}(\hat{Y}_{M1})$	-15.7	-15.7	92.7
$v_{R}(\hat{Y}_{M1})$	3.1	3.1	95.0
$v_{R\pi\xi}(\hat{Y}_{M1})$	2.8	2.8	96.8
$v_{\xi}(\hat{Y}_{M2})$	-0.9	-0.9	94.8
$v_{R\pi}(\hat{Y}_{M2})$	23.9	23.9	96.8
$v_{R\pi\xi}(\hat{Y}_{M2})$	23.6	23.6	97.0
$v_{\xi}(\hat{Y}_{MC1})$	-1.0	-1.0	94.9
$v_{R}(\hat{Y}_{MC1})$	-0.7	-0.7	94.8
$v_{\xi}(\hat{Y}_{MC2})$	-1.1	-1.1	94.8
$v_{R\pi}(\hat{Y}_{MC2})$	2.2	2.2	95.2
$v_{R\pi\xi}(\hat{Y}_{MC2})$	1.8	1.9	95.2
$v(\hat{Y}_{DR})$	0.4	0.4	94.9

5.2. Simulation Study II

In this simulation, we consider a case in which $R_j \neq \pi_j, j \in S_{np}$. The same finite population of size $N = 100{,}000$ is used as in simulation study I along with a stratified, probability sample S_p of size $n = 250$. A volunteer panel of expected size $m = 1{,}250$ is selected from the population using Poisson sampling with selection probabilities $\pi_i{}'$ defined as follows:

$$\pi_i = 0.085 \exp(-0.085X),$$
$$\pi_i{}' = \frac{m\pi_i}{\sum_{i=1}^{N} \pi_i}.$$

With this definition of π_i, the probability of being in S_{np} decreases with increasing X. This kind of selection for the volunteer panel will generally result in $R_j \neq \pi_i$, for a unit $j \in S_{np}$ matched to a unit $i \in S_p$.

As in simulation I, single nearest neighbor matching without replacement based on the variable X is adopted to conduct matching for the probability sample. The matched estimator, the matched, calibrated estimator, the doubly robust estimator and their variance estimators under cases (1) and (2) are computed. The above procedure is repeated 5,000 times. The relative biases, the variances and the mean squared errors are listed in Table 3. Also, the same relative biases and 95% CI coverages of variance estimators as those in simulation study I are displayed in Table 4.

In Table 3 the matched estimators, $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$, have biases of about -5%. These biases are largely corrected by calibrating with $\hat{Y}_{MC1}$ and $\hat{Y}_{MC2}$. The t-statistics for testing whether the biases are zero are over 100 in absolute value for $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$ but are only about -3 for the calibrated estimators. Despite the somewhat large t-statistics for the calibrated estimators, calibration is very effective in reducing bias due to using incorrect S_{np} inclusion probabilities. The calibrated estimates, consequently, have substantially smaller MSEs than $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$ because of their reduced bias. The doubly robust estimator, $\hat{Y}_{DR}$, is also approximately unbiased; however, its variance and MSE are 50% higher than those of $\hat{Y}_{MC1}$ and $\hat{Y}_{MC2}$.

In Table 4 the variance estimates for $\hat{Y}_{M1}$ and $\hat{Y}_{M2}$ are biased estimates of the empirical variance and severe underestimates of the MSEs. This leads to CIs that cover only about 56% to 66% of the time for the first four variance estimates in Table 4. Since $v_{R\pi}(\hat{Y}_{M2})$ and $v_{R\pi\xi}(\hat{Y}_{M2})$ overestimate the empirical variances by about 22%, their CIs do cover the population totals in 96.8% and 96.9% of samples. The fact that calibrating removes the bias of the matching estimators plus the low biases of the variance estimators for $\hat{Y}_{MC1}$ and

Table 3. Simulation Study II: Percent relative biases, variances and mean squared errors of the point estimators

Estimators	Relative bias (%)	Variance ($\div 10^7$)	MSE ($\div 10^7$)	Ratio to min MSE
$\hat{Y}_{M1}$	-5.2	8.83	31.04	3.9
$\hat{Y}_{M2}$	-5.1	7.99	29.86	3.8
$\hat{Y}_{MC1}$	-1.3	7.66	7.67	1.0
$\hat{Y}_{MC2}$	-1.3	7.66	7.67	1.0
$\hat{Y}_{DR}$	-0.2	11.31	11.35	1.5

Table 4. Simulation Study II: Percent relative biases and 95% confidence interval coverages of the variance estimators

Estimators	RB.Empvar (%)	RB.MSE (%)	CI coverage (%)
$v_{\xi}(\hat{Y}_{M1})$	-18.2	-76.7	56.2
$v_{R}(\hat{Y}_{M1})$	8.7	-69.1	64.9
$v_{R\pi\xi}(\hat{Y}_{M1})$	8.4	-69.2	66.1
$v_{\xi}(\hat{Y}_{M2})$	-9.5	-75.8	57.0
$v_{R\pi}(\hat{Y}_{M2})$	22.9	-67.1	96.8
$v_{R\pi\xi}(\hat{Y}_{M2})$	22.5	-67.2	96.9
$v_{\xi}(\hat{Y}_{MC1})$	1.6	1.4	95.1
$v_{R}(\hat{Y}_{MC1})$	-5.3	-5.4	94.2
$v_{\xi}(\hat{Y}_{MC2})$	-5.6	-5.8	94.5
$v_{R\pi}(\hat{Y}_{MC2})$	-2.4	-2.5	94.5
$v_{R\pi\xi}(\hat{Y}_{MC2})$	-2.7	-2.9	94.5
$v(\hat{Y}_{DR})$	-3.1	-3.4	94.3

$\hat{Y}_{MC2}$ leads to CI coverage of 94.2% to 95.1%. The variance estimator for $\hat{Y}_{DR}$ has CI coverage of 94.3% and performs well. These coverage probabilities are estimated reasonably precisely in the simulation of 5,000 samples. The half-widths of normal approximation 95% CIs on the empirical coverage rates themselves range from 0.62 percentage points when coverage is 95% to 1.39 percentage points when coverage is 60%. (These half-width lengths also apply to the results for Simulation I and for the real data simulation presented in Section 6.)

6. Illustration with Real Population

To further assess the performance of the matching estimators, they are applied to data obtained from the 2015 US Behavioral Risk Factor Surveillance Survey (BRFSS) (Center for Disease Control and Prevention 2023), which is a sample from the US population 18 years and older. The file contains information about whether persons used the internet in the past 30 days (`INTERNET`). The BRFSS is part of a national state-by-state system of surveys used to monitor health conditions in the United States. Data are collected through telephone household interviews. The analytic variables Y in this study are whether respondents were ever diagnosed with a heart attack (`CVDINFR4`), were ever told by a medical professional that they have diabetes (`DIABETE3`), and were ever told they had a stroke (CVDSTRK3). Although each of these analysis variables is binary, use of linear estimators, as studied in previous sections, is standard survey practice, largely because of their convenience for data analysts.

Covariates associated with Y are sex, age, race, marital status, physical weight, employment status, education level, income level, whether respondents smoked at least 100 cigarettes in their entire life, and whether respondents participated in any physical activities or exercises in the past 30 days in 2015. All of the variables are shown in Table 5.

After deleting cases with either a missing, a don't know or a refused response to any of these variables, 315,669 persons are available for this study. Two weights are provided with the data set: `X_WT2RAKE`, which is a design weight and `X_LLCPWT`, which is a raked, final weight. According to the documentation Center for Disease Control and

Table 5. Covariates used in the BRFSS simulation study

Variables	Type	Description
SEX	2 categories	Respondents sex: 1=Male; 2=Female
X_AGE	6 categories	Imputed age in six groups: 1=Age 18 to 24; 2=Age 25 to 34; 3=Age 35 to 44; 4=Age 45 to 54; 5=Age 55 to 64; 6=Age 65 or older
X_RACE	8 categories	Computed race-ethnicity grouping: 1=White only, non-Hispanic; 2=Black only, non-Hispanic; 3=American Indian or Alaskan Native only, Non-Hispanic; 4=Asian only, non-Hispanic; 5=Native Hawaiian or other Pacific Islander only, Non-Hispanic; 6=Other race only, non-Hispanic; 7=Multiracial, non-Hispanic; 8=Hispanic
MARITAL	6 categories	Marital status: 1=Married; 2=Divorced; 3=Widowed; 4=Separated; 5=Never married; 6=A member of an unmarried couple
WEIGHT2	Continuous	Reported weight in pounds: 50–999
EMPLOY1	8 categories	Employment status: 1=Employed for wages; 2=Self-employed; 3=Out of work for 1 year or more; 4=Out of work for less than 1 year; 5=A homemaker; 6=A student; 7=Retired; 8=Unable to work
EDUCA	6 categories	Education level: 1=Never attended school or only kindergarten; 2=Grades 1 through 8 (Elementary); 3=Grades 9 through 11 (Some high school); 4=Grade 12 or GED (High school graduate); 5=College 1 year to 3 years (Some college or technical school); 6=College 4 years or more (College graduate)
INCOME2	8 categories	Income level: 1=Less than USD 10,000; 2=USD 10,000 to less than USD 15,000; 3=USD 15,000 to less than USD 20,000; 4=USD 20,000 to less than USD 25,000; 5=USD 25,000 to less than USD 35,000; 6=USD 35,000 to less than USD 50,000; 7=USD 50,000 to less than USD 75,000; 8=USD 75,000 or more
SMOKE100	2 categories	Smoked at least 100 cigarettes?: 1=Yes; 2=No
EXERANY2	2 categories	Exercise in past 30 days?: 1=Yes; 2=No
INTERNET	2 categories	Internet use in the past 30 days?: 1=Yes; 2=No
CVDINFR4	2 categories	Ever diagnosed with heart attack?: 1=Yes; 2=No
DIABETE3	2 categories	Ever told you have diabetes?: 1=Yes; 2=No
CVDSTRK3	2 categories	Ever told you had a stroke?: 1=Yes; 2=No

Prevention (2017). BRFSS rakes the design weight to eight margins (gender by age group, race/ethnicity, education, marital status, tenure, gender by race/ethnicity, age group by race/ethnicity, and phone ownership). The raking also serves as a noncoverage/nonresponse adjustment. Because of the asymptotic equivalence of the GREG and raked estimators shown by Deville and Särndal (1992), the earlier theory in Sections 3 and 4 should apply to estimators based on `X_LLCPWT`.

In this data set of 315,669 persons, 256,949 people who had used the internet in the past 30 days are considered as the web (nonprobability) subset. Using the `X_LLCPWT` weights, the web population is only 84% (81% unweighted) of the target population, indicating that the effect of coverage error could be substantial. Moreover, the weighted distributions of the

categorical covariates among all respondents in the web, non-web, and full populations are given in Table 6. Categories of some variables are combined in Table 6 and in the simulation compared to the categories in Table 5 because they are small. Table 7 gives the proportions that reported a heart attack, diabetes, or a stroke in the web, non-web, and full populations.

As shown in Tables 6 and 7, there are differences between the target population and the web and non-web populations in the estimated distributions of some of the covariates. For example, 0.19 of the full population are age 65 or older, 0.14 of the web population

Table 6. Distributions of the categorical variables and means of the continuous variable, body weight, in the web, non-web, and full populations

	Variables	Web Population	Non-web Population	Target Population
SEX	Male	0.50	0.49	0.50
	Female	0.50	0.51	0.50
X_AGE	Age 18 to 24	0.13	0.02	0.11
	Age 25 to 34	0.20	0.05	0.18
	Age 35 to 44	0.19	0.09	0.17
	Age 45 to 54	0.19	0.16	0.18
	Age 55 to 64	0.16	0.22	0.17
	Age 65 or older	0.14	0.45	0.19
X_RACE	Non-black, non-Hispanic	0.90	0.84	0.89
	Black only, non-Hispanic	0.10	0.16	0.11
MARITAL	Married or member of an unmarried couple	0.59	0.47	0.58
	Divorced	0.11	0.15	0.11
	Widowed, separated, never married	0.30	0.38	0.31
EMPLOY1	Employed for wages, self-employed	0.65	0.30	0.59
	Out of work	0.05	0.06	0.05
	Other (homemaker, student, retired, unable to work)	0.30	0.63	0.35
EDUCA	Grade 11 or less	0.08	0.39	0.13
	Grade 12 or equivalent	0.25	0.37	0.27
	College 1 year to 3 years	0.35	0.18	0.32
	College 4 years or more	0.33	0.06	0.29
INCOME2	Less than USD 25,000	0.21	0.60	0.27
	USD 25,000 to less than USD 50,000	0.24	0.27	0.24
	USD 50,000 to less than USD 75,000	0.17	0.07	0.16
	USD 75,000 or more	0.38	0.06	0.33
SMOKE100	Smoked at least 100 cigarettes	0.59	0.51	0.58
	Not smoked at least 100 cigarettes	0.41	0.49	0.42
EXERANY2	Exercise in past 30 days	0.22	0.40	0.25
	No exercise in past 30 days	0.78	0.60	0.75
WEIGHT2	Body weight in pounds	180.50	176.90	180.00

Table 7. Proportions of the web, non-web, and total populations that have been told by a medical professional that they have three health conditions

Condition	Web	Non-web	Total pop
Heart attack (CVDINFR4)	0.031	0.107	0.043
Diabetes (DIABETE3)	0.093	0.233	0.116
Stroke (CVDSTRK3)	0.020	0.076	0.029

are, and 0.45 of the non-web are 65+. In the target population, 0.59 are employed for wages, 0.65 are in the web population, but only 0.30 of the non-web are. About 8% of the web population have a grade 11 education or less while 13% of the full population does; 33% of the web population attended four or more years of college while 29% of the full population did. For the analysis variables in Table 7, 4.3% of the target population have ever been diagnosed with a heart attack while 3.1% of the web population and 10.7% of the non-web population have. Similar differences occur for diabetes and stroke. Although the percentage point differences are small between the web and full populations, the relative differences are substantial. For example, heart attacks in the web population are 72% (0.031/0.043) of those in the full population; diabetes in the web population is 80% of the full population rate; strokes in the web population are 72% of those in the full population. Consequently, calibrating the matched sample may reduce bias and variance as long as the covariates in Table 6 are predictive of the Y's. However, it is clear that weighting a sample from the web population will have to achieve a considerable amount of bias correction in order to produce good estimates for the full, target population.

Also noteworthy are the substantial differences between the web and non-web subpopulations. The non-web people are older, more likely to be Black and non-Hispanic, more likely to not be in the labor force, less educated, lower income, and more likely to have smoked than the web persons. The non-web people are also much more likely to have had heart attacks, diabetes, and strokes. Our focus is on using a sample from the web population to make estimates for the full population, but any attempt to use a sample from the web population to represent the non-web population seems doomed to failure. In general, a nonprobability sample that has serious coverage problems cannot be expected to produce good estimates for poorly covered domains.

To apply the proposed matching method, simple random samples are selected from the BRFSS web subsample and from the BRFSS full sample. Using equal probability sampling preserves any differences between the web and full samples and, in particular, any coverage defects in the web sample. The size of the S_p probability sample was $n = 500$ while the size of the initial S_{np} web sample was $M = 3{,}000$. The BRFSS raked weights for persons in S_p were adjusted to equal $\tilde{w}_j = (N/n) * \texttt{X\_LLCPWT}$ where $N = 315{,}669$. Since the BRFSS design weights did not include a nonresponse adjustment and, consequently, did not sum to an estimate of the size of the target population, we computed a nonresponse-adjusted design weight for each person in S_p as $\tilde{w}_{\pi j} = (N/n) * \texttt{X\_WT2RAKE} * f_{NR}$ where f_{NR} is the sum of `X_LLCPWT` over the sum of `X_WT2RAKE`.

The samples, S_p and S_{np}, are combined and the propensity of being in S_p is estimated via logistic regression. The n closest matches in S_{np}, found using the R package Matching with the same matching criterion as described in Subsection 5.1, are retained for estimation. The

matching reduces the size of S_{np} to be the same ($n = 500$) as that of S_p. The weights $\tilde{w}_j$ and $\tilde{w}_{\pi j}$ from the matching person in S_p are assigned to person j in S_{np}. These weights were used to calculate estimated proportions, $\hat{\bar{Y}}_{M1}$, $\hat{\bar{Y}}_{M2}$, $\hat{\bar{Y}}_{MC1}$, $\hat{\bar{Y}}_{MC2}$, and their associated variance estimators. Estimators of the proportions of persons who reported heart attacks, diabetes, or strokes were computed based on the estimators of totals divided by $\hat{N} = \sum_{S_{np}} \tilde{w}_j$. Because of the way full-sample BRFSS weights are constructed, the variation $\hat{N}$ from sample to sample is minimal so that $\hat{N}$ is treated as a constant for variance estimation.

For $\hat{\bar{Y}}_{MC1}$ and $\hat{\bar{Y}}_{MC2}$ the calibration model used main effects for `SEX`, `X_AGE`, `MARITAL`, `EMPLOY1`, `EDUCA`, `INCOME2`, `EXERANY2`, and `SMOKE100` plus the continuous variable `WEIGHT2`. After some testing, the race variable was not included since it did not improve predictions once the other covariates were in the model. Calibration was done with the R survey package (Lumley 2020).

We also computed two versions of the doubly robust estimator for comparison. The two alternatives differed in the covariates used in the propensity and calibration models. The first, $\hat{\bar{Y}}_{DR1}$, used propensity and calibration models with the same covariates as the calibration model for $\hat{\bar{Y}}_{MC1}$ and $\hat{\bar{Y}}_{MC2}$. The second, $\hat{\bar{Y}}_{DR2}$, used propensity and calibration models that have a much smaller set of covariates: an intercept, `SEX`, `X_AGE`, `EMPLOY1`, and `EDUCA`. Although not reported here, we also ran simulations for a DR estimator that had a propensity model with an intercept, the interactions of `INCOME2` with `X_AGE`, `EDUCA` with `X_AGE`, and `INCOME2` with `EDUCA`. These interactions were determined from a regression tree analysis, and the covariates were recoded for the interactions to be binary. `INCOME2` was recoded to less than or greater than or equal to USD 25,000; `X_AGE` to less than 55 years or greater than or equal to 55 years; `EDUCA` to less than high school or high school or more. The logistic propensity model for being in S_{np} based on the merged data set of S_p and S_{np} was estimated using the method described in Wang et al. (2021). For this DR alternative, the same calibration model was used as f*or* $\hat{\bar{Y}}_{MC1}$ and $\hat{\bar{Y}}_{MC2}$. Because the summary results for this DR estimator were similar to those for $\hat{\bar{Y}}_{DR1}$, we have omitted them.

This process was repeated 5,000 times for each of the three analysis variables. The relative biases, the variances and the mean squared errors (MSEs) of the three point estimators across the 5,000 samples are summarized in Table 8. For all three analysis variables the biases of $\hat{\bar{Y}}_{M1}$ and $\hat{\bar{Y}}_{M2}$ are positive, ranging from 4.8% for diabetes with $M1$ to 15.7% for heart attack for $M2$. Recall that $M1$ is a type of π-estimator with the π-weight taken from the matched case in the probability sample. In this example, $M2$ is a raked estimator with the weight being the raked weight from the matched case in S_p. In contrast, the $MC1$, $MC2$, and $DR1$ estimators have serious negative biases, ranging from -21.6% to -17.5%. The DR estimator with fewer covariates, $\hat{\bar{Y}}_{DR2}$, has the smallest absolute relbias for two of the three variables: -6.5% for heart attack and -9.1% for stroke. Its -9.8% relbias for diabetes is an improvement over that of $\hat{\bar{Y}}_{DR1}$ but is still undesirably large. Although modeling suggested that the extensive list of covariates used in $MC1$, $MC2$, and $DR1$ was needed to correct for coverage bias, the simulations illustrate the well-known phenomenon that overloading a regression model with predictors leads to instability. $DR2$ has the smallest MSE for all three variables and might be preferred on that basis, but its bias results in poor confidence interval coverage as shown next in Table 9. All of the relbiases in Table 8 are highly significantly different from zero based on t-tests.

Table 8. Simulation study with BRFSS population: Monte Carlo percent relative biases, variances and mean squared errors of the point estimators

Estimator	Relative bias (%)	Variance ($\times 10^4$)	MSE ($\times 10^4$)	Ratio to min MSE
Heart attack				
$\hat{\bar{Y}}_{M1}$	12.4	2.57	2.85	1.51
$\hat{\bar{Y}}_{M2}$	15.7	3.76	4.22	2.23
$\hat{\bar{Y}}_{MC1}$	-20.5	1.73	2.51	1.32
$\hat{\bar{Y}}_{MC2}$	-20.3	2.03	2.78	1.47
$\hat{\bar{Y}}_{DR1}$	-21.6	1.61	2.47	1.30
$\hat{\bar{Y}}_{DR2}$	-6.5	1.81	1.89	1.00
Diabetes				
$\hat{\bar{Y}}_{M1}$	4.8	5.71	6.02	1.09
$\hat{\bar{Y}}_{M2}$	6.4	8.07	8.62	1.56
$\hat{\bar{Y}}_{MC1}$	-20.1	4.47	9.88	1.79
$\hat{\bar{Y}}_{MC2}$	-19.8	5.26	10.53	1.91
$\hat{\bar{Y}}_{DR1}$	-20.6	3.78	9.48	1.72
$\hat{\bar{Y}}_{DR2}$	-9.8	4.22	5.52	1.00
Stroke				
$\hat{\bar{Y}}_{M1}$	11.2	1.73	1.83	1.38
$\hat{\bar{Y}}_{M2}$	15.5	2.59	2.80	2.10
$\hat{\bar{Y}}_{MC1}$	-18.4	1.33	1.62	1.21
$\hat{\bar{Y}}_{MC2}$	-17.5	1.59	1.85	1.39
$\hat{\bar{Y}}_{DR1}$	-20.3	1.17	1.52	1.14
$\hat{\bar{Y}}_{DR2}$	-9.1	1.26	1.33	1.00

Table 9. Simulation study with BRFSS population: Percent relative biases and 95% confidence interval coverages of the variance estimators

Estimator	RB.Empvar (%)	RB.MSE (%)	CI coverage (%)
Heart attack			
$v_\xi(\hat{\bar{Y}}_{M1})$	-28.4	-35.5	89.6
$v_{R\pi}(\hat{\bar{Y}}_{M1})$	-4.0	-13.4	92.8
$v_{R\pi\xi}(\hat{\bar{Y}}_{M1})$	-10.1	-18.9	92.6
$v_\xi(\hat{\bar{Y}}_{M2})$	-26	-33.9	89.6
$v_{R\pi}(\hat{\bar{Y}}_{M2})$	-22.7	-31.1	89.5
$v_{R\pi\xi}(\hat{\bar{Y}}_{M2})$	-26.9	-34.8	89.4
$v_\xi(\hat{\bar{Y}}_{MC1})$	7.5	-25.8	81.6
$v_R(\hat{\bar{Y}}_{MC1})$	-14.4	-40.9	73.1
$v_\xi(\hat{\bar{Y}}_{MC2})$	-8.1	-33.1	80.4
$v_{R\pi}(\hat{\bar{Y}}_{MC2})$	-5.0	-30.9	77.2

Table 9. Continued.

Estimator	RB.Empvar (%)	RB.MSE (%)	CI coverage (%)
$v_{R\pi\xi}(\hat{\bar{Y}}_{MC2})$	13.8	-17.2	84.4
$v(\hat{\bar{Y}}_{DR1})$	-22.4	-49.3	68.4
$v(\hat{\bar{Y}}_{DR2})$	-19.6	-22.9	82.2
Diabetes			
$v_{\xi}(\hat{\bar{Y}}_{M1})$	-26.3	-30.1	89.8
$v_{R\pi}(\hat{\bar{Y}}_{M1})$	3.1	-2.2	94.8
$v_{R\pi\xi}(\hat{\bar{Y}}_{M1})$	2.1	-3.1	94.9
$v_{\xi}(\hat{\bar{Y}}_{M2})$	-25.0	-29.8	89.9
$v_{R\pi}(\hat{\bar{Y}}_{M2})$	-8.9	-14.8	92.9
$v_{R\pi\xi}(\hat{\bar{Y}}_{M2})$	-9.7	-15.5	92.9
$v_{\xi}(\hat{\bar{Y}}_{MC1})$	-5.3	-57.2	72.9
$v_{R}(\hat{\bar{Y}}_{MC1})$	-13.4	-60.9	68.4
$v_{\xi}(\hat{\bar{Y}}_{MC2})$	-19.6	-59.8	71.7
$v_{R\pi}(\hat{\bar{Y}}_{MC2})$	1.5	-49.3	75.6
$v_{R\pi\xi}(\hat{\bar{Y}}_{MC2})$	8.3	-45.9	79.4
$v(\hat{\bar{Y}}_{DR1})$	-19.2	-67.8	61.6
$v(\hat{\bar{Y}}_{DR2})$	-18.8	-37.9	80.0
Stroke			
$v_{\xi}(\hat{\bar{Y}}_{M1})$	-25.6	-29.9	87.9
$v_{R\pi}(\hat{\bar{Y}}_{M1})$	-1.8	-7.5	90.1
$v_{R\pi\xi}(\hat{\bar{Y}}_{M1})$	-9.7	-14.9	90.5
$v_{\xi}(\hat{\bar{Y}}_{M2})$	-24.7	-30.2	87.8
$v_{R\pi}(\hat{\bar{Y}}_{M2})$	-24.6	-30.1	87.4
$v_{R\pi\xi}(\hat{\bar{Y}}_{M2})$	-29.8	-35.0	87.4
$v_{\xi}(\hat{\bar{Y}}_{MC1})$	-2.3	-19.8	79.7
$v_{R}(\hat{\bar{Y}}_{MC1})$	-16.1	-31.1	73.2
$v_{\xi}(\hat{\bar{Y}}_{MC2})$	-18.3	-29.8	78.8
$v_{R\pi}(\hat{\bar{Y}}_{MC2})$	-13.3	-25.6	76.9
$v_{R\pi\xi}(\hat{\bar{Y}}_{MC2})$	-1.7	-15.6	82.2
$v(\hat{\bar{Y}}_{DR1})$	-19.7	-38.3	70.1
$v(\hat{\bar{Y}}_{DR2})$	-17.7	-22.1	80.1

Table 9 shows the percent relative biases of the variance estimators with respect to the empirical variance of each estimator of the proportion and with respect to the empirical MSE. These are labeled RB.Empvar (%) and RB.MSE (%). For the most part, the relative biases are negative. With respect to the MSE, all are negative owing to the biases of the point estimators of the proportions which inflate the MSEs. The coverage rates for 95% normal approximation confidence intervals is generally poor because the intervals are centered at the wrong place due to the biases of the estimators of proportions. Only the combinations of $\hat{\bar{Y}}_{M1}$ with $v_{R\pi}$ and $v_{R\pi\xi}$ have coverage rates above 90%.

Finally, as an experiment we also increased the sample sizes to $n = 1{,}000$ for the nonprobability sample and $M = 5{,}000$ for the initial probability sample. The increased sample sizes had no effect on the biases of the point estimates of means. (Results are omitted here.)

In summary, substantial differences between the web and non-web subpopulations result in serious coverage problems for the nonprobability sample S_{np}, causing all estimators to be biased. Furthermore, calibration adjustment after matching does not reduce the biases of the matched estimators. The better of the doubly robust estimators does produce the smallest MSEs in the simulations, but, by no means does it reduce the biases of estimated means to a negligible level.

