Does Field Substitution Impact the Educational Profile of the Belgian Health Interview Survey Net Sample?

Stefaan Demarest (stefaan.demarest@sciensano.be)
Sciensano https://orcid.org/0000-0003-2823-1372
Finaba Berete
WIV: Sciensano
Youri Baeyens
Statbel
Geert Molenberghs
University of Hasselt
Sabine Drieskens
Sciensano
Rana Charafeddine
Sciensano
Elise Braekman
Sciensano
Herman Van Oyen
Sciensano
Guido Van Hal
University of Antwerp

Research

Keywords: health surveys, participation, socio-economic differences, field substitution

Posted Date: November 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-100731/v1

Abstract

Background Although controversial as a sampling technique, field substitution of non-respondents is applied in the Belgian Health Interview (BHIS) since its start in 1997. The target number of participants to obtain is predefined and set at 10,000 individuals. Based on data derived from the National Register, non-participating households are substituted by at most three households matched on statistical sector, age group of the households' reference person (administrative contact of the household) and household size, thus creating a cluster. In this study, the impact of field substitution on the educational composition of the net sample is assessed.

Methods The educational level of the household' reference person derived from the Census 2001 and Census 2011, was used as a proxy for socio-economic position and was linked with respectively BHIS 2001 and BHIS 2013 paradata on the use of field substitution using a unique identifier. Given the high level of missing data on the educational level (+/-16%) in the Census, regression based multiple imputations (m=5) procedures were applied, presuming missingness at random. Response rates by educational level at any stage of the substitution process stage were calculated. Differences in response rates were assessed by applying the Delta method.

Results At any stage of the substitution process, the participation rate was the lowest in the lowest educated households and significantly higher in the middle and highest educated households. Throughout the substitution process, the participation rate dropped from 51.6% to 42.7% for low educated households and from 61.7% to 46.3% for high educated households.

Conclusions It is concluded that field substitution introduces higher levels of non-participation but does not affect the educational composition of the net sample.

Background

The Belgian Health Interview Survey (BHIS), commissioned by the federal and regional authorities in Belgium, is a cross-sectional survey for which the net sample size to be obtained is fixed: 10,000 individuals have to be interviewed throughout the data collection phase of one calendar year. The BHIS is a household survey: households are invited to participate and at most four household members are interviewed, a process that continues until the predefined number of individuals participate. The BHIS is one of the rare health surveys that applies field substitution during data collection: non-participating households are substituted – if needed several times – by similar households. The National Register (NR) is used as the sampling frame and the matching of the initial household and the substitute-households is based on information available in the Register: the age of the households' reference person, the size of the household and the statistical sector of the residence (the finest geographical area for which statistical data are available). Field substitution is preferred since applying it assures that the net sample meets the size of the target sample (1–3) and, at the same time, possible non-response bias in the survey estimates is attenuated (4–13). Although often