7. Conclusion

In this article we present several alternative estimators when a nonprobability sample, S_{np}, is matched to a probability sample, S_p. The general setting is that the nonprobability sample is weighted by assigning the weight from an S_p unit to its matched unit in the nonprobability sample. Particular cases are (A) the weight from S_p is its π-weight, (B) the weight from S_p is a GREG weight, (C) case (A) with the nonprobability sample being calibrated with a linear model, and (D) case (B) with S_{np} calibrated with a linear model. Under some restrictive conditions that may be difficult to satisfy in practice, these estimators can be approximately unbiased. The key requirement is that the actual propensity of a unit's being observed in the nonprobability sample should be equal or close to the inclusion probability of the unit to which it is matched in the probability sample.

Three simulation studies illustrated several points about the matched estimator and the doubly robust estimator, which is included for comparison. Study I used artificial data where the variable to be analyzed follows a linear model with a single covariate X, which was also used to create strata. The sample designs for both S_p and S_{np} were stratified simple random sampling with the design for S_{np} treated as unknown. In this case, matching on X was reliable and all estimators were unbiased. In fact, three of four of the matching estimators had a smaller MSE than the doubly robust estimator.

The second simulation used the same artificial population and S_p sample design as Study I, but S_{np} was selected with probabilities (treated as unknown) that decreased with X. In this example, the inclusion probabilities for the nonprobability sample are far from those in the probability sample used for matching. Consequently, the matched estimators without calibration are biased. However, calibration corrects the biases and the matched, calibrated estimator has a smaller MSE than the doubly robust estimator.

The third simulation used a real population (BRFSS) in which persons who had accessed the internet in the previous 30 days were treated as a nonprobability sample from the full US adult population. Since there was no control over how the nonprobability units were selected, this mirrored a situation that would be faced in practice. The prevalence of three health conditions was estimated. The prevalences differed considerably between the part of the population that was covered by S_{np} and the part that was not. The persons who did not use the internet were older, less educated, lower income, and less healthy than the internet users. These differences led to all estimators in the study being biased. Calibrating the matching estimators on a list of covariates did not correct their biases. In addition,

doubly robust (DR) estimation, which has been touted as one of the better options, needs to be carefully implemented to be effective. Including too many covariates in a DR estimator can destabilize the estimator (as in any regression) even though an extensive list of predictors may, in principle, be related to undercoverage of the population by the nonprobability sample.

The failure in the real data study has several, potential contributing factors, including poor matches between the nonprobability and probability units, inadequate models for the propensity of being observed in the nonprobability sample, and poor calibration models for predicting the health characteristics analyzed. However, the facts that the nonprobability sample does not cover the target population, and the noncovered units differ both on the distributions of the analytic variables and covariates is the critical problem. Some diagnostics have been devised for detecting nonignorability of selection of a nonprobability sample (e.g., see Andridge et al. 2019; Little et al. 2019). These diagnostics will signal non-ignorability if the means of covariates in S_{np} and the target population are sufficiently different. Thus, they might be a way forward in the BRFSS application.

However, if the variables to be analyzed differ between S_{np} and the target population but covariate distributions do not, the diagnostics will not alert an analyst to trouble, and poor inferences will still be made from the nonprobability sample. The type of coverage error in the BRFSS study is an example of what can happen in nonprobability samples, generally, and may be a problem that no amount of sophisticated mathematics is likely to correct.

8. Appendix

This appendix shows the details of variance calculations given in earlier sections. Several assumptions are used in the results below. These apply as N and $n \to \infty$.

(1) $\pi_j = O(N/n)$, $R_j = O(N/n)$ and $n/N \to 0$.
(2) $\tilde{\mathbf{A}}_U$ and $\tilde{\mathbf{A}}_U^*$ are $O(N)$.
(3) $V_\pi(\mathbf{X}_p) = O(N^2/n)$.
(4) $V_R(\mathbf{X}_{np}) = O(N^2/n)$.
(5) When $R_j = \pi_j$, $N^{-1}\tilde{\mathbf{A}}_p$ and $N^{-1}\tilde{\mathbf{A}}_{np}(\pi)$ both converge in probability to $N^{-1}\tilde{\mathbf{A}}_u = N^{-1}\sum_U \mathbf{x}_j\mathbf{x}_j^T/\tilde{\sigma}_j^2$.
(6) $N^{-1}\mathbf{A}_{np}^*(\tilde{w})$ converges in probability to $N^{-1}\tilde{\mathbf{A}}_u^* = N^{-1}\sum_U \mathbf{x}_j\mathbf{x}_j^T/\tilde{\sigma}_j^*$.
(7) When $R_j = \pi_j$, $\tilde{\mathbf{A}}_p^{-1}\sum_{S_{np}} \frac{\mathbf{x}_j y_j}{\pi_j \tilde{\sigma}_j^2} \xrightarrow{p} \tilde{\mathbf{B}}_U$ and $\left[\tilde{\mathbf{A}}_{np}^*(\tilde{w})\right]^{-1}\sum_{S_{np}} \frac{\mathbf{x}_j y_j}{\pi_j \sigma_j^{*2}} \xrightarrow{p} \mathbf{B}_U^*$.
(8) When $R_j = \pi_j$, $\sqrt{n}\,(\hat{\mathbf{X}}_p - \mathbf{X}_U)/N$, $\sqrt{n}\,(\hat{\mathbf{X}}_{np}(\pi) - \mathbf{X}_U)/N$, and $\sqrt{n}\,(\hat{\mathbf{X}}_{np}(\tilde{w}) - \mathbf{X}_U)/N$ are asymptotically multivariate normal with mean $\mathbf{0}$.

8.1. ξ-expectation of the With-Replacement Variance Estimator under Case (1)

To compute the ξ-expectation of $v_R\left(\hat{Y}_{M1}\right)$ in Subsection 3.3 under case (1), define $r_j = \tilde{w}_j y_j - \frac{1}{n}\sum_{j' \in S_{np}} \tilde{w}_{j'} y_{j'}$. Since $\tilde{w}_j = \pi_j^{-1}$, this can be rewritten as

$$r_j = \frac{n-1}{n}\frac{y_j}{\pi_j} - \frac{1}{n}\sum_{j' \neq j \in S_{np}} \frac{y_{j'}}{\pi_{j'}}.$$

The ξ-expectation of r_j^2 is then

$$E_\xi\left(r_j^2\right) = V_\xi(r_j) + \left[E_\xi(r_j)\right]^2$$

$$= \left(\frac{n-1}{n}\right)^2 \frac{\sigma_j^2}{\pi_j^2} + \frac{1}{n^2}\left(\sum_{j' \neq j \in S_{np}} \frac{\sigma_{j'}^2}{\pi_{j'}^2}\right) + \left\{\left(\frac{\mathbf{x}_j}{\pi_j} - \frac{1}{n}\sum_{j' \in S_{np}} \frac{\mathbf{x}_{j'}}{\pi_{j'}}\right)^T \beta\right\}^2.$$

Adding and subtracting σ_j^2/π_j^2 in the second term, summing over S_{np}, and doing some algebra leads to

$$E_\xi(v_{R\pi}) = \sum_{S_{np}} \frac{\sigma_j^2}{\pi_j^2} + \frac{n}{n-1}\sum_{j \in S_{np}}\left\{\left(\frac{\mathbf{x}_j}{\pi_j} - \frac{1}{n}\sum_{j' \in S_{np}} \frac{\mathbf{x}_{j'}}{\pi_{j'}}\right)^T \beta\right\}^2$$

as noted in Subsection 3.3. That is, $v_{R\pi}$ is an overestimate of the model variance under Equation (1). However, because $\hat{Y}_{M1}$ is model-biased, $v_{R\pi}$ will not appropriately estimate the ξ mean square error despite its overestimating the ξ-variance.

To derive the $R\pi\xi$-variance, note that

$$V_{R\pi\xi}\left(\hat{Y}_{M1}|S_p, S_{np}\right) = V_{R\xi}\left(\hat{Y}_{M1}|S_p, S_{np}\right)$$

$$= E_R\left\{V_\xi\left(\hat{Y}_{M1}|S_p, S_{np}\right)\right\} + V_R\left\{E_\xi\left(\hat{Y}_{M1}|S_p, S_{np}\right)\right\}.$$

Using the independence of the Y's under Equation (1), the first term is $\sum_U \sigma_j^2/\pi_j$. The second term is $V_R\left\{E_\xi\left(\hat{Y}_{M1}|S_p, S_{np}\right)\right\} = V_R\left(\hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)^T\beta\right) = \beta^T V_R\left(\hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)\beta$. Combining gives the expression shown in Equation (9).

8.2. *Variance of Matched Estimator $\hat{Y}_{M2}$ Under Case (2)*

Following similar steps to those in Särndal et al. (1992, sec. 6.6) and using condition (7), $\hat{Y}_{M2}$ can be approximated as

$$\hat{Y}_{M2} \doteq \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)\tilde{\mathbf{B}}_U, \tag{29}$$

where $\tilde{\mathbf{B}}_U = \left(\sum_U \frac{\mathbf{x}_j\mathbf{x}_j^T}{\pi_j\sigma_j^2}\right)^{-1}\left(\sum_U \frac{\mathbf{x}_j y_j}{\pi_j\sigma_j^2}\right)$.

Using the formula for total variance across the R and π distributions (denoted by $V_{R\pi}$) gives

$$V_{R\pi}(\hat{Y}_{M2}) = E_R V_\pi(\hat{Y}_{M2}|S_{np}) + V_R E_\pi(\hat{Y}_{M2}|S_{np}). \tag{30}$$

Working term by term in Equation (30) and using the approximation to $\hat{Y}_{M2}$ in Equation (29), we have

$$E_R V_\pi(\hat{Y}_{M2}|S_{np}) \doteq E_R V_\pi\left(\hat{Y}_{np}(\pi) + (\mathbf{X}_U - \hat{\mathbf{X}}_p)^T\tilde{\mathbf{B}}_U|S_{np}\right) = \tilde{\mathbf{B}}_U^T V_\pi(\hat{\mathbf{X}}_p)\tilde{\mathbf{B}}_U,$$

because $\hat{Y}_{np}(\pi)$ has zero R-variance given that S_{np} is fixed. To get the second term in Equation (30), note that $V_R E_\pi\left(\hat{Y}_{M2}|S_{np}\right) \doteq V_R\left(\hat{Y}_{np}(\pi)\right)$ assuming that $\hat{\mathbf{X}}_p$ is π-unbiased.

Combining these results, the variance across the R- and π-distributions is

$$V_{R\pi}\left(\hat{Y}_{M2}\right) \doteq V_R\left(\hat{Y}_{np}(\pi)\right) + \tilde{\mathbf{B}}_U^T V_\pi\left(\hat{\mathbf{X}}_P\right)\tilde{\mathbf{B}}_U.$$

as shown in Equation (13).

Turning to the $R\pi\xi$-variance, the total variance formula is given by Equation (8). The $E_\pi V_\xi$ term is $\sum_{S_{np}} \tilde{\sigma}_j^2 E_\pi\left(g_j^2\right)/\pi_j^2$. Using a Taylor series approximation as in Särndal et al. (1992, sec. 6.6), we have

$$g_j \doteq \pi_j^{-1}\left[1 + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_U^{-1}\mathbf{x}_j/\tilde{\sigma}_j^2\right]. \tag{31}$$

It follows that

$$\begin{aligned} E_\pi V_\xi\left(\hat{Y}_{M2}\right) &\doteq \sum_{S_{np}} \frac{\sigma_j^2}{\pi_j^2}\left\{1 + \frac{\mathbf{x}_j^T}{\tilde{\sigma}_j^2}\tilde{\mathbf{A}}_U^{-1} E_\pi\left[\left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)\left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T\right]\tilde{\mathbf{A}}_U^{-1}\frac{\mathbf{x}_j}{\tilde{\sigma}_j^2}\right\} \\ &= \sum_{S_{np}} \frac{\sigma_j^2}{\pi_j^2}\left\{1 + \frac{\mathbf{x}_j^T}{\tilde{\sigma}_j^2}\tilde{\mathbf{A}}_U^{-1} V_\pi\left(\hat{\mathbf{X}}_p\right)\tilde{\mathbf{A}}_U^{-1}\frac{\mathbf{x}_j}{\tilde{\sigma}_j^2}\right\}. \end{aligned} \tag{32}$$

Thus,

$$E_R E_\pi V_\xi\left(\hat{Y}_{M2}\right) = \sum_U R_j \frac{\sigma_j^2}{\pi_j^2} + \sum_U R_j \frac{\sigma_j^2}{\pi_j^2}\frac{\mathbf{x}_j^T}{\tilde{\sigma}_j^2}\tilde{\mathbf{A}}_U^{-1} V_\pi\left(\hat{\mathbf{X}}_p\right)\tilde{\mathbf{A}}_U^{-1}\frac{\mathbf{x}_j}{\tilde{\sigma}_j^2}.$$

Under the order assumptions at the beginning of this appendix, the first term above is $O(N^2/n)$ while the second is $O(N^2/n^2)$. Thus, we use the approximation $E_R E_\pi V_\xi\left(\hat{Y}_{M2}\right) \doteq \sum_U R_j\sigma_j^2/\pi_j^2$.

The second term in Equation (8) is $E_R V_\pi E_\xi\left(\hat{Y}_{M2}\right)$. Expanding and collecting terms gives

$$\begin{aligned} V_\pi E_\xi\left(\hat{Y}_{M2}\right) &= V_\pi\left(\sum_{S_{np}} \frac{g_j}{\pi_j}\mathbf{x}_j\beta\right) \\ &= V_\pi\left(\sum_{S_{np}} \frac{\mathbf{x}_j}{\pi_j}\beta + \left(\mathbf{X}_U - \mathbf{X}_p\right)^T \tilde{\mathbf{A}}_p^{-1}\tilde{\mathbf{A}}_{np}\beta\right). \end{aligned}$$

Under condition (v) above, $\tilde{\mathbf{A}}_p^{-1}\tilde{\mathbf{A}}_{np}$ converges to the $C \times C$ identity matrix and $E_R V_\pi E_\xi\left(\hat{Y}_{M2}\right) = \beta^T V_\pi\left(\hat{\mathbf{X}}_p\right)\beta$.

The third term in Equation (8) gives $V_R E_\pi E_\xi\left(\hat{Y}_{M2}\right)$. Frist, compute $E_\pi E_\xi\left(\hat{Y}_{M2}\right) = E_\pi\left(\sum_{S_{np}} \frac{g_j}{\pi_j}\mathbf{x}_j\beta\right)$. Using the approximation to g_j in Equation (31), $E_\pi\left(g_j\right) \doteq 1$ and $E_\pi E_\xi\left(\hat{Y}_{M2}\right) \doteq \hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\beta$. Consequently, the third term is $V_R E_\pi E_\xi\left(\hat{Y}_{M2}\right) \doteq \beta^T V_R\left(\hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)\beta$.

Combining results for the three terms in Equation (8) gives

$$V_{R\pi\xi}\left(\hat{Y}_{M2}\right) \doteq \sum_{U} R_j \frac{\sigma_j^2}{\pi_j^2} + \beta^T V_{\pi}\left(\hat{\mathbf{X}}_p\right)\beta + \beta^T V_R\left(\hat{\mathbf{X}}_{\boldsymbol{np}}(\pi)\right)\beta$$

as shown in Equation (15).

8.3. Approximation to $\left(\hat{Y}_{MC2}\right)$ in Case (2)

When S_p has case (2) weights, $\tilde{w}_j = g_j/\pi_j$ with g_j defined in Equation (3). The matched estimator after calibration then equals

$$\hat{Y}_{MC2} = \sum_{S_{np}} g_j^* g_j y_j/\pi_j,$$

where

$$g_j^* = 1 + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\tilde{w})\right)^T \left[\tilde{\mathbf{A}}_{np}^*(\tilde{w})\right]^{-1} \mathbf{x}_j/\sigma_j^{*2}.$$

Multiplying g_j^* by g_j defined in Equation (3) and substituting in the formula for $\hat{Y}_{MC2}$ gives

$$\hat{Y}_{MC2} = \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \sum_{S_{np}} \frac{\mathbf{x}_j y_j}{\pi_j \tilde{\sigma}_j^2} + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w})\right)^T \left[\tilde{\mathbf{A}}_{np}^*(\tilde{w})\right]^{-1} \sum_{S_{np}} \frac{\mathbf{x}_j y_j}{\pi_j \sigma_j^{*2}}$$

$$+ \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_p^{-1} \sum_{S_{np}} \frac{\mathbf{x}_j \mathbf{x}_j^T y_j}{\pi_j \tilde{\sigma}_j^2 \sigma_j^{*2}} \left[\tilde{\mathbf{A}}_{np}^*(\tilde{w})\right]^{-1} \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w})\right). \tag{33}$$

Using conditions (5), (6), and (7), the orders of the second, third, and fourth terms in Equation (33) are $O_p(N/\sqrt{n})$, $O_p(N/\sqrt{n})$, and $O_p(N\sqrt{n})$. The calibrated estimator can then be approximated by

$$\hat{Y}_{MC2} \doteq \hat{Y}_{np}(\pi) + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \mathrm{B}_U + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{\boldsymbol{np}}(\tilde{w})\right)^T \mathrm{B}_U^*. \tag{34}$$

8.4. Variance of Matched Estimator $\hat{Y}_{MC2}$ in Case (2)

To compute the ξ model variance under case (2), we break $\sum_U \mathbf{x}_j y_j/\tilde{\sigma}_j^2$ and $\sum_U \mathbf{x}_j y_j/\sigma_j^{*2}$ into sums over S_{np} and $U - S_{np}$. Equation (34) can then be expressed as

$$\hat{Y}_{MC2} \doteq \sum_{S_{np}} y_j \left(\frac{1}{\pi_j} + F_j\right) + \sum_{U-S_{np}} y_j F_j,$$

where

$$F_j = \left(\mathbf{X}_U - \hat{\mathbf{X}}_p\right)^T \tilde{\mathbf{A}}_U^{-1} \frac{\mathbf{x}_j}{\tilde{\sigma}_j^2} + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w})\right)^T \tilde{\mathbf{A}}_U^{*-1} \frac{\mathbf{x}_j}{\tilde{\sigma}_j^2}\ .$$

Applying conditions (2) and (8), $F_j = O_p(n^{-1/2})$. Since units in S_{np} and $U - S_{np}$ are independent under model (1), the ξ-variance is

$$V_{\xi}\left(\hat{Y}_{MC2}\right) \doteq \sum_{S_{np}} \sigma_j^2 \left(\frac{1}{\pi_j} + F_j\right)^2 + \sum_{U-S_{np}} \sigma_j^2 F_j^2 \tag{35}$$

$$= \sum_{S_{np}} \left(\frac{\sigma_j^2}{\pi_j} \right)^2 \left[1 + O_p\left(N/n^{3/2} \right) \right]. \tag{36}$$

The $R\pi\xi$-variance can be calculated using the total variance formula in Equation (8). First, when $R_j = \pi_j$, $E_R E_\pi V_\xi \left(\hat{Y}_{MC2} | S_p, S_{np} \right) = \sum_U \left(\sigma_j^2 / \pi_j \right)$. The second term in Equation (8) is

$$E_R V_\pi E_\xi \left(\hat{Y}_{MC2} \right) \doteq E_R V_\pi \left[\hat{\mathbf{X}}_{np}(\pi)^T \beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_p \right)^T \beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w}) \right)^T \beta \right]$$

$$= \beta^T V_\pi \left(\hat{\mathbf{X}}_p \right) \beta.$$

The third term in Equation (8) is

$$V_R \left[E_\pi E_\xi \left(\hat{Y}_{MC2} \right) \right] = V_R E_\pi \left[\hat{\mathbf{X}}_{np}(\pi)^T \beta + \left(\mathbf{X}_U - \hat{\mathbf{X}} p \right)^T \beta + \left(\mathbf{X}_U - \hat{\mathbf{X}}_{np}(\tilde{w}) \right)^T \beta \right]$$

$$= V_R \left\{ E_\pi \left[\hat{\mathbf{X}}_{np}(\pi) - \hat{\mathbf{X}}_{np}(\tilde{w}) \right]^T \beta \right\}.$$

Rewriting the term in brackets above leads to

$$\hat{\mathbf{X}}_{np}(\pi) - \hat{\mathbf{X}}_{np}(\tilde{w}) = \sum_{S_{np}} \frac{(1 - g_j)\mathrm{x}_j}{\pi_j}$$

$$= \left(\hat{\mathbf{X}}_p - \mathbf{X}_U \right)^T \tilde{\mathbf{A}}_p^{-1} \sum_{S_{np}} \frac{\mathbf{x}_j \mathbf{x}_j^T}{\pi j \tilde{\sigma}_j^2}$$

$$= \left(\hat{\mathbf{X}}_p - \mathbf{X}_U \right)^T \tilde{\mathbf{A}}_p^{-1} \tilde{\mathbf{A}}_{np}(\pi).$$

Applying condition (5) implies that $V_R \left[E_\pi E_\xi \left(\hat{Y}_{MC2} \right) \right] \doteq 0$. Combining results for the three terms in Equation (8) yields

$$V_{R\pi\xi} \left(\hat{Y}_{MC2} \right) \doteq \sum_U \frac{\sigma_j^2}{\pi_j} + \beta^T V_\pi \left(\hat{\mathbf{X}}_p \right) \beta.$$

An estimator of this variance is

$$\upsilon_{R\pi\xi} \left(\hat{Y}_{MC2} \right) \doteq \sum_{S_{np}} \left(\frac{\hat{e}_j^*}{\pi_j} \right)^2 + \hat{\hat{B}}_{np}(\pi)^T \upsilon_\pi \left(\hat{\mathbf{X}}_p \right) \hat{\hat{B}}_{np}(\pi)$$

as shown in Equation (27).

9. References

Andridge, R.R., B.T. West, R.J.A. Little, P.S. Boonstra, and F. Alvarado-Leiton. 2019. "Indices of non-ignorable selection bias for proportions estimated from non-probability samples." *Journal of the Royal Statistical Society,* 68(5): 1465–1483. Available at: https://doi.org/10.1111/rssc.12371 (accessed September 2022).

Baker, R., J.M. Brick, N.A. Bates, M.P. Battaglia, M.P. Couper, J.A. Dever, K. Gile, and R. Tourangeau. 2013. *Report of the AAPOR task force on non-probability sampling.* The American Association for Public Opinion Research, Deerfield, IL. Available at: https://www.aapor.org/aapor_main/media/mainsitefiles/nps_tf_report_final_7_revised_fnl_6_22_13.pdf (accessed September 2022).

Brick, J.M., and D. Williams. 2013. "Explaining rising nonresponse rates in cross-sectional surveys." *The Annals of the American Academy of Political and Social Science* 645(1): 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Caliendo, M, and S. Kopeinig. 2008. "Some practical guidance for the implementation of propensity score matching." *Journal of Economic Surveys* 2(1): 31–72. DOI: https://doi.org/10.1111/j.1467-6419.2007.00527.x.

Center for Disease Control and Prevention. 2017. *Weighting the BRFSS Data*. Available at: https://www.cdc.gov/brfss/annual_data/2017/pdf/weighting-2017-508.pdf.

Center for Disease Control and Prevention. 2023. *Behavioral Risk Factor Surveillance Survey*. Available at: http://www.cdc.gov/BRFSS.

Chen, Y., P. Li, and C. Wu. 2020. "Doubly robust inference with non-probability survey samples." *Journal of the American Statistical Association* 115: 2011–2021. DOI: https://doi.org/10.1080/01621459.2019.1677241.

Cochran, W.G. 1953. "Matching in analytical studies." *American Journal of Public Health* 43: 684–691. DOI: https://doi.org/10.2105/AJPH.43.6_Pt_1.684.

Dehejia, R. and S. Wahba. 2002. "Propensity-score matching methods for nonexperimental causal studies." *The Review of Economic and Statistics* 84(1): 151–161. Available at: http://www.mitpressjournals.org/doi/pdf/10.1162/003465302317331982 (accessed September 2022).

Deville, J.C., and C. Särndal. 1992. "Calibration estimators in survey sampling." *Journal of the American Statistical Association* 87(418): 376–382. DOI: https://doi.org/10.2307/2290268.

Elliott, M.R. and R. Valliant. 2017. "Inference for nonprobability samples." *Statistical Science* 32: 249–264. Available at: https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Inference-for-Nonprobability-Samples/10.1214/16-STS598.full. (accessed September 2022).

Gessendorfer, J., Beste, J., J. Drechsler, and J. Sakshaug. 2018. "Statistical Matching as a Supplement to Record Linkage: A Valuable Method to Tackle Nonconsent Bias?. *Journal of Official Statistics* 34(4): 909–933. DOI: https://doi.org/10.2478/jos-2018-0045.

Hansen, M.H., W.G. Madow, and B.J. Tepping. 1983. "An evaluation of model-dependent and probability sampling inferences in sample surveys." *Journal of the American Statistical Association* 78: 776–793.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. *Evaluating online nonprobability surveys, vendor choice matters: widespread errors found for estimates based on blacks and hispanics.* Technical report, Pew Research. Available at: https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/ (accessed September 2022).

Kim, J.-K., S. Park, Y. Chen, and C. Wu. 2021. "Combining non-probability and probability survey samples through mass imputation." *Journal of the Royal Statistical*

Society, Series A: Statistics in Society 184: 941–963. DOI: https://doi.org/10.1111/rssa.12696.

Lee, S. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22(2): 329–349. Available at: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/propensity-score-adjustment-as-a-weighting-scheme-for-volunteer-panel-web-surveys.pdf (accessed September 2022).

Little, R.J.A., B. West, P.S. Boonstra, and J. Hu. 2019. "Measures of the degree of departure from ignorable sample selection." *Journal of Survey Statistics and Methodology* 8(5): 932–964. DOI: https://doi.org/10.1093/jssam/smz023.

Lumley, T. 2020. "Survey: analysis of complex survey samples." R package version 4.1-1. Available at: https://cran.r-project.org/web/packages/survey/index.html (accessed September 2022).

MacKinnon, J.G., and H. White. 1985. "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics* 29(3): 305–325. DOI: https://doi.org/10.1016/0304-4076(85)90158-7.

Rao, J., W. Yung, and M Hidiroglou. 2002. "Estimating equations for the analysis of survey data using poststratification information." *Sankhyā Series A* 64: 364–378.

Rivers, D. 2007. "Sample matching for web surveys: Theory and application." In Proceedings of the Section on Survey Research Methods, July, Salt Lake City, Utah, USA. Available at: http://www.websm.org/uploadi/editor/1368187629Rivers_2007_Sampling_for_web_surveys.pdf (accessed September 2022).

Rivers, D., and D. Bailey. 2009. "Inference from matched samples in the 2008 U.S. national elections." In Proceedings of the American Statistical Association, Section on Survey Research Methods: 627–639. August, Washington D.C., USA. Available at: http://www.asasrms.org/Proceedings/y2009f.html (accessed September 2022).

Rosenbaum, P.R., and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55. Available at: https://doi.org/10.1093/biomet/70.1.41.

Rothman, K.J., S. Greenland, and T.L. Lash. 2008. *Modern Epidemiology*. Lippincott, Williams and Wilkins, 3rd edition.

Rubin, D. 1973. "Matching to remove bias in observational studies." *Biometrics* 29(1): 159–183. DOI: https://doi.org/10.2307/2529684.

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

Schonlau, M., Soest, A.V., A. Kapteyn, and M. Couper. 2009. "Selection bias in web surveys and the use of propensity scores. *Sociological Methods a Research* 37(3): 291–318. DOI: https://doi.org/10.1177/0049124108327128.

Sekhon, J.S. 2011. "Multivariate and propensity score matching software with automated balance optimization: The Matching package for R." *Journal of Statistical Software* 42(7): 1–52. DOI: https://doi.org/10.18637/jss.v042.i07.

Smith, J., and P. Todd. 2005. "Does matching overcome LaLonde's critique of non-experimental estimators?" *Journal of Econometrics* 125(1): 305–353. DOI: https://doi.org/10.1016/j.jeconom.2004.04.011.

Terhanian, G., and J. Bremer. 2012. "A smarter way to select respondents for surveys?" *International Journal of Market Research* 54(6): 751–780. DOI: https://doi.org/10.2501/IJMR-54-6-751-780.

Toluna. 2023. *Toluna delivers real-time consumer insights at the speed of the on-demand economy*. Available at: https://www.greenbook.org/company/Toluna.

U.S. Census Bureau. 2023. *American Community Survey*. Available at: https://www.census.gov/programs-surveys/acs.

Valliant, R. 2020. "Comparing alternatives for estimation from nonprobability samples." *Journal of Survey Statistics and Methodology* 8: 231–263. DOI: https://doi.org/10.1093/jssam/smz003.

Valliant, R. and J.A. Dever. 2011. "Estimating propensity adjustments for volunteer web surveys." *Sociological Methods and Research* 40: 105–137. DOI: https://doi.org/10.1177/0049124110392533.

Valliant, R., J.A. Dever, and F. Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York, 2nd edition.

Valliant, R., J.A. Dever, and F. Kreuter. 2020. *PracTools: Tools for Designing and Weighting Survey Samples*. R package version 1.2.8. Available at: https://CRAN.R-project.org/package = PracTools (accessedSeptember 2022).

Vavreck, L., and D. Rivers. 2008. "The 2006 cooperative congressional election study." *Journal of Elections, Public Opinion and Parties* 18(4): 35–66. DOI: https://doi.org/10.1080/17457280802305177.

Wang, L. 2020. *Improving External Validity of Epidemiologic Analyses by Incorporating Data from Population-Based Surveys*. PhD thesis, University of Maryland. Available at: http://hdl.handle.net/1903/26125, doi:10.13016/pogq-glbs (accessed September 2022).

Wang, L., Graubard, B.I., H. Katki, and Y. Li. 2020. "Improving external validity of epidemiologic cohort analyses: a kernel weighting approach." *Journal of the Royal Statistical Society, Statistics in Society, Series A* 183(3):1293–1311. Available at: https://rss.onlinelibrary.wiley.com/doi/10.1111/rssa.12564 (accessed September 2022).

Wang, L., R. Valliant, and Y. Li. 2021. "Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts." *Statistics in Medicine*: 1–14. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9122 (accessed September 2022).

Received September 2021
Revised May 2022
Accepted October 2022

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 79–101, http://dx.doi.org/10.2478/JOS-2023-0004

Using Eye-Tracking Methodology to Study Grid Question Designs in Web Surveys

Cornelia E. Neuert[1], *Joss Roßmann*[1], *and Henning Silber*[1]

Grid questions are frequently employed in web surveys due to their assumed response efficiency. In line with this, many previous studies have found shorter response times for grid questions compared to item-by-item formats. Our contribution to this literature is to investigate how altering the question format affects response behavior and the depth of cognitive processing when answering both grid question and item-by-item formats. To answer these questions, we implemented an experiment with three questions in an eye-tracking study. Each question consisted of a set of ten items which respondents answered either on a single page (large grid), on two pages with five items each (small grid), or on ten separate pages (item-by-item). We did not find substantial differences in cognitive processing overall, while the processing of the question stem and the response scale labels was significantly higher for the item-by-item design than for the large grid in all three questions. We, however, found that when answering an item in a grid question, respondents often refer to surrounding items when making a judgement. We discuss the findings and limitations of our study and provide suggestions for practical design decisions.

Key words: Web surveys; response behavior; cognitive processing; question design; eye-tracking methodology.

1. Introduction and Background

The use of grid questions is popular in self-administered surveys, such as web surveys. In a grid question format, respondents receive a series of substantially related items that share the same response scale. The items are usually presented in rows, and the response entry fields are presented in columns (Liu and Cernat 2018). An alternative approach of presenting items sharing the same response scale is the item-by-item design, where items are presented as stand-alone questions (Couper et al. 2013). Between those two extreme points of presenting a series of items are design choices that break the series of target items in smaller groups; for instance, by presenting a set of ten target items in two grids with five items (e.g., Couper et al. 2001; Grady et al. 2019). Each of these formats has benefits and drawbacks.

From a survey designers' perspective, the grid question format is an efficient way to ask multiple questions with the same response scale in a time- and space-saving manner (Couper et al. 2001; Couper et al. 2013; Tourangeau et al. 2004). From a respondent's perspective, the survey length and, thus, completion time is perceived to be shorter, and so is the perceived burden of answering the survey (Heerwegh 2009). Also, grouping items into a grid allows respondents to compare their answers as the content is perceived as belonging conceptually together (Heerwegh 2009; Tourangeau et al. 2004). The latter,

[1] GESIS – Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072 Mannheim, Germany. Emails: cornelia.neuert@gesis.org, joss.rossmann@gesis.org and henning.silber@gesis.org

which goes back to the principle of proximity of Gestalt psychology (Jenkins and Dillman 1995), facilitates comparative judgments and increases consistency of answers compared to when each item is considered in isolation (Couper 2008). However, this may also have disadvantages, such as artificially high inter-item correlations (Silber et al. 2018) due to respondents consistently giving the same answer to each item (a form of satisficing called straightlining or nondifferentiation; Krosnick and Alwin 1988).

Compared to an item-by-item design, in which each item is usually presented on a new page, grid questions present a large amount of information on one page. The amount of information and the effort required to answer the items increases with the size of a grid. An increasing number of items in rows and more response entry fields in columns imply a larger matrix, making navigation more difficult (Couper et al. 2013; Grady et al. 2019). With a cognitively more demanding task, respondents may get more easily confused and distracted, thereby increasing their actual or perceived response burden (Couper et al. 2013; Liu and Cernat 2018).

According to the theory of survey satisficing, the complexity of grid formats might encourage respondents to minimize time and effort for answering them thoroughly (Couper et al. 2013; Krosnick 1991). Congruent with that assumption, a large body of experimental research has shown that grid questions can have negative effects on data quality, which has been shown by higher rates of missing or non-substantive answers (i.e., "don't know"; Mavletova and Couper 2015; Roßmann et al. 2018; Toepoel et al. 2009), higher levels of non-differentiated answers (i.e., straightlining; DeBell et al. 2021; Roßmann et al. 2018; Tourangeau et al. 2004), and higher breakoff rates compared to item-by-item formats (Couper et al. 2013; Liu and Cernat 2018; Tourangau et al. 2004). Although nondifferentiation, item nonresponse, and similar response behaviors are generally viewed as undesirable response effects, there is the possibility that the less differentiated responses, and greater expressions of uncertainty (e.g., selecting "don't know") are closer to "truth"– that is, that grids actually help respondents to understand that their responses to individual items are (legitimately) close to each other, or recognize legitimate uncertainty; and that separating questions into an item-by-item format artificially magnifies differences between responses. While we acknowledge that this alternative interpretation is also plausible, we follow the general view that classifies the response behaviors as undesirable.

The faster completion times of grids compared to item-by-item question formats may also represent a form of superficial cognitive response processing and might increase measurement error (Couper et al. 2001; Peytchev 2005, cited in Couper et al. 2013; Roßmann et al. 2018; Tourangeau et al. 2004). Comparing response times between an item-by-item and a grid design, Roßmann et al. (2018) have shown that the response time for the first item did not differ between the two formats. This finding leads to the question of whether the longer response times within the item-by-item designs result from deeper cognitive processing of the item itself or from the response task of reading the question stem and the response scales each time in the item-by-item format.

In this study, we used eye-tracking methodology to gain more insights into the cognitive information process and response behavior when answering grid versus item-by-item question formats. Therefore, we employed an experiment with three questions. The respondents were randomly assigned to one of three question formats (item-by-item, two small grids, one large grid) while their eye movements were monitored.

2. Research Questions and Hypotheses

We investigated the following two research questions:

1. Does altering the question format affect the depth of cognitive processing when answering grid questions?
2. Does the differential depth of cognitive processing explain differences in response quality between the three different question designs?

Answering survey questions requires respondents to pass through four stages of cognitive processing (Tourangeau and Rasinski 1988; Tourangeau et al. 2000): (1) question comprehension, (2) retrieval of relevant information, (3) use of the information to arrive at a judgment, and (4) reporting of an answer within the response options provided. Each of these four stages of cognitive processing can be challenging for respondents, and thus, may contribute to the emergence of response effects. The theory of survey satisficing complements this framework by incorporating motivational components of the respondents (Krosnick 1991; Roßmann and Silber 2020). It states that respondents might alter their response behavior from complete and thorough execution of the four cognitive steps (i.e., optimizing) to less diligent or incomplete execution (i.e., satisficing) contingent on three factors: task difficulty, ability, and motivation. The higher the difficulty of answering a survey question and the lower a respondent's ability and motivation to perform the task, the higher is the chance of respondents showing satisficing response behavior (Krosnick 1991).

Referring to this theoretical framework, we propose specific hypotheses about how the design of grid questions can affect cognitive processing, and in consequence the survey response. While the design of grid questions could affect all four stages of cognitive processing, it seems likely that it mainly affects the processes of question comprehension and reporting of an answer. By using eye-tracking data, we can differentiate between different steps in the response process by observing response times and eye fixations for each part of the question (see Figure 1).

2.1. Hypotheses for the Stage of Question Comprehension

Comprehension includes such processes as attending to the question and instructions and identifying the information sought (Tourangeau et al. 2000). For grid questions, the comprehension stage requires respondents to attend to the question stem and the item texts. Particularly, the grid format promises efficient processing of the "question stem" compared to an item-by-item presentation. While in the former question design, respondents must attend to the question stem only once, they need to check whether the question is the same for each item in the item-by-item format. Thus, the more items are grouped on a single page of a questionnaire; the less the relative effort respondents need to invest in processing the question stem.

Hypothesis 1

"The fewer the number of items presented on a survey page; the more time is spent on processing the question stem on average across all target items."

Accordingly, the time spent processing the question stem should be highest for an item-by-item format, where the question stem is repeatedly displayed with each item, and

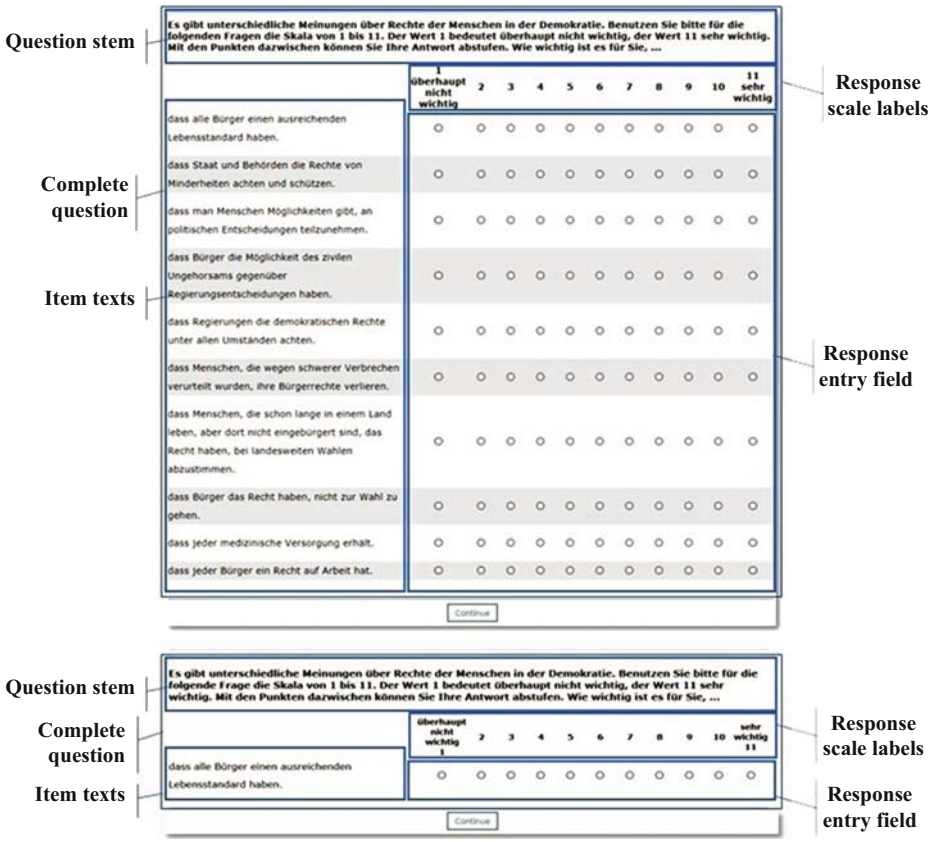


Fig. 1. Illustration of the different parts of the question and Areas of Interest (AOI) for the analysis of eye-tracking data. Above is an example of the question parts/AOIs for the grid question design of Question 2, below for the item-by-item design.

lowest for the presentation in one large grid, where the question stem is displayed only once.

In contrast, grouping more items on a page should not lead to more efficient processing of *item texts* under the condition of optimizing response behavior, that is, if respondents are able and motivated to thoroughly read and answer the items. However, grouping many items on a survey page can increase the actual or perceived complexity and burden of a grid question (Couper et al. 2013; Liu and Cernat 2018). The higher complexity of grid questions may discourage respondents, and thus increase the chances that they alter their response strategy to satisficing. In this regard, it seems plausible that the likelihood of superficial or incomplete processing of item texts increases with each additional item that is grouped on a survey page. This can be further reinforced because respondents can use the previously answered items as sources of information and orientation which allows them to answer the following items in a similar way without thinking thoroughly about each single one.

Hypothesis 2

"The fewer the number of items presented on a survey page; the more time is spent on processing the item texts on average, across all target items."

Accordingly, the time spent processing the item texts should be highest for an item-by-item format, and lowest in one large grid.

2.2. Hypotheses for the Stage of Reporting an Answer

The stage of reporting an answer includes two groups of processes: mapping the answer onto the response options and editing the response (Tourangeau et al. 2000). Regarding grid questions, the reporting and response selection stage particularly concerns the processing of the (numeric or verbal) response scale labels and the response entry fields (i.e., the response options). As for the processing of the question stem in the comprehension stage, the grid format promises efficient processing of the "response scale labels" compared to an item-by-item presentation. In the former, respondents must attend to response scale labels only once, whereas they need to repeatedly check whether they have changed or not in the item-by-item format. Thus, the more items are grouped on a single survey page using a common set of response scale labels; the relatively less effort respondents need to invest in processing them per item; thereby increasing item-efficiency.

Hypothesis 3

"The fewer the number of items presented on a survey page; the more time is spent on processing the response scale labels on average, across all target items."

Accordingly, the time spent processing the response scale labels should be highest for an item-by-item format, and lowest for the presentation in one large grid.

For the "response entry fields", the higher complexity that results from the larger size of the grid most likely increases the chance that respondents experience navigational difficulties in reporting responses. In other words, the fewer items are presented on a survey page, the easier it should be for respondents to select the answer that applies to them from the available response entry fields.

Hypothesis 4

"The fewer the number of items presented on a survey page; the less time is spent on processing the response entry fields on average, across all target items."

Also, the grouping of items in grids may encourage respondents to edit their responses for inter-item consistency or other criteria. This would additionally increase the processing time of the response entry fields in a grid.

However, some respondents may be discouraged by the daunting size of the grid and alter their response behavior to satisficing. As a consequence, the likelihood of incomplete or careless processing of the response entry fields might increase with each additional item that is grouped on a survey page. Thus, satisficing in grids may to some extent offset the higher processing time that results from "response editing".

Besides fixation durations, it is also relevant to investigate survey responding that is related to satisficing, such as nondifferentiation (Couper et al. 2013; Roßmann et al. 2018; Zhang and Conrad 2014). In line with the assumption that presenting items together increases the likelihood that they are perceived and answered in the same context, previous research has found that respondents differentiated their answers more when

answering item-by-item and less when the items were presented in grids (e.g., Roßmann et al. 2018).

Hypothesis 5

"Answering items in grids, compared to item-by-item formats, leads to less differentiation (e.g., more straightlining) in the responses, across all target items."

To better understand the response process when grid questions are answered, we also analyzed additional indicators of respondent behavior. First, we examined whether respondents answered the items in grids sequentially. We defined responding as sequential, when a respondent read and answered the first item, then read and answered the second item, then the following item, and so on until the last item on the page. Conversely, non-sequential responding involves skipping items or going back to previous items while reading through the list (see Figure 3 for examples). We also observed if respondents read all or several items on a survey page before starting to answer and whether respondents changed their response to an item after having read other items. For these different response behaviors, we proceeded exploratively and did not postulate hypotheses.

3. Methods

3.1. Experimental Design

In this study, we implemented a question format experiment with three questions. Each of these questions was presented either as a single large grid question with ten items on one page, as two small grids with five items on each of the two pages, or in an item-by-item design, in which each of the ten items was presented on a separate page. The respondents were randomly assigned to one of the three formats for each of the three questions (see Table A.1 in Online Appendix A for details on respondents' sociodemographic characteristics per question). Further, the questions were either presented with a five-point response scale or a 11-point response scale with labels at the end points and numbers in between. As the length of the scale was not the focus of the present research, and as the randomization regarding the response scale length was independent of the randomization regarding question format, we combined the two response scales for the comparison across formats presented here (Question 1 "Trust in Institutions": $\chi^2 = .46$; df $= 2, p = .978$; Question 2 "People's rights: $\chi^2 = .062$; df $= 2$, $p = .970$; Question 3 "BFI-10": $\chi^2 = .110$; df $= 2$, $p = .947$; see Tables A.2 and A.3 in Online Appendix A for an overview of the main results by response scale length).

3.2. Survey Questions

To ensure comparability between questions, we implemented three questions with ten items each. In surveys, grid questions with up to ten items are often used, and this number of items can still be presented on one page of a personal computer without scrolling (see Toepoel et al. 2009). Also, ten items could easily be split up into two almost equally sized grids with five items on each page. We selected published scales that differed in item text length. The first set of items asked about "trust in institutions" (Question 1; GLES 2019). The items are very short and state different institutions such as the European Commission or the Federal Constitutional Court (see Online Appendix B for question wordings of the three sets of questions). The

second set of items asked about "people's rights in a democracy" (Question 2; ISSP Research Group 2016). The items consist of rather long sentences. The third set of items is the "BFI-10" (Question 3), a ten-item scale measuring the Big Five personality traits extraversion, agreeableness, conscientiousness, emotional stability, and openness (Rammstedt and John 2007). The items include complete sentences, but the statements are relatively short.

3.3. Participants

The study was conducted at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany, between April and May of 2017. We recruited 132 respondents from the respondent pool maintained by the institute or by word of mouth. An equal share of women and men was recruited, but no quotas for other demographics such as age and education were set. However, the intention was to obtain a sample as diverse as possible. Technical difficulties prevented recording of eye movements for one respondent, and in each of the questions the eye-tracking data of 13 to 18 respondents were of no satisfactory quality as we observed shifts between the text on the screen and the eye gaze data. These respondents were excluded from the analyses, leaving 103 respondents with good quality of recordings in all three experimental questions and 125 respondents with good recordings in at least one question. Of those 125 respondents, 51% were female; 38% were between 18 and 24 years, 27% between 25 and 34 years, 11% between 35 and 44 years, 10% between 45 and 54 years, 8% between 55 and 64 years, and 5% were 65 years or older; 6% had a school-leaving certificate from lower secondary education after 9 years of education ("Volks-/ Hauptschulabschluss" – ISCED Level 244), 22% from lower secondary education after ten years of education ("Mittlere Reife/Realschulabschluss"– ISCED Level 244), and 71% from upper secondary education providing access to tertiary education ("Fachhochschulreife/Allgemeine Hochschulreife" – ISCED Level 344) or tertiary/university education ("Universitätsabschluss" – ISCED Level 64 or 74). More than one third of the participants (38%) had participated in at least one web survey during the last three months. To evaluate the effectiveness of random assignment and the sample composition across conditions, we conducted several χ^2-tests for the reported sociodemographic characteristics mentioned previously. Except for sex in Question 3, no significant differences between sociodemographic characteristics were observed (see online Table A.1). To ensure that this does not affect our conclusions; we included sex as covariate in the analyses of response times, fixation durations, and response differentiation for Question 3.

3.4. Eye-Tracking Equipment and Procedures

We used the Senso Motoric Instruments (SMI) RED250 mobile Eye Tracker to record participants' eye movements and "BeGaze" version 3.6.57 for data analysis. The RED250 mobile Eye Tracker was mounted on the bottom frame of a 22″ TFT desktop monitor (resolution 1280x1024). The documentation of the RED250 mobile describes its accuracy to be within 0.4° and its tracking range of 32x21 at 60 centimeters distance. Eye movements were recorded at a sampling rate of 250 Hz. The online questionnaire was programmed with a font size of 16 pixels and double-spaced text with a line height of 40 and 32 pixels for the question text and response categories, respectively. The online questionnaire did not feature a "back" button.

Before the web survey started, respondents completed a calibration exercise (in which they followed black circles displayed at nine different points of the screen with their eyes). The questionnaire, which contained several experiments, took on average 30 minutes to complete. During this time, an experimenter stayed in the room next door to observe respondents' eye movements on a second computer screen for reasons of quality assurance. Participants were paid an incentive of EUR 20 for taking part in the study.

3.5. Measures and Analytical Strategies

We tested our hypotheses on the effects of question design on cognitive processing by analyzing indicators of cognitive effort measured by response times and eye-tracking data. For collecting *response times,* we used UCSP, Universal Client Side Paradata (Kaczmirek 2005). Response times were measured in milliseconds from the time a question appeared on the screen to the time respondents clicked on the next button to move on to the next question. For the small grid (two pages) and the item-by-item condition (ten pages), response times from the individual survey pages were summed up. Eye-tracking data provide information on the question answer process by recording where respondents look, for how long, and in what order while reading question stems, item texts, and response options (Galesic and Yan 2011; Romano Bergstrom and Schall 2014). Eye-tracking can be used as a proxy for depth of cognitive processing (Rayner 1998). The analysis of eye movements is based on two common assumptions (Just and Carpenter 1980; Rayner 1998). The first one, called the "immediacy assumption", states that objects fixated by the eyes are processed immediately (i.e., the mind follows the eye). The second one, called the "eye-mind assumption", states that the eye remains fixated on an object, as long as it is being processed (i.e., the eye follows the mind). Taken together, these two assumptions state that there is a close relationship between fixation duration and processing duration. A longer fixation duration indicates a longer response process. A long response process can be due to thorough consideration and recalling, but it can also indicate difficulties during the answer process. Those difficulties might arise from unknown or difficult terms, difficulties in arriving at an answer or selecting one of the response options (Galesic and Yan 2011; Kamoen et al. 2017; Neuert and Lenzner 2017). To measure "cognitive effort", we compare fixation durations on predefined areas of interest (AOI) to be able to compare these measures across the different question formats. Each question was conceptionally divided in five AOIs: (1) the complete question, (2) the question stem, (3) response scale labels, (4) item texts, and (5) response entry fields (see Figure 1). The AOIs on each individual page in the small grid and item-by-item format were summed up to compare fixation durations across formats. For response times and fixation durations, we excluded those respondents from the analyses who had response times below or above the mean plus/minus two standard deviations (see, e.g., Mayerl 2013).

To determine whether cognitive effort measured by response latencies and fixation durations were associated with the question format, we employed OLS regression models.

To study respondents' response behavior, we investigated how much respondents varied their answers to the items within the experimental questions. Nondifferentiation is found when respondents do not differentiate in their answers but give similar (or identical) responses to all items. The level of differentiation can be investigated by the probability of

differentiation P_d (Krosnick and Alwin 1998), which indicates the variability of the responses. P_d is calculated as $P_d = 1 - \sum_{i=1}^{n} P_i^2$, where P_i is the proportion of the values rated on a given point of a response scale and n is the number of rating points. If $P_d = 0$, respondents answered all items by selecting the same response, while a higher P_d means that different response options were given. We also measured the coefficient of variation (CV) as an indicator of the extremity of the responses (McCarty and Shrum 2000). CV is computed as $CV = \frac{s}{\bar{x}}$, where s is the standard deviation and $\bar{x}$ is the mean of the responses over items. The CV indicates the distance between the responses given. A CV of zero indicates straightlining response behavior, while larger values indicate that respondents differentiated their answers to a greater extent. As a measure of perceived difficulty, respondents were asked after each experimental question to rate how difficult answering the question was on a fully labeled five-point scale ranging from "extremely difficult" to "not difficult at all." Analyses were conducted using Stata version 16.1.

To further analyze respondents' behavior, two student assistants coded the eye gaze videos with regard to the following response patterns: (1) sequential responding, that is, did respondents answer grid questions in a sequential order; answering one item following the next?; (2) how many items did respondents read before answering the first item?; (3) answer change, that is, did respondents change their response after having read other items? Agreement between the two raters was 95% and Cohen's Kappa (1960) was found to be .87, which is "almost perfect", according to Landis and Koch's (1977, 165) criteria. To make those response patterns comparable between the small and large grid format, we summed the results for the two small grids so that both numbers are based on ten items for each question. For sequential responding, we included two measures for the small grid: answering both pages of the small grid sequentially and answering at least one of the two pages sequentially. As the response behavior is only comparable across grid questions, which present several items on the same survey page, these analyses were restricted to the two grid formats.

4. Results

4.1. Overall Cognitive Effort – Response Times and Question Fixation Durations

Before considering fixation durations on specific parts of the questions, we compared response times and fixation durations for the complete question as indicators of overall cognitive effort by question format. Response times were measured from the time a question appeared on the screen to the time respondents clicked on the next button to move on to the next question. Hence, anything that respondents did in between is included in this indicator. In contrast, fixation duration corresponds to the time a respondent spent fixating the question which might therefore be a more accurate measure of cognitive question processing (Just and Carpenter 1980; Staub and Rayner 2007). With regard to response times (in seconds) for Questions 1 ("Trust in Institutions") and 2 ("People's rights"), there was the general trend observable that respondents needed the least amount of time when the items were presented in a large grid (ten items per page) followed by small grids (five-items per page), and the most amount of time when the questions were presented item-by-item. However, the differences were not statistically significant (see Table 1). For Question 3 ("BFI-10"), our analysis showed that response times were significantly shorter

Table 1. Means and standard errors of cognitive effort indicators.

Cognitive effort indicators	Large grid with ten items (1 page)		Small grid with 5 items each (2 pages)		Item-by-item (10 pages)		Test	
	M	*(SE)*	*M*	*(SE)*	*M*	*(SE)*	*F value*	*p*
Response times								
Question 1	49.8	30.6	55.2	32.2	57.5	31.8	1.16	.204
Question 2	94.2	49.5	101.9	51.2	107.9	57.3	1.69	.189
Question 3	58.7[c]	28.7	54.9[c]	28.7	71.1[ab]	29.6	5.79	.001
Fixation durations – Complete question								
Question 1	38.3	25.8	37.8	27.6	44.9	27.6	2.06	.132
Question 2	73.2	42.1	78.7	43.7	79.7	49.8	0.62	.539
Question 3	46.6[c]	26.0	41.5[c]	26.7	54.4[ab]	26.9	3.76	.013
Fixation durations – Question stem								
Question 1	3.9[bc]	.62	7.5[a]	.62	6.7[a]	.65	9.45	.001
Question 2	8.2[bc]	.87	12.2[a]	.88	12.5[a]	1.03	7.02	.001
Question 3	1.4[c]	.27	1.9	.32	2.8[a]	.28	5.86	.001
Fixation durations – Item texts								
Question 1	11.2	1.2	10.7	1.4	8.9	1.5	1.46	.238
Question 2	35.1	2.4	34.5	2.5	37.4	2.6	.35	.703
Question 3	18.4	1.1	15.4	1.2	18.1	1.2	1.30	.279
Fixation durations – Response scale labels								
Question 1	3.8[c]	.47	4.5[c]	.49	7.7[ab]	.54	15.6	.001
Question 2	4.7[bc]	.55	6.4[ac]	.56	10.0[ab]	.66	19.1	.001
Question 3	5.9[c]	.92	6.9[c]	1.1	12.0[ab]	.97	8.03	.001
Fixation durations – Response entry fields								
Question 1	15.6	1.0	14.5	1.0	12.9	1.1	1.72	.184
Question 2	17.4[c]	.89	18.4[c]	.89	13.7[ab]	.96	7.02	.001
Question 3	19.2 [bc]	1.0	15.1[a]	1.1	13.3[a]	1.0	5.56	.001

Note: Question 1 = Trust in institutions, Question 2 = People's Rights, Question 3 = BFI-10. Reported are estimated marginal means from linear regression models. For Question 3, we report estimated marginal means controlling for sex. Superscripts present a significant difference ($p < .05$) compared to (a) large ten-item grid, (b) small five-item grids, or (c) item-by-item presentation. To compare the response times and fixation durations across formats, response times and fixation durations from the individual survey pages in the small grid and in the item-by-item design were summed up.

for the two grid designs ($M_{large} = 58.7$; $M_{small} = 54.9$) than for the item-by-item design ($M_{single} = 71.1$; $F = 5.79$; $p = .001$).

Comparing fixation durations on the complete question across designs showed a similar relationship. Fixation durations were slightly higher for the item-by-item presentation than for the grid formats, however, the differences were not statistically significant for Questions 1 and 2. For Question 3, we found statistically significant differences for both the small ($M_{small} = 41.5$) and the large grid ($M_{large} = 46.6$; $t = 2.06$, $p = .041$) compared to the item-by-item format ($M_{single} = 54.4$; $t = 3.31$, $p = .001$).

4.2. Question Fixation Durations for the Stage of Question Comprehension

Regarding the stage of question comprehension, our results showed that cognitive processing of the "question stem" was lowest for the large grid in all three experimental questions, followed by the presentation in two small grids with five items each, and the presentation in the item-by-item-design in Questions 2 ("People's rights") and 3 ("BFI-10"), as expected in Hypothesis 1 (see Table 1). For Question 1 ("Trust in institutions"),

fixation durations on the question stem were higher in the small grids than in the item-by-item format (although not significantly different).

Contrary to Hypothesis 2, which stated that the depth of processing of "item texts" is expected to decrease with the number of items presented on a survey page, we did not find any significant differences across designs.

4.3. Question Fixation Durations for the Stage of Reporting An Answer

Regarding the depth of processing of the "response scale labels", we found that fixation durations were significantly higher for the item-by-item format than for the presentation in both the large and the small grid in all three questions. This is in line with Hypothesis 3. However, the expected relation that depth of processing increases, the fewer items of the question are presented on a survey page, only holds true for Question 2 ("People's rights": $M_{large} = 4.7$; $M_{small} = 6.4$; $M_{single} = 10.0$; $F = 19.1$; $p = .001$). For Questions 1 ("Trust in institutions") and 3 ("BFI-10"), there were no statistically significant differences between the two grid designs.

Regarding the stage of reporting and response selection, we did not find that the fewer items presented on a survey page, the less time is spent processing the "response entry fields" as expected in Hypothesis 4. Respondents processed the response entry fields more extensively in the large grid design than in the item-by-item-design in Questions 2 ("People's rights") and 3 ("BFI-10"). For the small grids, findings were mixed. In Question 1 ("Trust in institutions"), we did not find statistically significant differences in fixation durations.

4.4. Observations of Response Behavior

We also investigated two indicators of satisficing, the probability of differentiation (P_d) and the coefficient of variation (CV). The results are shown in Table 2. Regarding the indicator of differentiation P_d, we did not find any significant differences across the three presentation formats.

Table 2. Indicators of satisficing (means)

Indicators of satisficing	Large grid with 10 items (1 page)		Small grid with 5 items each (2 pages)		Item-by-item (10 pages)		Test	
	M	*(SE)*	*M*	*(SE)*	*M*	*(SE)*	*F value*	*p*
P_d								
Question 1	.55	.030	.51	.031	.56	.031	0.94	0.39
Question 2	.63	.023	.61	.023	.66	.025	1.05	0.35
Question 3	.74	.016	.74	.016	.74	.016	0.32	0.81
CV								
Question 1	.22	.022	.22	.022	.21	.022	0.04	0.95
Question 2	.32	.016	.33	.017	.34	.019	0.19	0.83
Question 3	.46[c]	.019	.42	.019	.40[a]	.018	1.62	0.19

Note: Question 1 = Trust in institutions, Question 2 = People's rights, Question 3 = BFI-10. Reported are estimated marginal means from linear regression models. For Question 3, we report estimated marginal means controlling for sex. Superscripts present a significant difference ($p < .05$) compared to (a) large 10-item grid, (b) small 5-item grids, or (c) item-by-item presentation.

For the indicator CV, we also did not find significant differences across the three experimental conditions for Questions 1 ("Trust in institutions") and 2 ("People's rights"). For Question 3 ("BFI-10"), however, we observed that the variation of answers across the items was lower in the item-by-item design ($M_{single} = .40$) than in the large grid design ($M_{large} = .46$, $t = -2.07$, $p = .040$). Hence, Hypothesis 5, that answering items in grids, compared to item-by-item formats, leads to less differentiation in the responses cannot be confirmed in the current study. Figure 2 illustrates two different response styles, with the respondent on the left-hand side showing no differentiation in the responses selected (straightlining) while the respondent on the right-hand side differentiated the responses to a greater extent. Observing the gaze videos provided the interesting finding that respondents may actually spend more cognitive effort than one might initially expect when showing a straightlining response behavior.

Finally, we analyzed how respondents answered the items presented in the two grid formats to gain more knowledge about response patterns when answering grid question formats. For the small grid, we report the number of respondents who answered the ten items displayed on two pages sequentially, and the number of respondents who answered at least one page with five items in a sequential order (see Table 3). Comparing sequential responding for all ten items, we did not observe differences between the large and the small grid for Questions 1 ("Trust in institutions": 37% large vs. 29% small) and Question 2 ("People's rights": 42% large vs. 40% small). However, for Question 3 ("BFI-10"), we found that half of the respondents in the large grid (51%) answered the items sequentially, while only 12% in the small grid condition did show this response behavior. When considering those respondents in the small grid condition who answered at least one page with five items in a sequential order, we observed that this amount is quite large, between 57% and 68%, but no statistically significant differences compared to the large grids were found.

Figure 3 illustrates both a sequential as well as a non-sequential response style. As shown by the eye movement patterns, the respondents on the left-hand side read and answered the items sequentially one by one; the respondents on the right-hand side first read all item texts and then moved to the response options and answered all items one after another or jumped back and forth on the question parts.

We also investigated whether respondents perceived the items in a grid as belonging conceptually together. The response behavior that respondents read all items before beginning to answer them only occurred in Question 1 ("Trust in institutions": 16% large vs. 41% small). In Questions 2 and 3, respondents did not read through all items before selecting a response for the first item. Comparing how many items respondents read on average before they started answering, revealed that respondents read on average between one and three items in the large grid and between two and four items in the small grid condition (Question 1 "Trust in institutions": 3.2 large vs. 4.4 small; Question 2 "People's rights": 1.2 large vs. 2.4 small; Question 3 "BFI-10": 1.1 large vs. small 2.5 small). It must be noted that the average number of items read on the two separate pages of the small grids were summed up, resulting in a systematic overestimation compared to the large grid conditions.

Interestingly, the overall number of items read before beginning to answer is higher in the question asking for "trust in institutions" (Question 1) than in the other two questions. Notably, the question asking for "trust in institutions" had the shortest item text regarding the number of characters but was perceived as most difficult by respondents ($M = 2.11$)

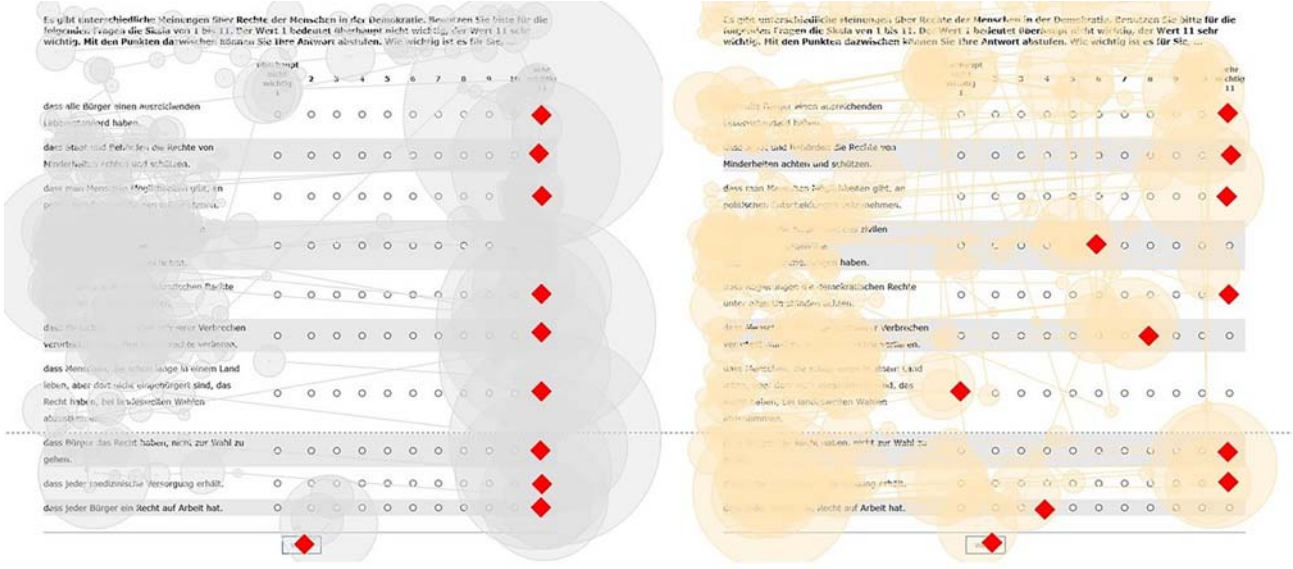


Fig. 2. Gaze Plots of two different respondents (large grid condition) showing either straightlining behavior or not.

Note: The figure shows the gaze plots of two different respondents answering Question 2. Each gaze plot displays the eye movements of one respondent. The circles indicate fixations, and the lines between the circles indicate saccades. The size of the circle is proportional to the fixation time, meaning larger circles indicate longer fixations. The squares indicate the mouse clicks of the respondents. The respondent on the left-hand side showed straightlining behavior; the respondent on the right-hand side differentiated responses to a greater extent.

Table 3. Response behavior for grid formats, by question and grid size.

Question	1 – Trust in Institutions		2 – People's rights		3 – BFI-10	
	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)	Large grid with 10 items (1 page)	Small grid with 5 items each (2 pages)
% Sequential responding						
– Yes, all 10 items	37.2 (16)	28.6 (12)	42.2 (19)	39.5 (17)	51.2 (22)	12.2 (5)
	χ^2=.718; df=1		χ^2=.066; df=1		χ^2=14.61; df=1**	
– Yes, at least on 1 page with 5 items	37.2 (16)	57.1 (24)	42.2 (19)	60.5 (26)	51.2 (22)	68.3 (28)
	χ^2=3.39; df=1		χ^2=2.93; df=1		χ^2=2.56; df=1	
Mean number of items read before beginning to answer						
	3.2 (43)	4.4 (42)	1.2 (45)	2.4 (43)	1.1 (43)	2.5 (41)
	$F(1,83)$=.412		$F(1,86)$=6.011**		$F(1,82)$=27.083**	
% All items read before beginning to answer						
– No	83.7 (36)	59.5 (25)	100 (45)	97.7 (42)	100. (43)	97.6 (40)
– Yes	16.3 (7)	40.5 (17)	0 (0)	2.3 (1)	0 (0)	2.4 (1)
	χ^2=6.139*; df=1		χ^2=1.059; df=1		χ^2=1.061; df=1	
% Answer change						
– No	72.1 (31)	78.6 (33)	71.1 (32)	81.4 (35)	81.4 (35)	87.8 (36)
– Yes	27.9 (12)	21.4 (9)	28.9 (13)	18.6 (8)	18.6 (8)	12.2 (5)
	χ^2=.479; df=1		χ^2=1.280; df=1		χ^2=.659; df=1	
n	43	42	45	43	43	41

Note: * $p < .05$; ** $p < .01$; Parenthetical entries are cell sizes. For sequential responding, we report both the comparison of the large grid with (1) reading all ten items of the small grids (summed up) and with (2) answering at least one page of two sequentially.

compared to Question 2 (M = 1.54, $p < .001$), and Question 3 (M = 0.81, $p < .001$). This finding can be interpreted as an indication that respondents used the information from the item texts of several other items to answer each item of the grid question.

Finally, we observed whether respondents tended to change their answers given to one item after having read other items in the grid. Across all three experimental questions between 16% and 25% of respondents changed at least one answer after having read the following items in the related grid (Question 1 "Trust in institutions": small 21% vs. 28% large; Question 2 "People's rights": small 19% vs. 29% large; Question 3 "BFI-10": small 12% vs. 19% large). Together with the findings on sequential reading and reading the item text of several other items before starting to answer, this might indicate that respondents used the additional information provided by the remaining items in a grid to give a response by applying the "near means related" heuristic (Tourangeau et al. 2004).

5. Discussion

5.1. Summary and Discussion of Findings

This study investigated the depth of cognitive processing when answering different grid question or item-by-item formats. We implemented three questions with ten items each in an experiment and tracked respondents' eye movements while they answered a web survey in

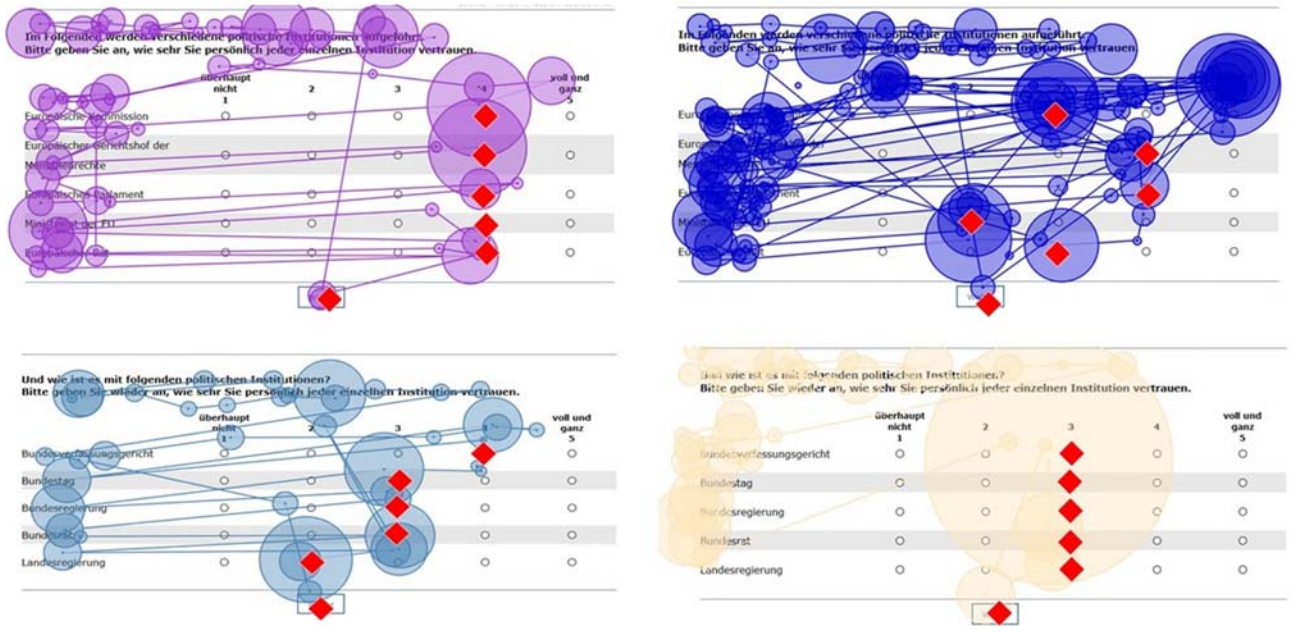


Fig. 3. Gaze Plots of different respondents (small grid condition) representing sequential (left-hand side) and non-sequential response styles (right-hand side).
Note: The figure shows the gaze plots of four different respondents answering Question 1. Each gaze plot displays the eye movements of one respondent. The circles indicate fixations, and the lines between the circles indicate saccades. The size of the circle is proportional to the fixation time, meaning larger circles indicate longer fixations. The two respondents on the left-hand side read and answered the items sequentially one by one; the respondents on the right-hand side showed non-sequential responding: above in the form of jumping back and forth on the question parts; below in the form of first reading all item texts and then moving to the response options and answering them in a row.

which the presentation was varied. For each question, we randomly assigned respondents to three question formats (item-by-item, small grid, or large grid). The eye-tracking data studied showed that the previous finding of longer response times (e.g., Callegaro et al. 2009; Couper et al. 2001; Roßmann et al. 2018; Toepoel et al. 2009) in the item-by-item format could be attributed to more extensive processing of the question stem and the response scale labels compared to the grid formats (Hypothesis 1 and 3). In contrast, we did not find differences with respect to processing of the item texts (Hypothesis 2). This indicates that respondents do not spend more time on the item texts in either format but need time to adjust to the new question context when the items are presented in an item-by-item format. Specifically, they have to read the same question stem and response scale labels multiple times, since they are presented to them with each item. Those findings suggest that the item-by-item format increases response burden compared to the grid formats.

With regard to the response process stages of reporting and response selection (Tourangeau et al. 2000), we observed that fixation durations on the area of the response entry fields were significantly longer in the large grid than in the item-by-item presentation in two out of three questions (Hypothesis 4). Since those two questions had longer item texts, one possible explanation might be that navigating within a grid is more difficult for long items than when the items are presented item-by-item on separate pages. Hence, selecting and reporting a response seems less burdensome for respondents in the item-by-item design.

By using eye-tracking methodology, we were able to observe the behavior of respondents more directly while they were answering the grid questions. These analyses suggest that respondents apply the "near means related" heuristic (Tourangeau et al. 2004; Silber et al. 2018), which is grounded in the proximity principle from Gestalt psychology (Koffka 1935; Wertheimer 1923). According to the principle of proximity, placing objects close to each other will let them be perceived as a group, and hence as not only physically but also conceptually related (Dillman et al. 2014). Consequently, items presented in the grid format were likely perceived, processed, and answered in the same context. Presenting multiple items together on a page can facilitate respondents' cognitive processing. If the respondent is not familiar with the topic or when the meaning of the question is not clear, respondents might try to capture the content using the surrounding items to improve their understanding of the question (Krosnick and Presser 2010).

Consistent with the "near means related" heuristic, many respondents in our survey did not respond to the grid questions sequentially but instead read multiple items before answering the first item. They also changed their answers later after reading other questions, suggesting that they reconsidered their answers after answering other items. Such response behavior was more pronounced for items presented in one large grid than for items presented in two small grids but was also visible there. For related items, grouping them may improve measurement (Krosnick and Presser 2010), for example, by increasing the consistency of responses among items and inter-item correlations (Couper 2008; Heerwegh 2009; Toepoel et al. 2009). In contrast, the grouping may also have negative effects on measurement. Although we did not observe differences regarding non-differentiation in our study (Hypothesis 5), previous research has consistently shown that a separate presentation like in an item-by-item design reduced undesired response effects, such as non-differentiation or item nonresponse (e.g., Roßmann et al. 2018; Toepoel et al.

2009). Also, higher inter-item correlations may be due to measurement error (Peytchev 2005, cited in Couper et al. 2013).

Across the three questions, we also observed some differences regarding response behavior, which might be related to the content of those questions. While the item texts of Question 1 only featured the names of institutions, the item texts of Questions 2 and 3 consisted of full sentences. The shorter item texts of Question 1 may have led more respondents to read multiple or even all item texts before answering the first item. Since our study did not experimentally vary the item length, further research is needed to understand the relationship between response behavior and item length in grid formats.

With respect to response burden, we found that both the item-by-item format and the grid format entail different burdensome elements. The item-by-item format requires respondents to adjust to a new context for each item. Specifically, presenting the question stem and response scale labels on each page makes respondents undergo repetitive reading tasks, which lowers response efficiency. In contrast, the grid format increases the complexity because respondents are confronted with multiple items on a single survey page. This may increase response burden due to navigational difficulties in the process of reporting the responses. With respect to response behavior, this study showed that respondents do not answer the items necessarily from top to bottom and in the presented order. Instead, some respondents read several items before beginning to answer the first item. Some respondents also changed their responses after they had read the following items. The extent of this behavior depended on the item texts of the questions. Finally, we observed that some respondents selected the identical responses for all items in a grid (i.e., straightlining) but still read the item texts attentively. With respect to strong satisficing, we expect that respondents skip the question comprehension stage altogether and provide responses haphazardly. Yet, the observation in the current study also fosters the notion of weaker forms of satisficing, in which respondents attend to the question stem and item text, but then decide to simplify the perceivably difficult task of reporting accurate and meaningful answers, for instance, by resorting to the "near means related" heuristic (Tourangeau et al. 2004; Silber et al. 2018). Hence, it might be worthwhile to investigate this and similar response behaviors in more depth with respect to how respondents arrive at selecting the same answer to all questions of an item sequence. Due to the low number of straightlining respondents in the lab setting, this study did not allow us to investigate this pattern further.

5.2. Limitations and Avenues for Future Research

Our study has several limitations. The most important is external validity since we designed our investigation as a lab study. Filling-out a questionnaire in a lab situation, in which the eye movements of respondents are recorded, may not perfectly reflect the behavior of respondents in a common survey interview environment. Also, the participants might have been rather engaged as they were willing to participate in a lab experiment and received an incentive of EUR 20 for their participation. Hence, the differences in cognitive processing might be less pronounced than in studies conducted in common survey settings. For instance, we did not find differences with respect to nondifferentiation, even though many previous studies have shown such differences (e.g., Mavletova et al. 2018; Roßmann et al. 2018; Tourangeau et al. 2004). Likewise, fixation durations may have been overestimated, and the number of respondents engaging in response behaviors such as

non-sequential responding or answer changes underestimated. Yet, those or similar response behaviors can be expected to occur outside the lab as well, since previous research found answer patterns that were likely due to answering items in the same context when they were presented in a grid (e.g., Couper et al. 2001; Toepoel et al. 2009; Tourangeau et al. 2004). Second, also attributable to being a lab study, is the comparatively smaller sample size than, for instance, in many online experiments. However, recruiting and testing more participants in an eye-tracking study would be laborious and expensive. A third limitation that should be addressed by additional research is that the scales we used were all endpoint-labeled with numbers in between. How the processing of the scales differs across grid formats when using fully labeled scales would be worthwhile examining in a follow-up study. Also, investigating the generalizability of our findings with a different number of items, such as eight or six items for the large grid and four or three for the small grid, is an avenue for future research. Fourth, we decided to place the two grids and every single item on separate pages in a so-called paging design. Future studies could explore whether similar results are obtained if they are presented on the same page in a scrolling design (see, e.g., Liu and Cernat 2018). A possible outcome of using the scrolling design could be that the items in the single item or the smaller grid formats might be more often answered in the same context due to the visual presentation on the same page. Finally, as the questions in our study were answered on a desktop PC screen only due to the eye-tracking system used, this study does not address the issue of responding on mobile devices, which seems to be another worthy avenue for future research.

5.3. Increasing Relevance of Mobile Devices in Web Surveys

Given the increasing number of respondents using smartphones or other mobile devices in answering web surveys (Gummer et al. 2019), design decisions on using grid versus item-by-item presentation have become increasingly important. This is especially true in the context of decisions regarding whether to use layouts that adapt to the device used by respondents (adaptive or responsive layouts) or to optimize layouts for use on a specific device (e.g., mobile first layouts). When grids are presented at full size on the small(er) screens of smartphones, this may require horizontal scrolling and zooming. Previous research has found that answering grids on smartphones compared to grids on personal computers increases breakoff rates and stimulates undesired response behaviors like straightlining (see Antoun et al. 2018 for a systematic review). In mobile-first unified designs and responsive designs, survey software often automatically adapts grids to screen size by converting them into a series of single items. Thus, some respondents will see the set of items as grids if they use a personal desktop or laptop computer, while other respondents will see them in an item-by-item format. This might result in systematic mode differences and measurement error. The same might apply to mixed-mode surveys, for example, paper versus web questionnaires (De Leeuw et al. 2018; Dillman et al. 2014). Respondents in our study answered grids solely on a desktop PC. Thus, we suggest that future studies could use eye-tracking methodology to investigate how adaptive (or responsive) layouts impact cognitive processing and response behavior in web surveys with multiple devices, and in particular, the suitability of grids on mobile devices.

5.4. Recommendations

Our findings have practical implications for researchers deciding between grids or item-by-item designs. In our study, both designs appear to come with format-specific limitations, which directly affect response burden. To ensure that the difference in question presentation format between respondents does not lead to potential measurement error, especially if smartphone participation in web surveys continues to rise, we recommend using the smallest screen as the basis of the format decision, which is in favor of the item-by-item design (see also Antoun et al. 2018; Liu and Cernat 2018; Mavletova et al. 2018). Another argument for the item-by-item design is that items in grid questions do not meet web content accessibility guidelines (WCAG; W3C 2018), which suggests that each question should be entirely understandable on its own. Yet, if a consistent use of item-by-item presentation is not possible, for instance due to restrictions in the available questionnaire length, we would recommend to break up larger grids into (as in our case two) smaller grids (see also Dillman et al. 2014) as they did not show substantial disadvantages compared to a large grid question, and they seem to be easier to navigate (this is in line with Grady et al. (2019) who recommend a small to medium grid size). Though, for some surveys, grid formats might be the best alternative (e.g., brand image research; Brosnan et al. 2021). This decision may depend on factors such as the question type, the complexity of the information, and the question content. For example, grid questions may help respondents quickly understand their response task for multiple items at once and thereby increase response efficiency.

6. Conclusion

This study showed that it takes respondents longer to answer a question in the item-by-item format than in the grid format because the former shows the question stem and the response scale repetitively, and respondents need to process both multiple times. The differences in the visual presentation and the shorter response times of grid questions did not result in more satisficing response behavior than in the item-by-item format, which might have been due to the lab setting in which participants are likely to be quite engaged. Finally, by using eye tracking, we were able to observe specific response styles (i.e., reading a few items before answering a grid and answer changes) when a question was presented in the grid format. An area for future research would be to investigate whether items presented in a grid format are more likely to be processed and answered in one context than when presented in an item-by-item format within a scrolling design, and whether these differences in cognitive processing and responding have a substantial impact on substantive analyses with the items is an area for future research.

7. References

Antoun, C., J. Katz, J. Argueta, and L. Wang. 2018. “Design Heuristics for effective Smartphone Questionnaires.” *Social Science Computer Review* 36(5): 557–574. DOI: https://doi.org/10.1177/0894439317727072.

Brosnan, K., B. Grün, and S. Dolnicar. 2021. "Cognitive load reduction strategies in questionnaire design." *International Journal of Market Research* 63(2): 125–133. DOI: https://doi.org/10.1177%2F1470785320986797.

Callegaro, M., J. Shand-Lubbers, and J.M. Dennis. 2009 "Presentation of a Single Item versus a Grid: Effects on the Vitality and Mental Health Scales of the SF-36v2 Health Survey." 64th Annual Conference of the American Association for Public Opinion Research (AAPOR), May 14, 2009: 5887–5897. Hollywood, Florida. Available at: http://www.asasrms.org/Proceedings/y2009/Files/400045.pdf (accessed July 2022).

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and psychological measurement* 20(1): 37–46. DOI: https://doi.org/10.1177/001316446002000104.

Couper, M.P. 2008. *Designing effective Web surveys*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511499371.

Couper, M.P., R. Tourangeau, F.G. Conrad, F.C. Zhang. 2013. "The Design of Grids in Web Surveys." *Social Science Computer Review* 31(3): 322–345. DOI: https://doi.org/10.1177/0894439312469865.

Couper, M.P., M.W. Traugott, and M.J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly*: 65(2): 230–253. DOI: https://doi.org/10.1086/322199.

DeBell, M., C. Wilson, S. Jackman, and L. Figueroa. 2021. "Optimal Response Formats for Online Surveys: Branch, Grid, or Single Item?" *Journal of Survey Statistics and Methodology* 9(1): 1–24. DOI: https://doi.org/10.1093/jssam/smz039.

De Leeuw, E.D., Z.T. Suzer-Gurtekin, and J.J. Hox. 2018. "The Design and Implementation of Mixed-mode Surveys." In *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, edited by T.P. Johnson, B. Pennell, I.A.L. Stoop, and B. Dorer: 387–409. Hoboken: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-mode Surveys: the Tailored Design Method* (4th edition). Hoboken: Wiley.

Galesic, M., and T. Yan. 2011. "Use of Eye Tracking for Studying Survey Response Processes." In *Social and Behavioral Research and the Internet*, edited by M. Das, P. Ester, and L. Kaczmirek: 349–370. New York: Taylor and Francis.

GLES. 2019. "Longterm-Online-Tracking, Cumulation 2009–2017 (GLES)." GESIS Data Archive, Cologne. ZA6832 Data file Version 1.1.0, DOI: https://doi.org/10.4232/1.13416.

Grady, R.H., R.L. Greenspan, and M. Liu. 2019. "What Is the Best Size for Matrix-Style Questions in Online Surveys?" *Social Science Computer Review* 37(3): 435–445. DOI: https://doi.org/10.1177/0894439318773733.

Gummer, T., F. Quoß, and J. Roßmann. 2019. "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?" *Social Science Computer Review* 37(3): 371–384. DOI: https://doi.org/10.1177/0894439318766836.

Heerwegh, D. 2009. "Mode Differences Between Face-to-Face and Web Surveys: an Experimental Investigation of Data Quality and Social Desirability Effects." *International Journal of Public Opinion Research* 21(1): 111–121. DOI: https://doi.org/10.1093/ijpor/edn054.

ISSP Research Group (2016): International Social Survey Programme: Citizenship II – ISSP 2014. GESIS Data Archive, Cologne. ZA6670 Data file Version 2.0.0. DOI: https://doi.org/10.4232/1.12590.

Jenkins C.R, and D.A. Dillman. 1995. "Towards a Theory of Self-Administered Questionnaire Design." In *Survey Measurement and Process Quality,* edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin: 165–196. New York Wiley.

Just, M.A., and P.A. Carpenter. 1980. "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological Review* 87: 329–354. DOI: https://doi.org/10.1037/0033-295X.87.4.329.

Kaczmirek, L. 2005. "A Framework for the Collection of Universal Client Side Paradata (UCSP)." Available at: http://kaczmirek.de/ucsp/ucsp.html (accessed January 2021).

Kamoen, N., B. Holleman, P. Mak, T. Sanders, and H. van den Bergh. 2017. "Why are Negative Questions Difficult to Answer? On the Processing of Linguistic Contrasts in Surveys." *Public Opinion Quarterly* 81(3): 613–635. DOI: https://doi.org/10.1093/poq/nfx010

Koffka, K. 1935. *Principles of Gestalt psychology*. New York: Harcourt.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–236. DOI: https://doi.org/10.1002/acp. 2350050305.

Krosnick, J.A., and D.F. Alwin. 1988. "A Test of the Form-Resistant Correlation Hypothesis. Ratings, Rankings, and the Measurement of Values." *Public Opinion Quarterly* 52(4): 526–538. DOI: https://doi.org/10.1086/269128.

Krosnick, J.A., and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by P.V. Marsden, and J.D. Wright: 263–314. Emerald Group Publishing.

Landis, J.R., and G.G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1): 159–174. DOI: https://doi.org/10.2307/2529310.

Liu, M., and A. Cernat. 2018. "Item-by-Item versus Matrix Questions: A Web Survey Experiment." *Social Science Computer Review* 36(6): 690–706. DOI: https://doi.org10.1177/0894439316674459.

Mavletova, A., and M.P. Couper. 2015. "A Meta-Analysis of Breakoff Rates in Mobile Web Surveys." In *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, edited by D. Toninelli, R. Pinter, and P. de Pedraza: 81–98. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.f.

Mavletova, A., M.P Couper, and D. Lebedev. 2018. "Grid and item-by-item formats in PC and mobile web surveys." *Social Science Computer Review* 36(6): 647–668. DOI: https://doi.org/10.1177%2F0894439317735307.

Mayerl, J. 2013. "Response Latency Measurement in Surveys. Detecting Strong Attitudes and Response Effects". *Survey Methods: Insights from the Field*. Retrieved December 17, 2020, from https://surveyinsights.org/?p = 1063. DOI: https://doi.org/10.13094/SMIF-2013-00005.

McCarty, J.A., and L.J. Shrum. 2000. "The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures." *Public Opinion Quarterly* 64(3): 271–298. DOI: https://doi.org/10.1086/317989.

Neuert, C.E., and T. Lenzner. 2017. "Incorporating Eye Tracking into Cognitive Interviewing to Pretest Survey Questions." *International Journal of Social Research Methodology* 19(5): 501–519. DOI: https://doi.org/10.1080/13645579.2015.1049448.

Peytchev, A. 2005. "How Questionnaire Layout Induces Measurement Error." Paper presented at the 60th annual meeting of the American Association for Public Opinion Research, May, 2005. Miami Beach, FL,USA. Available at: http://www.websm.org/db/12/3636/Bibliography/Causes%20of%20Context%20Effects:%20How%20Questionnaire%20Layout%20Induces%20Measurement%20Error/.

Rammstedt, B., and O.P. John. 2007. "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German." *Journal of Research in Personality* 41: 203–212. DOI: https://doi.org/10.1016/j.jrp. 2006.02.001.

Rayner, K. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." Psychological Bulletin 124: 372–422. DOI: https://doi.org/10.1037/0033-2909.124.3.372.

Romano Bergstrom, J., and A. Schall. 2014. *Eye Tracking in User Experience Design*. San Francisco, CA: Morgan Kaufmann.

Roßmann, J., T. Gummer, and H. Silber. 2018. "Mitigating Satisficing in Cognitively Demanding Grid Questions: Evidence from Two Web-Based Experiments." *Journal of Survey Statistics and Methodology* 6(3): 376–400. DOI: https://doi.org/10.1093/jssam/smx020.

Roßmann, J., and H. Silber. 2020. "Satisficing and Measurement Error." In *SAGE Research Methods Foundations*, edited by P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, and R.A. Williams. London: SAGE Publications. DOI: https://dx.doi.org/10.4135/9781526421036912794.

Silber, H., J. Roßmann, and T. Gummer. 2018. "When Near Means Related: Evidence from Three Web Survey Experiments on Inter-Item Correlations in Grid Questions." *International Journal of Social Research Methodology* 21(3): 275–288. DOI: https://doi.org/10.1080/13645579.2017.1381478.

Staub, A., and K. Rayner. 2007. "Eye movements and on-line comprehension processes." In *The Oxford Handbook of Psycholinguistics*, edited by G. Gaskell: 327–342. Oxford, UK: Oxford University Press.

Toepoel, V., M. Das, and A. van Soest. 2009. "Design of Web Questionnaires: The Effects of the Number of Items per Screen." *Field Methods* 21(2): 200–213. DOI: https://doi.org/10.1177/1525822X08330261.

Tourangeau R., M.P. Couper, and F. Conrad. 2004. "Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68(3): 368–393. DOI: https://doi.org/10.1093/poq/nfh035.

Tourangeau, R., and K. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103(3): 299–314. DOI: 10.1037/0033-2909.103.3.299.

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

Wertheimer, M. 1923. *Laws of organization in perceptual forms: A source book of Gestalt psychology*. London: Routledge.

W3C: World Wide Web consortium. 2018. "Web Content Accessibility Guidelines (WCAG) 2.1." Available at: https://www.w3.org/TR/WCAG21/ (accessed June 2021).

Zhang, C., and F. Conrad. 2014. "Speeding in Web Surveys: The Tendency to Answer very Fast and its Association with Straightlining." *Survey Research Methods* 8(2): 127–135. DOI: https://doi.org/10.18148/srm/2014.v8i2.5453.

Received January 2021
Revised June 2021
Accepted August 2022

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 103–121, http://dx.doi.org/10.2478/JOS-2023-0005

A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy

Andrew M. Raim[1], *Elizabeth Nichols*[1], *and Thomas Mathew*[1]

For operations such as a decennial census, the U.S. Census Bureau sends mail to potential respondents inviting a self-response. It is suspected that the mailing strategy affects the distribution of call volumes to the U.S. Census Bureau's telephone helplines. For staffing purposes, more uniform call volumes throughout the week are desirable. In this work, we formulate tests and confidence intervals to compare uniformity of call volumes resulting from competing mailing strategies. Regarding the data as multinomial observations, we compare pairs of call volume observations to determine whether one mailing strategy has multinomial cell probabilities closer to the uniform probability vector compared to another strategy. A motivating illustration is provided by call volume data recorded in three studies which were carried out in advance of the 2020 Decennial Census.

Key words: Entropy; multinomial distribution; self-response; uniform probability vector.

1. Introduction

1.1. Background

Beginning with the 1990 Census, a telephone questionnaire assistance operation has accompanied each decennial census. These operations included helpline agents representing the U.S. Census Bureau who fielded support questions and assisted the public in completing paper forms. An automated interactive voice response system was added after the 1990 Census to augment live agents. Since the 2000 Census, agents were additionally able to conduct interviews and collect census data by phone, although this has not been marketed as a response option. For the 2020 Census, the U.S. Census Bureau mailed letters and postcards to residential addresses in the country to request participation in the census, and encouraged responses on a large scale through the internet (U.S. Census Bureau 2017). Telephone helplines were highlighted in mailings, both as a means to assist with questions about the census and to serve as a mode of response in itself. From the perspective of the U.S. Census Bureau, an ideal distribution of helpline calls would be where a uniform volume of calls is received throughout the week for the duration of the

[1] U.S. Census Bureau, Washington, D.C., 20233, U.S.A. Emails: andrew.raim@census.gov, elizabeth.may.nichols@census.gov, and thomas.mathew@census.gov.

Acknowledgments: We thank Sarah Konya and Julia Coombs for providing the call data for the experiments studied in this work. Thanks to Darcy Steeg Morris, Julia Coombs, Kevin Zajac, Tommy Wright, Eric Slud, and the Decennial Research Objective and Methods (DROM) committee for their feedback on this report. Helpful comments from the reviewers and an Associate Editor are gratefully acknowledged. This article is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau. The article has been reviewed for disclosure avoidance and approved under CBDRB-FY21-CBSM001-004.

operation. More uniform call volumes would provide a more constant workload for helpline staff over the course of the operation. Furthermore, a reduction of large peaks in volume may reduce the need for increased staff.

The schedule of mailings influences when calls tend to occur and is an aspect of census design which U.S. Census Bureau can control. Chesnut (2003) and Zajac (2012) discuss call volumes received during the 2000 and 2010 Decennial Census, respectively, and note how they were affected by the mailing schedule. Helplines received 7.9 million calls in the 1990 Census, six million calls in the 2000 Census, and 4.5 million calls in the 2010 Census. Over 13 million calls were received by the system used in 2020. The interactive voice response system deflected most of them, but nearly five million incoming calls were handled by customer service representatives. Although volumes of calls have changed over the decades, patterns of calls to the helpline have not changed drastically (Nichols et al. 2019). Higher call volumes occur on the expected delivery date of mailed census notification letters and postcards. The first peaks occur after the initial mailout and second mailout, and another peak occurs the week of Census Day. There is also a trend in which Mondays and Tuesdays are the highest call volume days, with a gradual decline in call volume throughout the week and a large dropoff over the weekend. This pattern is more exaggerated when mail arrives on a Monday, as was the case in the 2000 Census (Chesnut 2003). The volume of calls diminishes after both Census Day and the arrival of all mailed notifications have occurred. Expecting similar patterns with increased volume for the 2020 Census, the U.S. Census Bureau adopted staggered mailings of census notifications to spread delivery to multiple days of the week and multiple weeks of the month, anticipating that daily calls to the helpline would be more uniform and thus easier to staff efficiently (Nichols et al. 2019). Note that other aspects of a mailing schedule–such as potential impact to response rates – are important to the U.S. Census Bureau as well; however, the remainder of this article focuses on call uniformity.

The U.S. Census Bureau conducts operations throughout the decade to prepare for the decennial census. Several studies leading up to the 2020 Census featured mailings with an invitation to respond via internet where data on subsequent calls to helplines were recorded. These data provide an opportunity to compare variations between mailing strategies and the uniformity of the resulting call volumes. Assuming that the data are realizations from a multinomial distribution, in this article we compare pairs of call volume observations and develop a methodology to assess whether one mailing strategy provides multinomial cell probabilities closer to the uniform probability vector compared to another mailing strategy.

1.2. On the Kullback-Leibler Divergence

To our knowledge, inference comparing the closeness of two discrete distributions to a discrete uniform distribution is not standard. Many conventional tests are primarily designed to detect departure from equality of two distributions. However, the equality of two distributions is not the primary interest in our application; our problem is that of comparing two sets of multinomial cell probabilities in terms of their relative closeness to the uniform probability vector, and it seems appropriate to consider statistics used for assessing goodness of fit. While several empirical distribution based statistics are available

for this (for example, the Kolmogorov-Smirnov statistic; see Li et al. 2005), a statistic based on the Kullback-Leiber divergence is an option that has been investigated in the literature. As is well known, the K-L divergence measure is a directed divergence measure that is not symmetric with respect to the two distributions (see Li et al. 2005; Cover and Thomas 2006). This is appropriate for our purpose since our focus is on a measure of closeness directed towards the uniformity of multinomial cell probabilities.

In a book on information theory, Cover and Thomas (2006) discuss the K-L distance, entropy, and related concepts, and discuss fundamental applications in information theory. Dorfinger et al. (2011) use entropy as a measure of uniformity to classify in real time whether traffic in computer networks is encrypted or not. Their approach makes a decision based on the difference between the estimated entropy of an observed payload and that of a uniformly distributed random payload of the same length. Liu and Wang (2004) and Cohen et al. (2006) consider an increasing convex ordering among discrete distributions; when this ordering holds, one particular consequence is that one of the distributions has a larger entropy than the other and is therefore closer to uniform. K-L divergence and entropy have found use in many areas of the statistics literature, including: to justify information criteria in assessing model fits (Konishi and Kitagawa 2008), to obtain variational approximations to complicated distributions such as the posterior in Bayesian analysis (Ormerod and Wand 2010; Blei et al. 2017), and as a basis for statistical inference (Pardo 2006; Girardin and Lequesne 2019). Paninski (2008) proposed a method to test whether a single multinomial distribution departs from discrete uniform; this work is based on a sparse setting with many categories and relatively few observations.

From the perspective of testing goodness-of-fit, Several authors have investigated the K-L divergence measure and its properties. The article by Girardin and Lequesne (2019) gives a review of the literature on entropy-based goodness-of-fit tests. A test that is asymptotically distribution-free is investigated in Song (2002). Regarding the K-L divergence measure based test, Song (2002) notes that "With its good power properties, the method provides an extremely simple and potentially much better alternative to the classical empirical distribution function (EDF)-based test procedures". A Monte Carlo simulation study comparing several goodness of fit tests is reported in Li et al. (2005). Along with a test based on K-L divergence, the tests considered by these authors include those based on the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling statistics. The authors note that "... the tests based on the directed divergence measure give a good approximation to the given significance levels and are more powerful than other tests against the given alternative distributions". In an earlier article comparing models that are relevant in economics, Vuong (1989) used the K-L divergence criterion to measure the closeness of a model to the true model. The author has developed likelihood-ratio based statistics for testing the null hypothesis that two competing models are equally close to the true model; the alternative hypothesis being that one model is closer. Our problem is similar to what Vuong (1989) has investigated. We believe that these articles provide a strong motivation for using the K-L divergence measure to address the problem we have investigated.

The rest of the article proceeds as follows. Section 2 introduces the call volume data and provides some exploratory analysis. Section 3 discusses test and confidence interval procedures based on the K-L divergence criterion in order to compare the closeness of two discrete

distributions to uniformity. Section 4 gives results of the data analysis. Finally, Section 5 concludes the article. Additional technical details of the statistical procedures are given in the appendices. Simulation studies to validate the procedures are provided as a online supplement.

2. Call Volume Data

A standard mailing strategy would be to send mail to all intended recipients on a common schedule. Alternatively, consider a staggered strategy where respondents are randomly partitioned into several groups which are sent mailings on different schedules. It is thought that compared to an unstaggered strategy, a staggered strategy leads to a more uniform distribution of calls, in the sense that the probability of a mail recipient calling the helpline on any day of the week is closer to being uniform (i.e., equal to 1/7). To explore this theory, we make use of data sets from three studies in which up to four mailings were sent to a target population and subsequent calls to census helplines were recorded. Each of these operations is referred to as a National Census Bureau Survey (NCBS) in mailing materials. The materials in these studies mimicked typical U.S. Census Bureau mailings for production surveys at that time. The first and fourth mailings included a one page letter with the help phone line and the survey URL. These mailings also included a thicker stock card with an authentication code for the web survey and the mailing address which showed through the envelope window. The second and third mailings were postcards which also included the phone line, URL, and authentication codes. The 2016 September NCBS (Eggleston and Coombs 2017) and 2016 June NCBS (Coombs 2017) are two studies where an unstaggered mailing strategy was utilized. A staggered mailing schedule was used in the 2017 March NCBS (Nichols et al. 2019); mail recipients were randomly assigned into either a Monday Mailout group, to whom three out of four mailings were initiated on Mondays, or a Thursday Mailout, where three out of four mailings were initiated on Thursdays.Table 1 displays the schedules for each mailing in the three studies. In all three operations, no live agents were present to answer the helpline and callers instead received a prerecorded message. Callers' identities were not recorded so that the data do not distinguish whether multiple calls were made by the same caller.

For each study, we designate day 1 as the day of the first mailing. For the 2017 March NCBS, where there are two mailing schedules, day 1 is the day of the very first mailing, Monday May 6. We then designate week 1 as days 1–7, week 2 as days 8–14, and so forth. We consider weeks 1–5 in each study, and disregard calls which occurred in week 6 or later because call activity becomes sparse. For each pair of studies, we compare the uniformity for week j of the first study to week j of the second study, for $j = 1, \ldots, 5$. It is possible to consider other methods of designating weeks, such as counting each Sunday as the start of a new week; however, our main interest is in call behavior relative to the mailing schedule. We also considered dropping weekends or consolidating Saturday and Sunday into a single "weekend" category, but decided to keep weekends intact. Changing designations of weeks could substantially change results, and such a choice should ideally not be based on the observed data. Table 2 reports the weekly call counts for each study. It is also possible to compare different weeks from pairs of studies, but this yields a large number of possible comparisons. According to our definition of weeks, there is very little opportunity to receive calls in week 1 of the 2017 March NCBS for the Thursday Mailout

Table 1. The schedule of mailings for each study. The 2017 March NCBS (a) was targeted to 8,000 recipients. Of these, half were assigned to the Monday mailout group, and half were assigned to the Thursday mailout group. The 2016 September NCBS (b) was targeted to 9,000 recipients. The 2016 June NCBS (c) was targeted to 8,000 recipients.

(a) 2017 March NCBS.

	Monday Mailout Group		Thursday Mailout Group	
Mailing	Date	Day of Week	Date	Day of Week
1	March 6, 2017	Monday	March 9, 2017	Thursday
2	March 9, 2017	Thursday	March 13, 2017	Monday
3	March 20, 2017	Monday	March 23, 2017	Thursday
4	March 27, 2017	Monday	March 30, 2017	Thursday

(b) 2016 September NCBS.

Mailing	Date	Day of Week
1	August 25, 2016	Thursday
2	September 1, 2016	Thursday
3	September 8, 2016	Thursday
4	September 15, 2016	Thursday

(c) 2016 June NCBS.

Mailing	Date	Day of Week
1	June 13, 2016	Monday
2	June 15, 2016	Wednesday
3	June 24, 2016	Friday
4	July 5, 2016	Tuesday

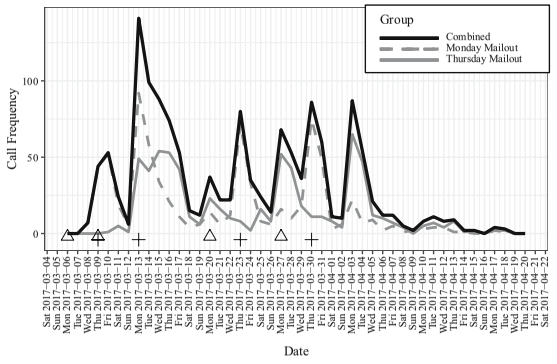
Table 2. Call counts by designated week of study.

			2017 March			
Week	2016 June	2016 Sept	Mon	Thu	Total	$\hat{\pi}_j$
1	353	490	127	7	134	0.9478
2	747	689	226	256	482	0.4414
3	757	970	151	83	234	0.6453
4	383	528	177	147	324	0.5463
5	273	129	48	145	193	0.2487
Total	2513	2806	729	638	1367	

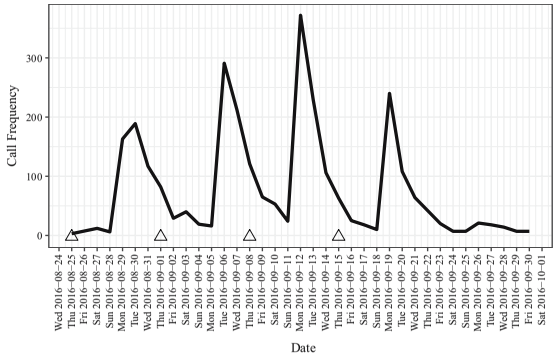
Note: The quantity $\hat{\pi}_j$ in the last column, defined in the next section, is the ratio of the entries under the Mon column to those in the Total column.

group (seven calls). Therefore, if increased uniformity is observed in week 1 for the 2017 March NCBS, it is likely due to some factor other than the staggered mailing schedule.

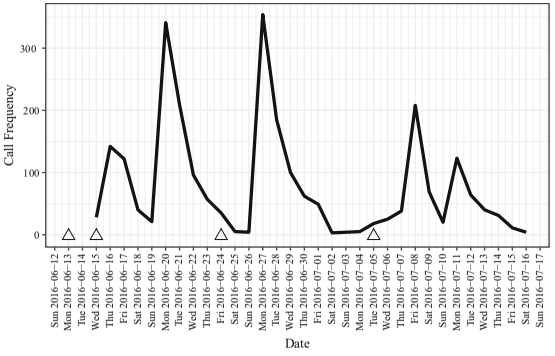
To compare uniformity of call volumes in the three studies, we first examine plots of call frequencies. Figure 1 presents daily call volumes for the three studies. Mailing dates are marked in each plot. Receipt times of the mailings were not known precisely; however, spikes in call volumes can be observed about three days after each mailing, or on the following workday if the



(a) 2017 March NCBS.



(b) 2016 September NCBS.



(c) 2016 June NCBS.

Fig. 1. Daily call volumes during the three NCBS studies. In (a), triangles (Δ) and pluses (+) along the x-axis represent mailing dates for Monday and Thursday Mailout groups, respectively. In (b) and (c), triangles (Δ) represent the mailing dates.

third day happened to fall on a weekend. Even with a staggered mailout, Figure 1a exhibits spikes on Mondays for both mailing schedules, as well as on the expected delivery date of Thursday for the Monday Mailout group. Because the expected delivery day is Monday for the Thursday Mailout group, call volumes primarily spike on Mondays and decrease throughout the rest of the week. For the 2016 June NCBS, spikes can again be seen either on Mondays or three days after a mailing if that day fell on a weekday. A similar pattern can be seen in the 2016 September NCBS; note that September 5, 2016 was a Labor Day holiday so that the expected spike in call volume shifted from Monday to the next day (September 6). Figure 2 presents call frequencies summed by day of week for each study. It appears that the distribution for the 2017 March NCBS (Monday and Thursday mailing groups combined) is flatter than either the distributions for the 2016 September NCBS or 2016 June NCBS. Figure 2d exhibits a large spike on Monday, and therefore appears to be the furthest from being uniform.

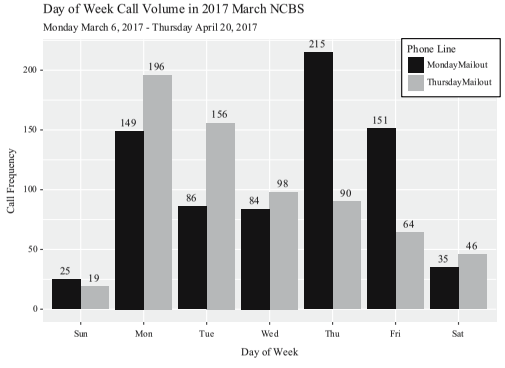
Section 3 describes a framework to carry out more formal statistical analyses. We then return to analysis of the data in Section 4 and report the results. It should be emphasized that the three data sets were collected in three separate NCBS studies, and not as the result of a controlled experiment. Therefore, significant differences in uniformity might be suspected to be caused by factors other than staggering which were not controlled. Such factors include timing of the mailings and timing of the study within the years 2016–2017. Although conclusions which can be drawn from the data are limited, the data provide a concrete example that illustrates the methodology.

3. Methodology

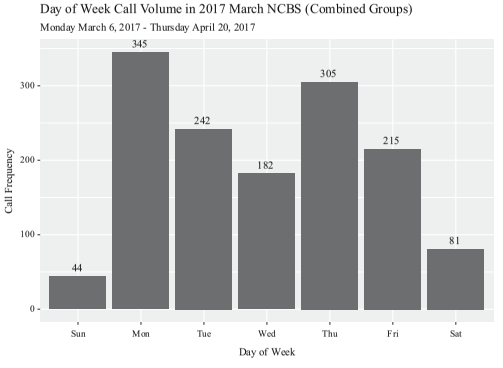
Suppose $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{q} = (q_1, \ldots, q_K)$ are probability distributions on categories labeled $1, \ldots, K$. In our application, categories $1, \ldots, K$ represent the $K =$ seven days of the week: Sun, Mon, $\ldots$, Sat, and $\boldsymbol{p}$ and $\boldsymbol{q}$ are probabilities of a mail recipient calling the helpline on those days, given that the call will occur during that week. In general, we can consider $\boldsymbol{p}$ and $\boldsymbol{q}$ to be probability vectors on any K categories. Let $D(\boldsymbol{p}, \boldsymbol{q}) = \sum_{k=1}^{K} p_k \log\left(p_k/q_k\right)$ denote the K-L divergence between distributions with probabilities $\boldsymbol{p}$ and $\boldsymbol{q}$. Let $\varepsilon(p) = -\sum_{k=1}^{K} p_k \log p_k$ denote the entropy of a distribution with probabilities $\boldsymbol{p}$. Write $\bar{e} = (1/K, \ldots, 1/K)$ and let $\boldsymbol{e}_k$ be a point mass on the kth category; that is, a vector with a 1 in the kth position and zeros in the remaining $K-1$ positions.

As noted in the introduction, the methodology developed in this section is based on the K-L divergence measure. We will say that $\boldsymbol{q}$ is "more uniform" than $\boldsymbol{p}$ if $\boldsymbol{q}$ is closer to $\bar{e}$ than $\boldsymbol{p}$ in terms of K-L divergence; in other words, if

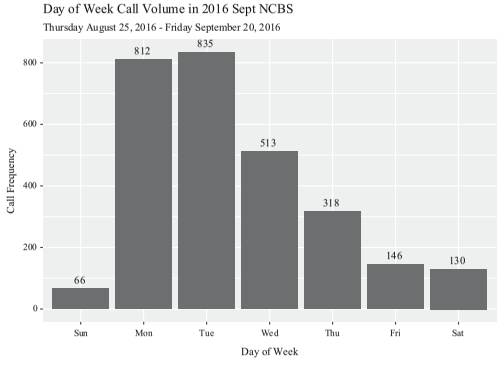
$$
\begin{aligned}
& D(\boldsymbol{p}, \bar{e}) > D(\boldsymbol{q}, \bar{e}) \\
& \Leftrightarrow \left[\sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log (1/K)\right] > \left[\sum_{k=1}^{K} q_k \log q_k - \sum_{k=1}^{K} q_k \log (1/K)\right] \\
& \Leftrightarrow \sum_{k=1}^{K} p_k \log p_k > \sum_{k=1}^{K} q_k \log q_k \\
& \Leftrightarrow \varepsilon(p) < \varepsilon(q). \qquad (1)
\end{aligned}
$$



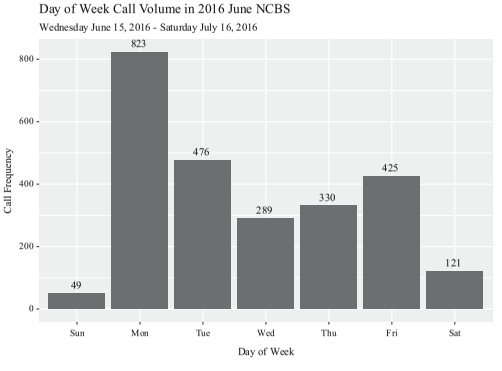
(a) 2017 March NCBS; Monday and Thursday Mailout groups separated.



(b) 2017 March NCBS; Monday and Thursday Mailout groups combined.



(c) 2016 September NCBS.



(d) 2016 June NCBS.

Fig. 2. Day-of-week call volumes for three NCBS studies. The 2017 March NCBS is displayed with Monday and Thursday Mailout groups separated in (a) and combined in (b).

It is well known that entropy is maximum at $\bar{e}$,

$$\varepsilon(p) \leq \varepsilon(\bar{e}) = \log K,$$

over any $\boldsymbol{p} = (p_1, \ldots, p_K)$, and minimized by the point masses,

$$\varepsilon(p) \geq \varepsilon(e_k) = 0, \text{ for any } k = 1, \ldots, K.$$

Suppose $\boldsymbol{p}$ and $\boldsymbol{q}$ are parameterized by unknown $\boldsymbol{\theta}$ to be discussed later in this section. Let $g(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{q}) - \varepsilon(\boldsymbol{p})$ represent the difference in entropy. Motivated Equation (1), we will consider testing hypotheses of the form

$$H_0 : g(\boldsymbol{\theta}) = 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) \neq 0, \tag{2}$$

$$H_0 : g(\boldsymbol{\theta}) \leq 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) > 0, \tag{3}$$

$$H_0 : g(\boldsymbol{\theta}) \geq 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) < 0. \tag{4}$$

If H_0 is rejected in Equation (3), for example, we conclude that $\boldsymbol{q}$ has a higher entropy, or equivalently that $\boldsymbol{q}$ is closer to the uniform probability vector. Note that these hypotheses are invariant to the order of elements in both $\boldsymbol{p}$ and $\boldsymbol{q}$; this is desirable for call volume analysis because we are primarily interested in comparing flatness of distributions, and not whether volumes have simply shifted to different days of the week. In addition to hypothesis testing, we consider point estimates and confidence intervals for the quantity $g(\boldsymbol{\theta})$. We provide additional information on the statistical properties in Appendix (Section 6).

We make note of several points before proceeding:

(A) As a guide to interpret the size of the effect $g(\boldsymbol{\theta})$, recall that $0 \leq \varepsilon(\boldsymbol{p}) \leq \log K$ for any $\boldsymbol{p}$, so that $-\log K \leq g(\boldsymbol{\theta}) \leq \log K$.

(B) Let $\varepsilon_a(\cdot)$ denote the entropy function where logarithms are taken under base a. Here, $\varepsilon_a(\boldsymbol{q}) - \varepsilon_a(\boldsymbol{p}) = [\log a]^{-1} g(\boldsymbol{\theta})$, so that the change of base serves only to scale our quantity of interest by a constant. Therefore, we will consider natural logarithms for the remainder of the article.

(C) It is possible to compare the entropies of $\boldsymbol{p}$ and $\boldsymbol{q}$ with different numbers of support points. If $\boldsymbol{p} = (p_1, \ldots, p_{K_1})$, $\boldsymbol{q} = (q_1, \ldots, q_{K_2})$, and $\bar{e}^{(K)} = (1/K, \ldots, 1/K)$, we obtain the analog to Equation (1) that $D(\boldsymbol{p}, \bar{e}^{(K_1)}) > D(\boldsymbol{q}, \bar{e}^{(K_2)}) \Leftrightarrow \varepsilon(\boldsymbol{p}) - \log K_1 < \varepsilon(\boldsymbol{q}) - \log K_2$. For example, multinomial outcomes from a month with $K_1 = 31$ days may be compared to a month with $K_2 = 28$ days.

(D) As noted in the introduction, and as is well known, the K-L divergence measure is a directed divergence measure that is not symmetric with respect to the two distributions. This is appropriate for our purpose since our focus is on a measure of closeness directed towards uniformity of the probability vectors.

Let $\boldsymbol{X} \sim \text{Mult}_K(m, \boldsymbol{p})$ denote that random variable $\boldsymbol{X}$ has a multinomial distribution

$$\text{P}(X = x) = \frac{m!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}, \; x \in \{0, 1, \ldots, m\}^K, \; \sum_{k=1}^{K} x_k = m.$$

Consider the comparison of two studies where a total of I mailing schedules were attempted among the two. For the ith mailing schedule, let J_i denote the total number of

weeks of the study. In our data, all J_i are equal and represented by a common J. Define $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijK})$ as the call counts observed on (Sun, Mon,. . ., Sat) on the jth week for the ith mailing schedule, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We will assume that

$$X_{ij} \overset{\text{ind}}{\sim} \text{Mult}_K\left(m_{ij}, p_{ij}\right), \tag{5}$$

where $\boldsymbol{p}_{ij} = (p_{ij1}, \ldots, p_{ijK})$ is the day-of-week distribution and $m_{ij} = \sum_{k=1}^{K} X_{ijk}$ is the total call count on the jth week for the ith mailing schedule. With no further assumptions about probabilities $\boldsymbol{p}_{ij}$, the unknown parameter is $\theta = \left(p_{11}^{-}, \ldots, p_{1J}^{-}, \ldots, p_{I1}^{-}, \ldots, p_{IJ}^{-}\right)$ where $p_{ij}^{-} = \left(p_{ij2}, \ldots, p_{ijK}\right)$. It is possible to extend the model to support regression; see Appendix (Subsection 6.2). Note that model (5) regards the totals m_{ij} as fixed, but regards the day of week of each call as an independent multinomial (random) trial. We will specifically consider two scenarios:

1. Two studies with $I = 2$ and one mailing schedule used in each study.
2. Two studies with $I = 3$; one mailing schedule was used in the first study and two were used in the second. Here we assume that $i = 1$ corresponds to the first and $i \in \{2, 3\}$ corresponds to the second.

All studies under consideration consist of $J =$ five weeks of data, and we will compare them on a week-by-week basis. Under Scenario 1, we are interested in

$$g_j(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{p}_{2j}) - \varepsilon(\boldsymbol{p}_{1j}), \quad j = 1, \ldots, J.$$

We will write $g(\boldsymbol{\theta})$ when referring to the entropy difference in a general sense, and $g_j(\boldsymbol{\theta})$ when a particular week is emphasized. For Scenario 2, let $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})$ be the overall day-of-week distribution for calls from the jth week of the second study. To combine data from the two mailing schedules, define random variables A and B as follows. For a given call, let $A \in \{1, \ldots, K\}$ indicate the day-of-week in which the call occurs and let $B \in \{2, 3\}$ indicate the mailing schedule to which the caller was assigned. Suppose $\text{P}(B = 2) = \pi_j$ and $\text{P}(B = 3) = 1 - \pi_j$. The overall probability of receiving a call on day of week $k \in \{1, \ldots, K\}$ during week j is

$$\begin{aligned} q_{jk} &= \text{P}(A = k) \\ &= \text{P}(A = k \mid B = 2)\text{P}(B = 2) + \text{P}(A = k \mid B = 3)\text{P}(B = 3) \\ &= \pi_j p_{2jk} + (1 - \pi_j) p_{3jk}. \end{aligned}$$

We may then write $\boldsymbol{q}_j = \pi_j \boldsymbol{p}_{2j} + (1 - \pi_j)\boldsymbol{p}_{3j}$, and our quantities of interest are

$$g_j(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{q}_j) - \varepsilon(\boldsymbol{p}_{1j}), \; j = 1, \ldots, J.$$

The π_j are unknown and therefore will be replaced by estimates $\hat{\pi}_j = m_{2j}/(m_{2j} + m_{3j})$. Our analysis will be carried out conditionally on the m_{ij} for the sake of tractability; however, modeling the m_{ij} as observed quantities would likely express additional variability in the results and may be considered in future work.

Appendix (Subsection 6.1) describes a statistic Z based on an estimate $g(\hat{\boldsymbol{\theta}})$ of $g(\boldsymbol{\theta})$. Denoting α as the desired significance level for a test and z_α as the $1-\alpha$ quantile of the

standard normal distribution, we obtain familiar tests for $g(\boldsymbol{\theta})$: reject H_0 in hypothesis (2) if $|Z| > z_{\alpha/2}$, reject H_0 in hypothesis (3) if $Z > z_\alpha$, or reject H_0 in hypothesis (4) if $Z < z_\alpha$. We also obtain familiar confidence limits for $g(\boldsymbol{\theta})$ with level $1-\alpha$: the two-sided confidence interval $g(\hat{\theta}) \pm z_{\alpha/2} \cdot \widehat{\mathrm{SE}}$, the lower confidence limit $g(\hat{\theta}) - z_\alpha \cdot \widehat{\mathrm{SE}}$, and the upper confidence limit $g(\hat{\theta}) + z_\alpha \cdot \widehat{\mathrm{SE}}$. Properties of the test and confidence bounds based on Z are investigated by simulation in the supplementary material. The tables in the supplementary material give the estimated type I error probabilities and powers of the one-sided test for $g(\boldsymbol{\theta})$ based on Z, empirical coverage probabilities of the lower confidence limit for $g(\boldsymbol{\theta})$, and the empirical widths associated with the lower confidence limits. The numerical results correspond to a 5% nominal level for the test, and a 95% nominal level for the confidence limit. The numerical results show that the estimated type I error probabilities and the estimated coverage probabilities approach their respective nominal levels as the sample size becomes large. Furthermore, the power of the test increases as $g(\boldsymbol{\theta})$ approaches its maximum value. These results are to be anticipated.

It should be noted that in addition to the K-L divergence criterion, one can think of numerous other criteria that are appropriate for assessing the relative closeness of two sets of multinomial cell probabilities to the uniform probability vector; for example, the power divergence family of criteria investigated in Cressie and Read (1984). An extensive numerical investigation will be necessary before recommendations can be made on specific criteria for practical use. We have not undertaken such a simulation study since there is already some precedence in using the K-L divergence measure, as noted in Subsection 1.2. Furthermore, the focus of our work is on an application relevant to the U.S. Census Bureau, rather than the comparison of various criteria. Thus we are not making a definite recommendation in favor of tests based on the K-L divergence criterion, compared to other tests; for example, Anderson-Darling and other empirical distribution function based tests.

4. Data Analysis

Subsections 4.1, 4.2, and 4.3 compare the three possible pairs among the three NCBS studies. In each section, the given pair is compared using tests and confidence intervals based on the Z statistic from Section 3. Code for all procedures has been implemented in the R programming language (R Core Team 2022).

Although each dataset has been used twice among the three analyses, we have not adjusted results for multiple comparisons (e.g.,Westfall et al. 2011). It should however be noted that a number of p-values in our analysis are quite small and such an adjustment will not matter.

4.1. 2016 September NCBS versus 2017 March NCBS

Our first comparison is between the 2016 September NCBS and 2017 March NCBS call volumes, which falls into Scenario 2. Take $\boldsymbol{X}_{1j}$ to be the call frequencies observed (Sunday, Monday,. . ., Saturday) on the jth week of the 2016 September NCBS. Accordingly, $\boldsymbol{X}_{2j}$ and $\boldsymbol{X}_{3j}$ are call frequencies on the jth week of the 2017 March NCBS for the Monday Mailout and Thursday Mailout groups, respectively. Estimates for π_j are given in Table 2. We test hypothesis (3) for each week $j = 1,. . ., 5$, which can be written as follows.

[Test 1] H_0: "The day-of-week distribution in week j resulting from the 2016 September NCBS mailing schedule has larger or equal entropy for the call probabilities than the day-of-week distribution resulting from the 2017 March NCBS mailing schedule" versus H_1: "Not".

Therefore, rejection of H_0 for week j suggests that the 2017 March NCBS mailing strategy leads to call probabilities closer to the uniform probability vector during that week.

Table 3a gives results of testing this hypothesis for weeks 1–5. Recall that quantity $g(\boldsymbol{\theta})$ is bounded, so that $-1.946 \leq g(\boldsymbol{\theta}) \leq 1.946$ for any $\boldsymbol{\theta}$. The U.S. Census Bureau uses $\alpha = 0.10$ as its standard significance level for hypothesis testing. H_0 can be rejected for weeks 1–4, but there is insufficient evidence to reject during week 5. The Z-statistic is a rather large negative value in week 5, suggesting that there is evidence that the 2016 September NCBS call distribution had higher entropy during that time. Table 4a displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{q}_j$ for weeks $j = 1, \ldots, 5$. We notice in week 1 that the 2017 March NCBS had a higher estimated entropy despite very low call probabilities on Monday and Tuesday; recall that it is very unlikely for Monday Mailout group respondents to receive the first mailing by this Monday or Tuesday, and impossible for the Thursday Mailout group. In week 5, large peaks are observed for the 2017 March NCBS on Monday

Table 3. Results for inference on the quantity g(θ). Estimates, standard errors (SEs), and Z-statistics are displayed in the first three columns. (a) and (b) give a p-value for hypothesis (4) and a level 0.90 lower confidence limit for g(θ). (c) gives a p-value for hypothesis (2) and a level 0.90 (two-sided) confidence interval for g(θ).

(a) 2016 September NCBS to 2017 March NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo
1	0.0823	0.0602	1.3681	0.0856	0.0052
2	0.2605	0.0402	6.4731	<0.0001	0.2090
3	0.1480	0.0411	3.6026	0.0002	0.0953
4	0.2273	0.0453	5.0166	<0.0001	0.1693
5	-0.3376	0.0775	-4.3563	1.0000	-0.4369

(b) 2016 June NCBS to 2017 March NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo
1	-0.0153	0.0631	-0.2426	0.5958	-0.0962
2	0.3463	0.0379	9.1467	<0.0001	0.2977
3	0.3905	0.0442	8.8356	<0.0001	0.3338
4	0.3523	0.0565	6.2409	<0.0001	0.2800
5	0.0253	0.0640	0.3956	0.3462	-0.0567

(c) 2016 June NCBS to 2016 September NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo	CI Hi
1	-0.0976	0.0451	-2.1630	0.0305	-0.1719	-0.0230
2	0.0857	0.0433	1.9782	0.0479	0.0144	0.1570
3	0.2425	0.0365	6.6499	<0.0001	0.1825	0.3025
4	0.1250	0.0607	2.0602	0.0394	0.0252	0.2247
5	0.3629	0.0535	6.7823	<0.0001	0.2749	0.4509

Table 4. Estimated probabilities for data analyses. Estimates are ordered from largest to smallest within each week. The corresponding day of week is shown to the right of each probability. The column labeled $\hat{\varepsilon}$ displays the entropy of the estimated probabilities.

(a) 2016 September NCBS vs. 2017 March NCBS.

Week	$\hat{p}$ (Week)							$\hat{\varepsilon}$
1	0.3857 Tue	0.3327 Mon	0.2388 Wed	0.0245 Sat	0.0122 Sun	0.0061 Thu	0.0000 Fri	1.2515
2	0.4224 Tue	0.3077 Wed	0.1190 Thu	0.0581 Sat	0.0421 Fri	0.0276 Sun	0.0232 Mon	1.4650
3	0.3835 Mon	0.2361 Tue	0.1247 Thu	0.1093 Wed	0.0670 Fri	0.0546 Sat	0.0247 Sun	1.6414
4	0.4545 Mon	0.2045 Tue	0.1212 Wed	0.1193 Thu	0.0473 Fri	0.0341 Sat	0.0189 Sun	1.5272
5	0.3256 Thu	0.1628 Mon	0.1550 Fri	0.1395 Tue	0.1085 Wed	0.0543 Sun	0.0543 Sat	1.7819
Week	$\hat{q}$ (Week)							$\hat{\varepsilon}$
1	0.3955 Fri	0.3284 Thu	0.1791 Sat	0.0522 Wed	0.0448 Sun	0.0000 Mon	0.0000 Tue	1.3338
2	0.2925 Mon	0.2054 Tue	0.1826 Wed	0.1535 Thu	0.1100 Fri	0.0311 Sat	0.0249 Sun	1.7255
3	0.3419 Thu	0.1581 Mon	0.1496 Fri	0.1026 Sat	0.0940 Tue	0.0940 Wed	0.0598 Sun	1.7894
4	0.2654 Thu	0.2099 Mon	0.1852 Fri	0.1636 Tue	0.1111 Wed	0.0340 Sat	0.0309 Sun	1.7545
5	0.4508 Mon	0.2798 Tue	0.1088 Wed	0.0622 Thu	0.0622 Fri	0.0259 Sat	0.0104 Sun	1.4443

(b) 2016 June NCBS vs. 2017 March NCBS.

Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.4023 Thu	0.3456 Fri	0.1133 Sat	0.0793 Wed	0.0595 Sun	0.0000 Mon	0.0000 Tue	1.3492
2	0.4565 Mon	0.2798 Tue	0.1285 Wed	0.0763 Thu	0.0469 Fri	0.0067 Sat	0.0054 Sun	1.3793
3	0.4676 Mon	0.2444 Tue	0.1321 Wed	0.0819 Thu	0.0647 Fri	0.0053 Sun	0.0040 Sat	1.3989
4	0.5431 Fri	0.1802 Sat	0.0992 Thu	0.0653 Wed	0.0522 Sun	0.0470 Tue	0.0131 Mon	1.4022
5	0.4505 Mon	0.2344 Tue	0.1465 Wed	0.1136 Thu	0.0403 Fri	0.0147 Sat	0.0000 Sun	1.4190
Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.3955 Fri	0.3284 Thu	0.1791 Sat	0.0522 Wed	0.0448 Sun	0.0000 Mon	0.0000 Tue	1.3338
2	0.2925 Mon	0.2054 Tue	0.1826 Wed	0.1535 Thu	0.1100 Fri	0.0311 Sat	0.0249 Sun	1.7255
3	0.3419 Thu	0.1581 Mon	0.1496 Fri	0.1026 Sat	0.0940 Tue	0.0940 Wed	0.0598 Sun	1.7894
4	0.2654 Thu	0.2099 Mon	0.1852 Fri	0.1636 Tue	0.1111 Wed	0.0340 Sat	0.0309 Sun	1.7545
5	0.4508 Mon	0.2798 Tue	0.1088 Wed	0.0622 Thu	0.0622 Fri	0.0259 Sat	0.0104 Sun	1.4443

(c) 2016 June NCBS vs. 2016 September NCBS.

Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.4023 Thu	0.3456 Fri	0.1133 Sat	0.0793 Wed	0.0595 Sun	0.0000 Mon	0.0000 Tue	1.3492
2	0.4565 Mon	0.2798 Tue	0.1285 Wed	0.0763 Thu	0.0469 Fri	0.0067 Sat	0.0054 Sun	1.3793
3	0.4676 Mon	0.2444 Tue	0.1321 Wed	0.0819 Thu	0.0647 Fri	0.0053 Sun	0.0040 Sat	1.3989
4	0.5431 Fri	0.1802 Sat	0.0992 Thu	0.0653 Wed	0.0522 Sun	0.0470 Tue	0.0131 Mon	1.4022
5	0.4505 Mon	0.2344 Tue	0.1465 Wed	0.1136 Thu	0.0403 Fri	0.0147 Sat	0.0000 Sun	1.4190
Week	$\hat{q}$ (Week)							$\hat{\varepsilon}$
1	0.3857 Tue	0.3327 Mon	0.2388 Wed	0.0245 Sat	0.0122 Sun	0.0061 Thu	0.0000 Fri	1.2515
2	0.4224 Tue	0.3077 Wed	0.1190 Thu	0.0581 Sat	0.0421 Fri	0.0276 Sun	0.0232 Mon	1.4650
3	0.3835 Mon	0.2361 Tue	0.1247 Thu	0.1093 Wed	0.0670 Fri	0.0546 Sat	0.0247 Sun	1.6414
4	0.4545 Mon	0.2045 Tue	0.1212 Wed	0.1193 Thu	0.0473 Fri	0.0341 Sat	0.0189 Sun	1.5272
5	0.3256 Thu	0.1628 Mon	0.1550 Fri	0.1395 Tue	0.1085 Wed	0.0543 Sun	0.0543 Sat	1.7819

and Tuesday. Many calls occurring on these peak days are from the Thursday Mailout group, as seen in Figure 1a, whose final mailing was initiated the previous Thursday (March 30). However, calls are also contributed from the Monday Mailout group, whose final mailing was the previous Monday (March 27).

4.2. 2016 June NCBS versus 2017 March NCBS

Our second comparison is between the 2016 June NCBS versus 2017 March NCBS call volumes, which falls into Scenario 2. Take $\boldsymbol{X}_{1j}$ to be the frequencies of the 2016 June

NCBS calls observed on (Sunday, Monday,. . ., Saturday). Take $\boldsymbol{X}_{2j}$, $\boldsymbol{X}_{3j}$, and π_j to be defined as in Subsection 4.1. We test hypothesis (3) for each week $j = 1, \ldots, 5$, which can be written as follows.

> [Test 2] H_0: "The day-of-week distribution in week j resulting from the 2016 June NCBS mailing schedule has larger or equal entropy for the call probabilities than the day-of-week distribution resulting from the 2017 March NCBS mailing schedule" versus H_1: "Not".

Here, rejection of H_0 for week j suggests that the 2017 March NCBS mailing strategy leads to call probabilities closer to the uniform probability vector during that week.

Table 3b gives results of testing this hypothesis for weeks 1–5. Table 4b displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{q}_j$ for weeks $j = 1, \ldots, 5$. There is strong evidence to reject H_0 for weeks 2, 3, and 4, but insufficient evidence to reject during weeks 1 and 5 at significance level $\alpha = 0.10$. Recall that the Thursday Mailout group had little opportunity to call during week 1, so the results for this week are primarily based on the Monday Mailout group.

4.3. 2016 June NCBS versus 2016 September NCBS

Our third comparison is between the 2016 September NCBS and 2016 June NCBS call volumes, which matches Scenario 1. Take $\boldsymbol{X}_{1j}$ to be the frequencies of 2016 June NCBS as in Subsection 4.2 and $\boldsymbol{X}_{2j}$ to be the frequencies of 2016 September NCBS calls as in Subsection 4.2. Because both studies used a single mailing strategy for all respondents, we assume a null hypothesis that neither strategy leads to a significantly more uniform call distribution. Therefore, we test hypothesis (2) for each week $j = 1, \ldots, 5$, which can be written as follows.

> [Test 3] H_0: "The day-of-week distribution in week j resulting from the 2016 June NCBS mailing schedule and that resulting from the 2016 September NCBS mailing schedule have equal entropies for the call probabilities" versus H_1: "Not".

Here, rejection of H_0 for week j suggests that the two mailing strategies do not lead to call probabilities that are equally close to the uniform probability vector during that week.

Table 3c gives results of testing this hypothesis for weeks 1–5, and Table 4c displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{p}_{2j}$ for weeks $j = 1, \ldots, 5$. The test can be rejected at significance level $\alpha = 0.10$ for all five weeks, although the evidence is much stronger in weeks 3 and 5. The Z-statistics are positive for weeks 2–5, indicating a larger entropy for the 2016 September NCBS except during week 1.

5. Discussion and Conclusions

In this work, we analyzed call volumes observed in several census studies to determine whether one mailing strategy produced more uniform call probabilities than another. To do this, we made use of statistical tests and confidence intervals determined from basic large sample theory. Our analysis found that a staggered strategy–as used in the 2017 March NCBS–yielded a significantly higher entropy than the two unstaggered strategies toward the middle of the study period, after both the Monday and Thursday Mailout groups

received the first mailing. However, the two unstaggered strategies–the 2016 September NCBS and 2016 June NCBS–also yielded significantly different entropies when compared to each other; this demonstrates that other aspects of mailing schedule design, aside from staggering, affect uniformity of calls from week to week. After the final mailing is sent, the choice of mailing schedule is expected to have a diminishing effect on call uniformity as the overall volume of calls diminishes. One way to attenuate starting and ending differences among the operations would be, say, to combine weeks 1 and 2 into a "beginning period", label week 3 as a "middle period", and combine weeks 4+ into an "ending period"; the methodology could be applied to the three periods instead of the individual weeks without any changes.

Although some evidence was found in the data that staggering increases call uniformity, the three featured studies were carried out separately and may have been subject to subtle variations other than mailing strategy. A controlled experiment would help to isolate this effect from other factors. Procedures such as the ones presented in this article could help to guide design and sample size determination for such an experiment.

The report by Nichols et al. (2019) represents the analysis-of-record based on the methodology in the present article. Marasteanu (2019) describes a staggered mailing strategy for the 2020 decennial census which makes use of findings by Nichols et al. (2019) and others, while also taking into account a number of important practical concerns. An unspecified model is used to express volumes to call centers and to illustrate the effects of staggered mailings on them over time.

While our model was based on independent multinomial observations, we could consider a regression model with appropriate covariates as described in Appendix (Subsection 6.2). Here, it becomes necessary to check model adequacy–e.g., via goodness-of-fit testing–before proceeding with inference on $g(\boldsymbol{\theta})$. However, were a sufficiently good predictive model available, it could be used to optimize over a class of mailing strategies and identify which one(s) achieved an optimal uniformity. This could be an objective in future analyses of U.S. Census Bureau operations, as a step beyond statistical inference.

6. Appendix

6.1. Inference on Entropy Difference

In order to discuss statistical procedures, let us generally write

$$g_j(\boldsymbol{\theta}) = \varepsilon(c_1\boldsymbol{p}_{1j} + \cdots + c_I\boldsymbol{p}_{Ij}) - \varepsilon(d_1\boldsymbol{p}_{1j} + \cdots + d_I\boldsymbol{p}_{Ij}), \quad j = 1, \ldots, J, \tag{6}$$

for given coefficients $\boldsymbol{c} = (c_1, \ldots, c_I)$ and $\boldsymbol{d} = (d_1, \ldots, d_I)$ which are distributions on $\{1, \ldots, I\}$. We do not encounter a situation where two studies use data from a common mailing schedule; therefore, we will have $c_i d_i = 0$ for $i = 1, \ldots, I$. In a multinomial analysis, one of our day-of-week categories is redundant because $\sum_{k=1}^{K} X_{ijk} = m_{ij}$ and $\sum_{k=1}^{K} p_{ijk} = 1$. Without loss of generality, we will consider the first category as the redundant one, and write $X_{ij}^{-} = (X_{ij2}, \ldots, X_{ijK})$ and $p_{ij}^{-} = (p_{ij2}, \ldots, p_{ijK})$. Denote $p^{-} = \left(p_{11}^{-}, \ldots, p_{1J}^{-}, \ldots, p_{I1}^{-}, \ldots, p_{IJ}^{-}\right)$, and its maximum likelihood estimator as $\hat{p}^{-}$, which replaces each p_{ij}^{-} with $\hat{p}_{ij}^{-} = X_{ij}^{-}/m_{ij}$. Under model (5), the unknown $\boldsymbol{\theta}$ is taken

assumed to be $\boldsymbol{p}^-$. Write $\boldsymbol{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that random variable $\boldsymbol{T} \in \mathbb{R}^n$ follows a normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and variance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. As discussed in Lehmann (2004, 314), we have a large sample normal distribution $\hat{p}^- \sim \mathrm{N}(\boldsymbol{p}^-, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \mathrm{Blockdiag}\left(m_{11}^{-1}\left[\mathrm{Diag}\left(p_{11}^-\right) - p_{11}^- p_{11}^{-\top}\right], \ldots, m_{IJ}^{-1}\left[\mathrm{Diag}\left(p_{IJ}^-\right) - p_{IJ}^- p_{1J}^{-\top}\right]\right) \quad (7)$$

is an $IJ(K-1) \times IJ(K-1)$ covariance matrix. Furthermore, the delta method (Lehmann 2004, 315) gives the large sample distribution

$$g_j\left(\hat{p}^-\right) \sim \mathrm{N}\left(g_j\left(p^-\right), \sigma^2_{g_j(p-)}\right), \quad \sigma^2_{g_j(p-)} = \left[\frac{\partial g_j(p^-)}{\partial p^-}\right] \Sigma \left[\frac{\partial g_j(p^-)}{\partial p^-}\right]^\top,$$

for $j = 1, \ldots, J$. After some algebra, we obtain the $1 \times IJ(K-1)$ gradient

$$\frac{\partial g_j(p^-)}{\partial p^-} = \begin{pmatrix} e_j \otimes \left[c_1 \nabla \varepsilon\left(\Sigma_{i=1}^I c_i p_{ij}^-\right) - d_1 \nabla \varepsilon\left(\Sigma_{i=1}^I d_i p_{ij}^-\right)\right] \\ e_j \otimes \left[c_I \nabla \varepsilon\left(\Sigma_{i=1}^I c_i p_{ij}^-\right) - d_1 \nabla \varepsilon\left(\Sigma_{i=1}^I d_i p_{ij}^-\right)\right] \end{pmatrix}^\top, \quad (8)$$

where $\nabla \varepsilon(\boldsymbol{x}) = (-\log(x_2/x_1), \ldots, -\log(x_K/x_1))^\top$, $\boldsymbol{e}_j$ is the jth column of a $J \times J$ identity matrix, and $\otimes$ denotes the matrix Kronecker product.

Let $\hat{p}^-$ be a maximum likelihood estimator of $\boldsymbol{p}^-$ and define the statistic $Z = g_j(\hat{p}^-)/\sigma_{g_j(\hat{p}^-)}$ with the expression for $\sigma^2_{g_j(p^-)}$. Now, $Z \sim \mathrm{N}(0,1)$ approximately under the restriction $g_j(\boldsymbol{p}^-) = 0$, which occurs at the boundary of the null hypotheses (2), (3), and (4). Therefore, we obtain the test and confidence interval procedures discussed at the end of Section 3 with $\widehat{\mathrm{SE}} = \sigma_{g(\hat{p}^-)}$.

Some insight into the behavior of $g(\boldsymbol{p}^-)$ can be seen from its gradient. Consider Scenario S1 with $J = 1$ and suppress the j index; we have

$$\begin{aligned} \frac{\partial g\left(p^-\right)}{\partial p^-} &= \begin{pmatrix} -\nabla \varepsilon\left(p_1^-\right)^\top & \nabla \varepsilon\left(p_2^-\right)^\top \end{pmatrix} \\ &= \log(p_{12}/p_{11}) \quad \cdots \quad \log(p_{1K}/p_{11}) \quad -\log(p_{22}/p_{21}) \quad \cdots \quad -\log(p_{2K}/p_{21}). \end{aligned}$$

Because the gradient separates into a component involving only $\boldsymbol{p}_1$ and similar one involving only $\boldsymbol{p}_2$, it suffices to comment only on the former. When $\boldsymbol{p}_1 \approx (1/K, \ldots, 1/K)$, it is seen that $-\nabla \varepsilon(\boldsymbol{p}_1^-) \approx \boldsymbol{0}$; therefore, $g(\boldsymbol{p}^-)$ increases very slowly to its maximum of log K as $\boldsymbol{p}_1$ approaches a discrete uniform distribution. On the other hand, when $\boldsymbol{p}_1 \to \boldsymbol{e}_2$, then $-\nabla \varepsilon(\boldsymbol{p}_1^-) \to (0, \infty, \ldots, \infty)$; therefore, when $\boldsymbol{p}_1$ is close to a point mass, small changes in $\boldsymbol{p}_1$ result in very large changes in some of the components of $g(\boldsymbol{p}^-)$.

This remark notes that changes in the quantity $g(\boldsymbol{p}^-)$ are smaller when the component distributions are closer to uniform. Differences in this setting may therefore be difficult to detect with the Z-statistic, which may warrant consideration of alternative distance measures in future work. Alternative distance measures might also be considered if one is thought to better express the cost of departure from uniformity than K-L divergence.

6.2. Regression Model

The framework can be extended to a more general multinomial regression model for additional flexibility. Suppose we wish to compare two mailing strategies based on I mailing schedules where calls are recorded in J weeks. Let $\boldsymbol{Y}_{ij} = (Y_{i11}, \ldots, Y_{ijK})$ be a vector of call counts received on the kth day in the jth week from the ith mailing schedule, for $i = 1, \ldots, I, j = 1, \ldots, J$, and $k = 1, \ldots, K$, with $m_{ij} = \sum_{k=1}^{K} Y_{ijk}$ total calls received. Let $w_{ijk} \in \mathbb{R}^D$ be a corresponding fixed covariate. The discrete choice model described by Agresti (2013, Sec. 8.5),

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Mult}_K\left(m_{ij}, p_{ij}\right), \quad p_{ijk} = \frac{\exp\left\{w_{ijk}^{\top}\beta\right\}}{\Sigma_{\ell=1}^{k} \exp\left\{w_{ij\ell}^{\top}\beta\right\}}, \tag{9}$$

is a form of multinomial regression which permits the covariate to vary by category. The independent variable w_{ijk} may contain indicators such as day of the week, week of the mailing schedule, holidays and other features of interest which vary by day. In this model, $\log\left(p_{ijk}/p_{ij\ell}\right) = \left(w_{ijk} - w_{ij\ell}\right)^{\top}\beta$ for any $k,\ell \in \{1, \ldots, K\}$. Regression model (9) allows a number of alternative formations; for example, rather than partitioning the data into J weeks (or other non-overlapping time periods) as desired in our application, one may also consider taking $J = 1$, the K multinomial categories to correspond to all days relevant to the analysis, and covariates w_{ijk} to include adjustments for day-of-week and/or day-of-study effects.

Under model (9), the unknown parameter $\boldsymbol{\theta}$ is taken to be the coefficient vector $\boldsymbol{\beta}$. To perform inference on (9) in the regression case, the delta method applied to the maximum likelihood estimator $\hat{\beta}$ yields large sample distribution

$$g\left(\boldsymbol{p}^{-}(\hat{\beta})\right) \sim \text{N}\left(g\left(p^{-}(\beta)\right), \sigma^2_{g\left(p^{-}(\beta)\right)}\right), \quad \sigma^2_{g\left(p^{-}(\beta)\right)} = \left[\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta}\right] \mathcal{I}^{-1}(\beta) \left[\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta}\right]^{\top}$$

where $\mathcal{I}(\beta)$ is the Fisher information matrix with respect to β under the outcomes $\boldsymbol{Y}_{ij}$. An estimate of $\mathcal{I}^{-1}(\beta)$ based on $\hat{\beta}$ may be produced by statistical software such as the `mlogit` package (Croissant 2020) in R, or may be computed directly via $\mathcal{I}(\beta) = \left[\frac{\partial p^{-}(\beta)}{\partial \beta}\right]^{\top} \sum^{-1} \left[\frac{\partial p^{-}(\beta)}{\partial \beta}\right]$ using $\boldsymbol{\Sigma}$ given in Equation (7). To compute the remaining terms of $\sigma^2_{g\left(p^{-}(\beta)\right)}$, we have

$$\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta} = \frac{\partial g\left(p^{-}\right)}{\partial p^{-}} \frac{\partial p^{-}(\beta)}{\partial \beta}$$

with $\frac{\partial g\left(p^{-}\right)}{\partial p^{-}}$ given by Equation (8) and

$$\frac{\partial p_{ijk}(\beta)}{\partial \beta_d} = \left\{\sum_{\ell=1}^{K} \exp\left\{\left(w_{ij\ell} - w_{ijk}\right)^{\top}\beta\right\}\right\}^{-2} \sum_{\ell=1}^{K} \exp\left\{\left(w_{ij\ell} - w_{ijk}\right)^{\top}\beta\right\}\left(w_{ij\ell d} - w_{ijkd}\right).$$

7. References

Agresti, A. 2013. *Categorical Data Analysis*. Hoboken, NJ: Wiley, 3rd edition.

Blei, D.M., A. Kucukelbir, and J.D. McAuliffe 2017. "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association* 112: 859–877. DOI: https://dx.doi.org/10.1080/01621459.2017.1285773.

Chesnut, J. 2003. "Telephone Questionnaire Assistance". Census 2000 Evaluation A.1.a, U.S. Census Bureau. Available at: https://www.census.gov/pred/www/rpts/A.1.a.pdf (accessed June 2022).

Cohen, A., J. Kolassa, and H. Sackrowitz. 2006. "A Test for Equality of Multinomial Distributions vs Increasing Convex Order". In *Recent Developments in Nonparametric Inference and Probability*, edited by J. Sun, A. DasGupta, V. Melfi, and C. Page. *Lecture Notes–Monograph Series*, Volume 50: 156–163. Institute of Mathematical Statistics, Beachwood, OH. DOI: https://dx.doi.org/10.1214/074921706000000662.

Coombs, J. 2017. *Analysis Report for the Small-Scale Mailout Testing Program June 2016 Test on the Placement and Length of the User ID for an Online Census Bureau Survey*. Census Program Internal Memorandum Series 2017.2.i, U.S. Census Bureau.

Cover, T.M., and J.A. Thomas. 2006. *Elements of Information Theory*. Hoboken, NJ: Wiley, 2nd edition.

Cressie, N., and T.R. Read. 1984. "Multinomial Goodness-of-Fit Tests". *Journal of the Royal Statistical Society: Series B* 46: 440–464. DOI: https://dx.doi.org/10.1111/j.2517-6161.1984.tb01318.x.

Croissant, Y. 2020. "Estimation of Random Utility Models in R: The mlogit Package". *Journal of Statistical Software* 95: 1–41. DOI: https://dx.doi.org/10.18637/jss.v095.i11.

Dorfinger, P., G. Panholzer, and W. John. 2011. "Entropy Estimation for Real-Time Encrypted Traffic Identification (Short Paper)". In *Traffic Monitoring and Analysis*, edited by J. Domingo-Pascual, Y. Shavitt, and S. Uhlig: 164–171. Berlin, Heidelberg. Springer. DOI: https://dx.doi.org/10.1007/978-3-642-20305-3_14.

Eggleston, C., and J. Coombs. 2018. *Effect of Data Use Statements and Postcard Format on Login Rates for a Mandatory Online Census Bureau Survey*. U.S. Census Program Internal Memorandum Series 2018.11.i, U.S. Census Bureau.

Girardin, V., and J. Lequesne. 2019. "Entropy-Based Goodness-of-Fit Tests–A Unifying Framework: Application to DNA Replication". *Communications in Statistics–Theory and Methods* 48: 62–74. DOI: https://dx.doi.org/10.1080/03610926.2017.1401084.

Konishi, S., and G. Kitagawa. 2008. *Information Criteria and Statistical Modeling*. New York, NY: Springer.

Lehmann, E.L. 2004. *Elements of Large-Sample Theory*. New York, NY: Springer.

Li, Y., R.S. Singh, and Y. Sun. 2005. "Goodness-of-Fit Tests of a Parametric Density Functions: Monte Carlo Simulation Studies". *Journal of Statistical Research* 39: 103–125. Available at: http://jsr.isrt.ac.bd/article/goodness-fit-tests-parametric-density-functions-monte-carlo-simulation-studies/.

Liu, X., and J. Wang. 2004. "Testing the Equality of Multinomial Populations Ordered by Increasing Convexity Under the Alternative". *Canadian Journal of Statistics* 32: 159–168. DOI: https://dx.doi.org/10.2307/3315939.

Marasteanu, I. 2019. "The Research and Methodology on Staggering the 2020 Census Mailings". In JSM Proceedings, Government Statistics Section: 2457–2467. Alexandria, VA: American Statistical Association.

Nichols, E., S. Konya, R. Horwitz, and A. Raim. 2019. *2020 Census Research and Testing Report: The Effect of the Mail Delivery Date on Survey Login Rates and Helpline Call Rates*. U.S. Census Bureau, Research and Methodology Directorate, Center for Behavioral Science Methods Research Report Series (Survey Methodology): 2019–01,

U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2019/adrm/rsm2019-01.pdf (accessed June 2022).

Ormerod, J.T., and M.P. Wand. 2010. "Explaining Variational Approximations". *The American Statistician* 64: 140–153. DOI: https://dx.doi.org/10.1198/tast.2010.09058.

Paninski, L. 2008. "A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data". *IEEE Transactions on Information Theory* 54: 4750–4755. DOI: https://dx.doi.org/10.1109/TIT.2008.928987.

Pardo, L. 2006. *Statistical Inference Based on Divergence Measures*. Boca Raton, FL: Chapman and Hall/CRC.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Song, K.S. 2002. "Goodness-of-Fit Tests Based on Kullback-Leibler Discrimination Information". *IEEE Transactions on Information Theory* 48: 1103–1117. DOI: https://dx.doi.org/10.1109/18.995548.

U.S. Census Bureau. 2017. *2020 Census Operational Plan*. Available at: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/operational-plan.html. Version 3.0. (accessed June 2022).

Vuong, Q.H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". *Econometrica* 57: 307–333. DOI: https://dx.doi.org/10.2307/1912557.

Westfall, P.H., R.D. Tobias, and R.D. Wolfinger. 2011. *Multiple Comparisons and Multiple Tests Using SAS*. Cary, NC: SAS Institute, Inc., 2nd edition.

Zajac, K. 2012. *2010 Census Telephone Questionnaire Assistance Assessment Report*. 2010 Census Planning Memoranda Series No. 231, U.S. Census Bureau. Available at: https://www.census.gov/2010census/pdf/2010_Census_TQA_Assessment.pdf (accessed June 2022).

Received June 2021
Revised March 2022
Accepted July 2022

sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 123–137, http://dx.doi.org/10.2478/JOS-2023-0006

A Two-Stage Bennet Decomposition of the Change in the Weighted Arithmetic Mean

Thomas von Brasch[1], *Håkon Grini*[1], *Magnus Berglund Johnsen*[1], *and Trond Christian Vigtel*[1]

The weighted arithmetic mean is used in a wide variety of applications. An infinite number of possible decompositions of the change in the weighted mean are available, and it is therefore an open question which of the possible decompositions should be applied. In this article, we derive a decomposition of the change in the weighted mean based on a two-stage Bennet decomposition. Our proposed decomposition is easy to employ and interpret, and we show that it satisfies the difference counterpart to the index number time reversal test. We illustrate the framework by decomposing aggregate earnings growth from 2020Q4 to 2021Q4 in Norway and compare it with some of the main decompositions proposed in the literature. We find that the wedge between the identified compositional effects from the proposed two-stage Bennet decomposition and the one-stage Bennet decomposition is substantial, and for some industries, the compositional effects have opposite signs.

Key words: Index theory; weighted arithmetic mean; decomposition.

1. Introduction

What are the driving forces underlying aggregate productivity growth? Why has the labour force participation rate changed during the last two decades? What has driven the change in annual earnings over the last year and why have import prices changed? All these questions have a common feature in that statistics on productivity, the labour force participation rate, earnings and import prices are often constructed using a weighted arithmetic mean formula.

A natural starting point for answering these questions is to decompose the change in the weighted mean. A frequently used decomposition is the Bennet (1920) decomposition, often also referred to as shift-share analysis. This decomposition enables within-group growth effects to be distinguished from between-group compositional effects. For example, when examining productivity dynamics in U.S. manufacturing plants between 1972 and 1987, Baily et al. (1992) find a positive contribution to growth due to increasing output shares among high-productivity plants and decreasing output shares among low-productivity plants. Daly and Hobijn (2017) show that compositional effects due to labour market status flows are important in explaining aggregate real wage growth in the U.S.

[1] Statistics Norway, Post Box 2633, Oslo 0131, Norway. E-mails: thomas.vonbrasch@ssb.no, knut.grini@ssb.no, magnus.johnsen@ssb.no, and trond.vigtel@ssb.no

Acknowledgments: We thank Bert M. Balk, Pål Boug, W. Erwin Diewert, Terje Skjerpen and three anonymous referees for valuable comments. In particular, thanks to Bert M. Balk for pointing out that we could simplify the proof of Proposition 1 and to W. Erwin Diewert for providing us with a comprehensive analysis of the alternative and approximate decomposition that we compare with our proposed exact two-stage Bennet decomposition. The usual disclaimer applies.

Analysing the fall in the U.S. labour force participation rate, Krueger (2017) finds that the population composition has shifted toward groups with lower participation rates, and that this accounts for well over half of the decline in the labour force participation rate between 1997 and 2017. Moreover, a large body of literature has identified the deflationary effects of international trade resulting from increased import shares from low-price countries, such as China; see, for example, Kamin et al. (2006), Thomas and Marquez (2009) and Benedictow and Boug (2017, 2021).

Although the Bennet decomposition is useful for identifying the overall contribution from compositional effects, it does not identify how much of the overall compositional effect that can be attributed to a particular group or subset. To overcome this shortcoming, the Bennet decomposition is often rewritten by subtracting a scalar A from each group in the between effect, where the scalar A typically represents some measure of the weighted mean, see Huerga (2010) and Balk (2021, 137). Foster et al. (2001) analyzed productivity developments and measured the between effect as the product of changes in the plant-level output share and the deviation of average plant-level productivity from the overall industry average. If the composition of firms changes such that the output share of a low-productivity plant increases, this will lower the aggregate weighted mean productivity level and thus contribute negatively to the compositional effect. Note that in these decompositions a plant may contribute negatively to the compositional effect even if there is no change in the output of that plant. The reason is that it is the output share, and not the output of the plant, that enters the decomposition, and the output share of a given plant may change because the output of all the other plants changes. Moreover, as pointed out by Balk (2021), the choice of the scalar A is arbitrary. Since any scalar may be subtracted from the Bennet decomposition, an infinite number of possible decompositions are available, and it is therefore an open question which of the possible decompositions should be applied.

In this article, we derive a decomposition that identifies the contribution to the overall change in the weighted mean from changes in prices and quantities. In contrast to the previously discussed one-stage Bennet decomposition, which relates to changes in weights, our proposed two-stage Bennet decomposition relates to the change in the underlying quantity variables. This difference is important because the weights can never change without a change in at least one of the underlying quantity variables. Moreover, a change in a single quantity variable will affect all the share variables. Therefore, examining the shares provides only limited information about the underlying driving forces. In the two-stage Bennet decomposition, all the terms related to the within effects are identical to those in the one-stage Bennet decomposition. Also, the overall between effect, or compositional effect, is identical to the overall between effect in the one-stage Bennet decomposition. The group-specific between effects, however, differ from those in the one-stage Bennet decomposition. The decomposition captures the intuitive property that the weighted mean increases if a group whose quantity variable is growing has a level that is above the weighted mean level. There are two ways in which compositional effects for a group will be zero: either the group-specific indicator equals the weighted mean level, and/or there is no change in the quantity variable of that group. The proposed decomposition is easy to employ and interpret and furthermore gives a better platform for comparing groups. Moreover, we show that the decomposition is invariant with respect to

treatment of time, and that it therefore satisfies the difference counterpart to the index number time reversal test; see ILO et al. (2004, 411).

To provide further support for the two-stage Bennet decomposition, we also compare it with a decomposition based on a quadratic approximation of the weighted mean, while considering the weighted mean as a non-linear function of underlying prices and quantities. It is shown that this alternative decomposition has some similar features to the two-stage Bennet decomposition. In particular, it also identifies how the weighted mean price level increases if products that are growing in quantity have price levels that are higher than the mean price level. However, this alternative decomposition is not exact, and the size of the approximation error depends on how much prices, quantities, and weights change.

To illustrate the two-stage Bennet decomposition, we use data on aggregate earnings growth in Norway between 2020Q4 and 2021Q4. In this empirical example, it is shown that the wedge between the identified compositional effects from the two-stage and one-stage Bennet decompositions is substantial, and for some industries, the compositional effects are of opposite signs. We also compare the two-stage Bennet decomposition with the decomposition based on a quadratic approximation. These two decompositions yield similar group-specific contributions to the overall compositional effects.

The article is structured as follows: Section 2 outlines the weighted mean formula, some of the most standard decompositions applied in the literature, our proposed decomposition and the decomposition based on a quadratic approximation. Section 3 contrasts and compares empirically our proposed decomposition with those used in the literature and the decomposition based on a quadratic approximation, using the case of earnings growth in Norway. Section 4 provides a conclusion.

2. Decomposing the Change in the Weighted Mean

Our point of departure is the weighted mean of indicators P_{it} across units i at time t of the form:

$$P_t = \sum_{i=1}^{N} S_{it} P_{it}, \tag{1}$$

with weights $S_{it} = \frac{X_{it}}{\sum_{j=1}^{N} X_{jt}}$, where the quantity variable $X_{it} \geq 0$ and $\sum_{j=1}^{N} X_{jt} > 0$. Note that the weights sum to unity. The weighted mean in Equation (1) has numerous applications within the fields of economics and measurement theory. In some applications, the weighted mean is also referred to as a unit value. To our knowledge, unit values were first introduced by Segnitz (1870). Although the weighted mean has been applied in a variety of fields, and the indicator and quantity variables may refer to "inter alia" wages, hours worked, productivity, output prices and so on, we will henceforth refer to P_{it} and X_{it} as representing prices and quantities, respectively, and unit i as product i. Note that for the weights to have a meaningful interpretation, the quantities involved must be comparable. In the following we are concerned with identifying the contribution to the change in the weighted mean of changes in both prices and quantities. Before we present our proposed decomposition, we start by recapitulating the most widely utilized decompositions in the literature.

2.1. The One-Stage Bennet Decomposition

Bennet (1920) provided a decomposition of the nominal value change into the sum of a price change and a quantity change. This decomposition stands in contrast to traditional index theory, which focuses on decomposing a value ratio into the product of a price index and a quantity index. Diewert (2005) analyzed the axiomatic and economic properties of the Bennet decomposition. When applied to Equation (1), the decomposition yields the identity:

$$\Delta P = \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} \bar{P}_i \Delta S_i, \tag{2}$$

where Δ is the difference operator and a bar over a variable represents the moving average operator between time t and v, that is, $\Delta x = x_t - x_v$ and $\bar{x} = 1/2(x_t + x_v)$, and the time subscript is dropped when it is superfluous, for notational convenience. This one-stage Bennet decomposition is standard in productivity and shift-share analysis, see for example, Baily et al. (1992) and OECD (2018). The terms $\bar{S}_i \Delta P_i$ and $\bar{P}_i \Delta S_i$ represent the contributions to the change in the weighted mean of a change in the price of product i, and in the quantity share of product i, respectively.

2.2. The One-Stage Bennet Decomposition with Extended Weight

The one-stage Bennet decomposition may be rewritten to create terms that explicitly capture the fact that the weighted mean price level increases if the quantity shares of high-priced products increase. Since the weights sum to unity, we can subtract the term $\sum_{i=1}^{N}(A\Delta S_i)$, for any given scalar A, such that:

$$\Delta P = \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} (\bar{P}_i - A)\Delta S_i. \tag{3}$$

In this case, the contribution to the change in the weighted mean of a change in the weight of product i is given by the term $(\bar{P}_i - A)\Delta S_i$. For example, this term captures the increase in the weighted mean price if products whose quantity shares are growing have an average price level between time v and t that is larger than the scalar A. In the literature, the scalar A has often been chosen to represent some measure of the mean price level. For example, the choice $A = \bar{P}$, where $\bar{P} = 1/2(P_t + P_v)$, is often used, see Huerga (2010) and Balk (2021, 137). Note that when $A = \bar{P}$ is chosen, the framework above is invariant with respect to treatment of time, that is, it satisfies the difference counterpart to the index number time reversal test; see Diewert and Fox (2010). The contribution to the change in the weighted mean of a change in the share of product i is then given by $(\bar{P}_i - \bar{P})\Delta S_i$. This term captures the fact that the weighted mean price level increases if products whose quantity shares are growing have a price level that is higher than the weighted mean price level. Conversely, the term also captures the fact that the weighted mean price level will decrease if products whose quantity shares are growing have a price level that is lower than the weighted mean price level.

There are shortcomings related to the two decompositions above if the purpose is to identify how much of the overall compositional effect that can be attributed to a particular

group or subset. First, and as argued by Balk (2021), the choice of the scalar A is completely arbitrary. Second, the weight of product i (S_{it}) may change even if there is no change in the quantity of product i (X_{it}), that is, if there is a change in the sum of all the other products. Since the change in the weight of product i may reflect more than just changes in the quantity of product i, neither the one-stage Bennet decomposition nor the one-stage Bennet decomposition with extended weight identifies the contributions to the change in the weighted mean that are due to changes in quantities.

2.3. The Two-Stage Bennet Decomposition

To create a decomposition that identifies the contributions to the change in the weighted mean from changes in quantities, we apply the Bennet decomposition also in a second stage. Using the weights $S_{it} = \frac{X_{it}}{\sum_{j=1}^{N} X_{jt}}$ and defining $Q_t = \sum_{j=1}^{N} X_{jt}$, the expression for the quantity variable X_{it} can be written as:

$$X_{it} = S_{it} Q_t, \tag{4}$$

with the change defined as $\Delta X_i = X_{it} - X_{iv}$. Although S_{it} and Q_t cannot vary independently, we apply the Bennet decomposition to the change in the quantity variable in Equation (4), which yields the identity:

$$\Delta X_i = \bar{S}_i \Delta Q + \bar{Q} \Delta S_i, \tag{5}$$

where $\Delta Q = Q_t - Q_v$ and $\bar{Q} = (Q_t + Q_v)/2$. Solving Equation (5) for ΔS_i gives:

$$\Delta S_i = \frac{1}{\bar{Q}} \left(\Delta X_i - \bar{S}_i \Delta Q \right) \tag{6}$$

Inserting Equation (6) into Equation (2) and collecting terms yields the following exact decomposition of the change in the weighted mean:

Proposition 1 (Two-Stage Bennet Decomposition of the Change in the Weighted Mean) Consider the weighted mean across units i at time t of the form: $P_t = \sum_{i=1}^{N} S_{it} P_{it}$, with weights $S_{it} = \frac{X_{it}}{\sum_{j=1}^{N} X_{jt}}$, where $X_{it} \geq 0$ and $Q_t = \sum_{j=1}^{N} X_{jt} > 0$. The change in the weighted mean between times t and v can be exactly decomposed as

$$\Delta P = \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} \left(\frac{1}{\bar{\bar{Q}}} \right) \left(\bar{P}_i - \bar{\bar{P}} \right) \Delta X_i \tag{7}$$

where $\bar{\bar{P}} = \sum_{i=1}^{N} \bar{S}_i \bar{P}_i$, Δ is the difference operator and a bar over a variable represents the moving average operator between times t and v, that is, $\Delta x = x_t - x_v$ and $\bar{x} = 1/2 \, (x_t + x_v)$.

Two features of the two-stage Bennet decomposition in Proposition 1 merit attention. First, the term that shows the contribution to the change in the weighted mean from the change in the quantity of product i is given by

$$\left(\frac{1}{\bar{\bar{Q}}} \right) \left(\bar{P}_i - \bar{\bar{P}} \right) \Delta X_i. \tag{8}$$

This term differs from that in the one-stage Bennet decomposition. It has a natural interpretation and captures the intuitive property that the weighted mean price level increases if products that are growing in quantity have price levels that are higher than the mean price level. $\bar{P}_i - \bar{\bar{P}}$ compares the price level of product i with a measure of the weighted mean price level $\bar{\bar{P}} = \sum_{i=1}^{N} \bar{S}_i \bar{P}_i$. There are thus two ways in which the compositional effects of product i can equal zero: the price of product i equals the weighted average price level, and/or there is no change in the quantity of product i.

Second, the two-stage Bennet decomposition does not hold a time subscript. In other words, the framework is invariant with respect to treatment of time and it therefore satisfies the difference counterpart to the index number time reversal test. The time reversal test for indices states that if the data for the two time periods are interchanged, then the resulting formula should equal the reciprocal of the original index; see, for example, ILO et al. (2004, 295). This test can be rephrased in the case where the formula is in the form of differences, such as the decomposition in Proposition 1: if the data for the two time periods are interchanged, then the resulting formula should equal the negative of the original formula. To illustrate this analytically, let the function $H(\mathbf{P}_t, \mathbf{P}_v, \mathbf{X}_t, \mathbf{X}_v)$ represent the formula for decomposing the change in the weighted mean, where $\mathbf{P}_t = (\mathbf{P}_{1t}, \mathbf{P}_{2t}, \ldots, \mathbf{P}_{Nt})$ and $\mathbf{X}_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})$. The function H passes the time reversal test if and only if $H(\mathbf{P}_t, \mathbf{P}_v, \mathbf{X}_t, \mathbf{X}_v) = -H(\mathbf{P}_v, \mathbf{P}_t, \mathbf{X}_v, \mathbf{X}_t)$. The proposed decomposition in Proposition 1 satisfies this counterpart to the time reversal test.

We commented above on the practice in the literature of choosing a scalar A when decomposing the change in the weighted mean, see Equation (3). Although the choice of A is arbitrary, it is nevertheless interesting to see whether it is possible to derive a value for A that is consistent with the two-stage Bennet decomposition. From Equation (3), the contribution to the change in the weighted mean due to a change in the quantity share of product i is given by the term $(\bar{P}_{it} - A)\Delta S_{it}$. In the two-stage Bennet decomposition, the contribution to the change in the weighted mean due to a change in the quantity of product i is given by the term $\left(\frac{1}{\bar{Q}}\right)(\bar{P}_{it} - \bar{\bar{P}}_t)\Delta X_i$. For these terms be equal, the scalar A must be given by:

$$A = \bar{\bar{P}} - \left(\frac{\Delta Q/\bar{Q}}{\Delta S_i/\bar{S}_i}\right)(\bar{P}_{it} - \bar{\bar{P}}) \tag{9}$$

However, the right-hand side of the equality sign is usually not a constant. This feature stands in contrast to Equation (3), where the property that A is a scalar and independent of i is central to deriving Equation (3) from Equation (2). In the case where the aggregate quantity is unchanged, that is, $\Delta Q = 0$, Equation (9) reduces to $A = \bar{\bar{P}}$, which is independent of i. Moreover, in this case the value of A is close to the choices commonly used in the literature. Several values for the scalar A have been applied, most frequently P_t, P_v and the average of the two, all of which are close to the average measure $\bar{\bar{P}}_t$. However, when the aggregate quantity changes, $\Delta Q \neq 0$, the factor $\left(\frac{\Delta Q/\bar{Q}}{\Delta S_i/\bar{S}_i}\right)$ may differ from zero, possibly leaving a sizable discrepancy between the two-stage Bennet decomposition and most common decompositions used in the literature. In particular, and as can be seen from Equation (5), ΔX_i may have the opposite sign to ΔS_i, depending on how much the aggregate quantity (Q) changes. As a result, the measured contributions from compositional effects in Equation (3) and the two-stage Bennet decomposition may

have opposite signs. In the empirical section, we examine in depth how large the discrepancy between the two decompositions may be in practice when aggregate earnings growth in Norway is decomposed. Before we embark on the empirical application, we first examine an alternative way to decompose the change in the weighted mean that has many similarities with the two-stage Bennet decomposition.

2.4. An Approximate Decomposition of the Change in the Weighted Mean

An alternative way to decompose the change in the weighted mean is to apply the quadratic approximation lemma (QAL) which, loosely defined, states that the average of two first order approximations is equivalent to a second order approximation. According to Theil (1975, 38), the quadratic approximation lemma provides an "approximation which is as simple as the linear approximation and as accurate as the quadratic approximation", where the term "quadratic approximation" refers to a second-order Taylor approximation, see also Diewert (2002) and references therein for applications of QAL.

Consider the weighted mean as a non-linear function of underlying prices and quantities, that is, $P_t = f\left(\mathbf{P}_t, \mathbf{X}_t\right)$, see also Von Brasch et al. (2017). The first-order Taylor series approximation around the initial period (period v) values for the price and quantity variables can be expressed as:

$$\Delta^v \approx \sum_{i=1}^{N} S_{iv}\Delta P_i + \sum_{i=1}^{N}\left(\frac{1}{Q_v}\right)(P_{iv} - P_v)\Delta X_i \tag{10}$$

The first-order Taylor series approximation around the end period (period t) values for the price and quantity variables can be expressed as:

$$\Delta^t \approx \sum_{i=1}^{N} S_{it}\Delta P_i + \sum_{i=1}^{N}\left(\frac{1}{Q_t}\right)(P_{it} - P_t)\Delta X_i \tag{11}$$

Applying QAL to the weighted mean, that is, taking the arithmetic average of the two approximations in Equation (10) and Equation (11), yields:

$$\Delta P \approx \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} \frac{1}{2}\left[\frac{1}{Q_t}(P_{it} - P_t) + \frac{1}{Q_v}(P_{iv} - P_v)\right]\Delta X_i \tag{12}$$

Like the two-stage Bennet decomposition, the decomposition based on QAL also holds the intuitive property that the weighted mean price level increases if products that are growing in quantity have price levels that are higher than the mean price level. However, in contrast to the two-stage Bennet decomposition, the prices are demeaned separately for each period and separately divided by their respective Q_t.

The decomposition based on QAL in Equation (12) is not exact. The size of the approximation error, measured by the difference between the change in the weighted mean and the right-hand side of Equation (12), is given by:

$$\sum_{i=1}^{N} \frac{1}{2}\left[\frac{1}{Q_t}P_{it} - \frac{1}{Q_v}P_{iv}\right]\Delta S_i \Delta Q \tag{13}$$

The smaller the change in Q, the smaller the approximation error. If either all weights or all prices are unchanged, the approximation error is zero. In the empirical application we consider the size of this approximation error and compare the decomposition based on QAL with the two-stage Bennet decomposition.

3. Empirical Application

The data used in the empirical application are obtained through the "a-ordning", which is a collaborative digital system shared by Statistics Norway, The Norwegian Tax Administration and the Norwegian Labour and Welfare Administration (NAV). It provides information about employment, remuneration in cash and in kind and taxes (The Norwegian Tax Administration 2022). Data for all industries and individuals are compiled monthly, and this is the main source Statistics Norway utilizes for producing statistics on earnings and the labour market.

We focus on the change in monthly earnings per full-time equivalent as the price variable from 2020Q4 to 2021Q4 and allow for compositional effects across industries using the number of full-time equivalents in each industry as the quantity variable. Table 1 shows the mean monthly earnings and the number of full-time equivalents in each industry and in the aggregate for 2020Q4 and 2021Q4.

Table 2 shows the results from using the one-stage Bennet decomposition in Equation (2), the one-stage Bennet decomposition with extended weight in Equation (3), the approximation in Equation (12) and our proposed two-stage Bennet decomposition in Proposition 1. As expected, the contribution to the change in the weighted mean from the change in earnings of each industry (and the aggregate) is identical across the four decompositions, as is the total compositional effect. We find that the wedge between the identified compositional effects from (1) the two-stage Bennet decomposition and (2) the one-stage Bennet decomposition and one-stage Bennet decomposition with extended weight is considerable, and for some industries, such as mining and quarrying and wholesale and retail trade, the compositional effects are of opposite signs. Such divergence in the signs of the compositional effects between the methods can be attributed to the changes in the share and quantity variables being of opposite signs (see Figure 2). Furthermore, the identified compositional effects in the one-stage Bennet decomposition generally have a greater absolute value than the compositional effects identified using the one-stage Bennet decomposition with extended weight and the two-stage Bennet decomposition. This is attributable to the use of $\bar{P}_i$ in the one-stage Bennet decomposition, compared to the use of the relative earnings level in the other two decompositions $\left(\left(\bar{P}_i - A\right)\right.$ and $\left.\left(\bar{P}_i - \bar{\bar{P}}\right)\right)$. The intuition for why the one-stage Bennet decomposition with extended weight and the two-stage Bennet decomposition yield somewhat similar results can be seen by rewriting the two decompositions in a manner that is easier to compare. The one-stage Bennet decomposition with extended weight may be written as

$$\Delta P = \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} \left(\bar{P}_i - \bar{\bar{P}}\right) \left[\frac{X_{it}}{\Sigma_{j=1}^{N} X_{jt}} - \frac{X_{iv}}{\Sigma_{j=1}^{N} X_{jv}} \right]$$

Table 1. Monthly earnings per full-time equivalent and number of full-time equivalents, 2020Q4 and 2021Q4.

	2020Q4		2021Q4	
	Monthly earnings (NOK)	Number of full-time equivalents	Monthly earnings (NOK)	Number of full-time equivalents
All industries	48,750	2,242,706	50,790	2,320,214
Agriculture, forestry and fishing	41,880	24,838	43,830	25,664
Mining and quarrying	74,290	59,890	76,310	61,292
Manufacturing	49,090	197,800	50,940	201,393
Electricity, water supply, sewerage, waste management	54,450	32,012	57,430	32,846
Construction	46,110	221,718	47,730	228,745
Wholesale and retail trade; repair of motor vehicles and motorcycles	44,220	262,080	46,320	267,773
Transportation and storage	47,860	110,862	49,350	114,705
Accommodation and food service activities	33,340	60,593	34,380	72,623
Information and communication	63,270	93,626	65,980	100,882
Financial and insurance activities	70,360	45,928	73,670	46,890
Real estate, professional, scientific and technical activities	60,060	146,836	63,460	155,418
Administrative and support service activities	41,000	115,691	42,760	122,725
Public administration and defence; compulsory social security	52,210	157,396	54,240	159,831
Education	46,750	195,500	48,710	197,832
Human health and social work activities	44,860	446,720	47,010	455,729
Other service activities	43,830	70,212	45,360	74,466
Unspecified	63,440	1,003	68,490	1,403

Source: Statbank Table 11419, Statistics Norway.

In comparison, the two-stage Bennet decomposition may be written as

$$\Delta P = \sum_{i=1}^{N} \bar{S}_i \Delta P_i + \sum_{i=1}^{N} \left(\bar{P}_i - \bar{\bar{P}}\right) \left[\frac{X_{it}}{\left(\Sigma_{j=1}^{N} X_{jt} + \Sigma_{j=1}^{N} X_{jv}\right)/2} - \frac{X_{iv}}{\left(\Sigma_{j=1}^{N} X_{jt} + \Sigma_{j=1}^{N} X_{jv}\right)/2} \right]$$

The difference between the contributions to the overall compositional effects in these two decompositions is thus caused by changes in quantities (full-time equivalents), not changes in prices. The compositional effects from each industry, identified using each of the four different decompositions, are illustrated in Figure 1. An add-in for carrying out

Table 2. Decomposition of change in monthly earnings (NOK), from 2020Q4 to 2021Q4.

Industry	All decompositions	One-stage Bennet		One-stage Bennet with extended weight		Approximation		Two-stage Bennet	
	Earnings contribution	Compositional effect	Total	Compositional effect	Total	Compositional effect	Total	Compositional effect	Total
All industries	2,074	-33	2,041	-33	2,041	-33	2,041	-33	2,041
Agriculture, forestry and fishing	22	-1	21	0	22	-3	19	-3	19
Mining and quarrying	54	-22	32	-7	47	16	69	16	69
Manufacturing	162	-70	92	1	163	0	162	0	162
Electricity, water supply, sewerage, waste management	42	-7	36	-1	42	2	45	2	45
Construction	160	-13	147	1	161	-9	151	-9	151
Wholesale and retail trade: repair of motor vehicles and motorcycles	244	-66	178	8	252	-11	233	-11	233
Transportation and storage	74	0	74	0	74	-2	72	-2	72
Accommodation and food service activities	30	145	175	-73	-42	-84	-54	-84	-54
Information and communication	115	112	227	24	139	47	163	47	163

Table 2. Continued

Industry	All decompositions	One-stage Bennet		One-stage Bennet with extended weight		Approximation		Two-stage Bennet	
	Earnings contribution	Compositional effect	Total	Compositional effect	Total	Compositional effect	Total	Compositional effect	Total
Financial and insurance activities	67	-19	48	-6	62	9	77	9	77
Real estate, professional, scientific and technical activities	225	93	319	17	242	45	270	45	270
Administrative and support service activities	92	55	147	-12	80	-24	68	-24	68
Public adm., defence, soc. security	141	-69	72	-3	138	4	145	4	145
Education	169	-91	78	6	175	-2	167	-2	167
Human health and social work activities	425	-127	298	13	439	-15	410	-15	410
Other service activities	49	35	84	-5	44	-10	39	-10	39
Unspecified	3	10	13	2	5	3	5	3	5

Source: Authors' own calculations using data from Statistics Norway.

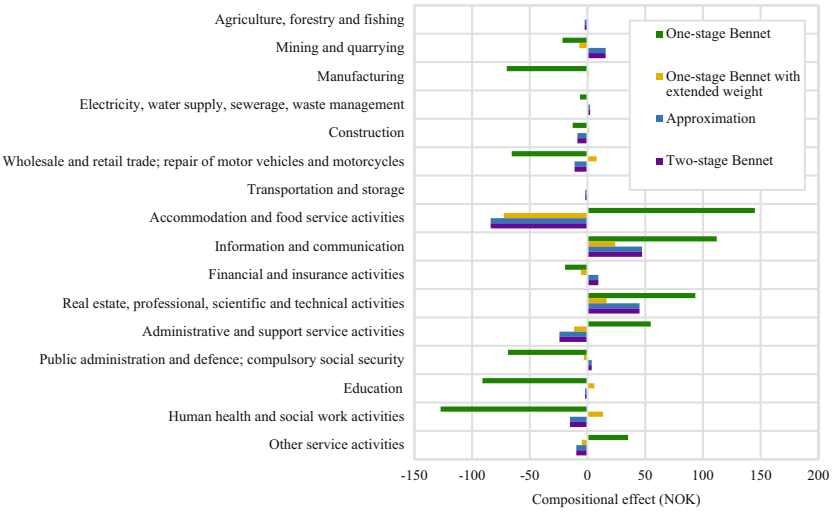


Fig. 1. Compositional effects across decompositions.
Note: See Table 2 for precise magnitudes of compositional effects for each industry and decomposition method. Note that "Unspecified" from Table 2 is not shown in this figure. Source: Authors' own calculations using data from Statistics Norway.

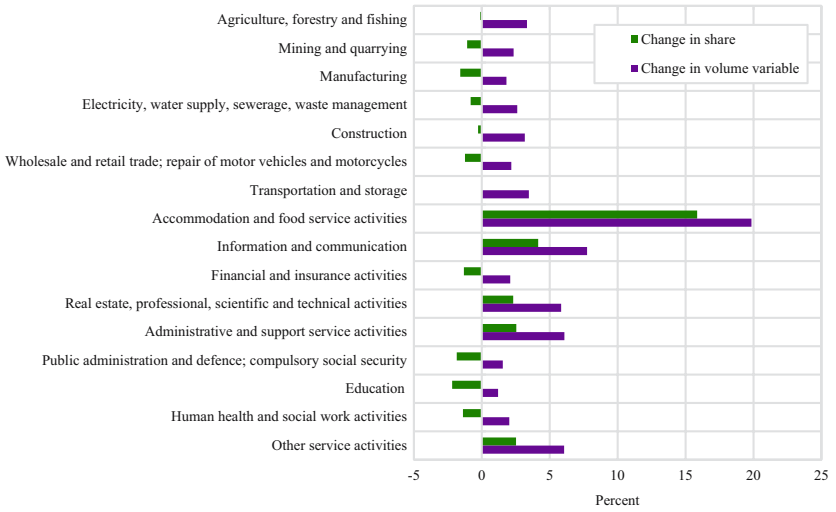


Fig. 2. Change in share and quantity variable, from 2020Q4 to 2021Q4.
Note: Change in share and quantity variable for each industry from 2020Q4 to 2021Q4, measured in percent. Note that "Unspecified" from Table 2 is not shown in this figure. Source: Authors' own calculations using data from Statistics Norway.

this decomposition in EViews, and a Stata replication code to generate the results in Table 2 and Figure 1, are available from the authors upon request.

Focusing on our proposed decomposition, the above results illustrate that there were both positive and negative compositional effects present in aggregate earnings growth in

Norway from 2020Q4 to 2021Q4. In the aggregate, these effects were negative, which is largely attributable to developments in the industry accommodation and food service activities. This industry was to a considerable extent impacted by the Norwegian government's actions to curb the spread of the coronavirus, as these measures particularly affected industries where social interaction is a necessity. However, the gradual softening of these measures throughout 2021 was accompanied by an increase in activity in these industries. As the industry accommodation and food service activities has a level of earnings lower than the weighted mean earnings level (see Table 1), the compositional effect becomes negative when the quantity variable of the industry increases. Conversely, the industries information and communication and real estate, professional, scientific and technical activities, in which the earnings level is higher than the aggregate earnings level, contributed to aggregate earnings growth with a noteworthy positive compositional effect, as these industries had an increase in the quantity variable. At the same time, several industries had a compositional effect that was close to zero.

Comparing the results of our proposed decomposition with the results produced by the other two exact methods, a few comments are warranted. First, the proposed decomposition provides an exact and intuitive description of the contribution to the overall compositional effect from each industry, in terms of both the size and the direction of the contribution. For example, Table 2 shows a large difference between the one-stage Bennet decomposition on the one hand and the one-stage Bennet decomposition with extended weight and two-stage Bennet decomposition on the other in the contribution from accommodation and food service activities. The compositional effect is far greater with the one-stage Bennet decomposition, and it has an opposite sign from the other two. Furthermore, in the industries mining and quarrying and wholesale and retail trade, amongst others, the two-stage Bennet decomposition and the one-stage Bennet decomposition with extended weight produce compositional effects with opposite signs, and in these cases, we argue that the two-stage Bennet decomposition provides a more intuitive result. For mining and quarrying, where the average earnings are higher than the average earnings for all industries, as shown in Table 1, and the quantity of full-time equivalents increases from 2020Q4 to 2021Q4, it is natural that the contribution from compositional effects should be positive, as the two-stage Bennet decomposition shows. The comparison yields the same conclusion for wholesale and retail trade, but with opposite signs. With a lower level of earnings than the aggregate earnings level and an increase in the number of full-time equivalents, it is apparent that this industry, all else being equal, should contribute with a negative compositional effect, just as provided by the two-stage Bennet decomposition. Figure 1 also shows the approximation from the decomposition based on QAL in Equation (12), which is very similar to the two-stage Bennet decomposition. The approximation error is small, at 0.03 percent relative to the size of the compositional effect from the two-stage Bennet decomposition, due to modest changes in the aggregate quantity, as well as the prices and weights (cf. Equation (13)). It should be noted that the choice of dimension along which the change in the weighted mean is decomposed is by no means arbitrary. When choosing this dimension, it is critical that it inherently captures the nature of the changes in the quantity variable.

4. Conclusion

In this article, we have derived an exact additive decomposition of the change in the weighted mean. Our proposed two-stage Bennet decomposition is easy to employ and interpret. We also show that it satisfies the difference counterpart to the index number time reversal test. The fundamental difference between our proposed decomposition and many of the decompositions used in the literature is that our measure of the contribution to compositional changes of a given product is based on the change in the quantity of that product. If there is no change in the quantity of a product, then that product does not contribute to a compositional change in the weighted mean. In contrast, in other decompositions, such as the one-stage Bennet decomposition, the measure of the contribution to compositional changes of a given product is based on the change in the quantity share of that product. Since the quantity share of a product may change because the quantities of other products change, this may lead to compositional changes stemming from a product whose quantity level is unchanged. We have also outlined an alternative decomposition based on a quadratic approximation of the weighted mean, when the weighted mean is regarded as a non-linear function of underlying prices and quantities. Although this alternative decomposition is not exact, it has some similar features to the two-stage Bennet decomposition.

When comparing our proposed decomposition to the standard decomposition used in the literature in the case of aggregate earnings growth in Norway from 2020Q4 to 2021Q4, we find that the wedge between the identified compositional effects is substantial, and for some industries the compositional effects are of opposite signs. We also compared the two-stage Bennet decomposition with the decomposition based on a quadratic approximation and found that these two decompositions generated similar group-specific contributions to the overall compositional effects.

5. References

Baily, M.N., C. Hulten, D. Campbell, T. Bresnehan, and R.E. Caves. 1992. "Productivity dynamics in manufacturing plants." *Brookings Papers on Economic Activity: Microeconomics*: 187–267. DOI: https://doi.org/10.2307/2534764.

Balk, B.M. 2021. *Productivity: Concepts, Measurement, Aggregation, and Decomposition.* Springer.

Benedictow, A., and P. Boug. 2017. "Calculating the real return on a sovereign wealth fund." *Canadian Journal of Economics* 50(2): 571–594. DOI: https://doi.org/10.1111/-caje.12270.

Benedictow, A., and P. Boug. 2021. "Exact and inexact decompositions of trade price indices." *Empirical Economics*. DOI: https://doi.org/10.1007/s00181-021-02078-4.

Bennet, T.L. 1920. "The Measurement of Changes in the Cost of Living." *Journal of the Royal Statistical Society* 83(3): 455–462. DOI: https://doi.org/10.2307/2340777.

Daly, M.C., and B. Hobijn. 2017. "Composition and Aggregate Real Wage Growth." *American Economic Review* 107(5): 349–352. DOI: https://doi.org/10.1257/aer.p20171075.

Diewert, W.E. 2002. "The quadratic approximation lemma and decompositions of superlative indexes." *Journal of Economic and Social Measurement* 28: 63–88.

Available at: https://ip.ios.semcs.net/articles/journal-of-economic-and-social-measurement/jem00200.

Diewert, W.E. 2005. "Index number theory using differences rather than ratios." *American Journal of Economics and Sociology* 64(1): 311–360. DOI: https://doi.org/10.1111/j.1536-7150.2005.00365.x.

Diewert, W.E., and K.J. Fox. 2010. "On measuring the contribution of entering and exiting firms to aggregate productivity growth. In *Price and Productivity Measurement*", edited by W.E. Diewert, B. Balk, D. Fixler, K.J. Fox, and A. Nakamura. 6: 41–66. Trafford Press.

Foster, L., J.C. Haltiwanger, and C.J. Krizan. 2001. "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." In *Developments in Productivity Analysis*, edited by C.R. Hulten, E.R. Dean, and M.J. Harper: 303–372. University of Chicago Press.

Huerga, J. 2010. "An Application of Index Numbers Theory to Interest Rates." In *Price Indexes in Time and Space: Methods and Practice*, edited by L. Biggeri, and G. Ferrari: 239–248. DOI: https://doi.org/10.1007/978-3-7908-2140-6_13.

ILO, IMF, OECD, Eurostat, United Nations, and World Bank. 2004. *Consumer price index manual – Theory and practice*. I.L. Office, Geneva. Available at: https://www.ilo.org/global/statistics-and-databases/publications/WCMS_331153/lang–en/index.htm.

Kamin, S.B., M. Marazzi, and J.W. Schindler. 2006. "The Impact of Chinese Exports on Global Import Prices." *Review of International Economics* 14(2): 179–201. DOI: https://doi.org/10.1111/j.1467-9396.2006.00569.x.

Krueger, A.B. 2017. "Where have all the workers gone? An inquiry into the decline of the U.S. Labor force participation rate." *Brookings Papers on Economic Activity* 2017*:* 1–87. DOI: https://doi.org/10.1353/eca.2017.0012.

OECD. 2018. *OECD Compendium of Productivity Indicators 2018*. OECD. DOI: https://doi.org/10.1787/pdtvy-2018-en.

Segnitz, E. 1870. "Über die Berechnung der sogenannten Mittel, sowie deren Anwendung in der Statistik und anderen Erfahrungswissenschaften." Jahrbücher für Nationalökonomie und Statistik, 14: 183–195.

The Norwegian Tax Administration. 2022. *About a-ordningen*. Available at: https://www.skatteetaten.no/en/business-and-organisation/employer/the-a-melding/-about-the-a-ordning/about-a-ordningen/ (accessed March 2022).

Theil, H. 1975. *Theory and measurement of consumer demand*. Vol(1), North-Holland Amsterdam.

Thomas, C.P., and J. Marquez, J. 2009. "Measurement matters for modelling US import prices." *International Journal of Finance and Economics* 14(2): 120–138. DOI: https://doi.org/10.1002/ijfe.370.

Von Brasch, T., B. Dapi, and V. Sparrman. 2017. *Sammensetningseffekter mellom næringer og veksten i gjennomsnittlig årslønn (Compositional effects and growth in annual earnings):* 2017/45. Statistics Norway. Available at: https://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/sammensetningseffekter-mellom-naeringer-og-veksten-i-gjennomsnittlig-arslonn.

Received October 2021
Revised April 2022
Accepted June 2022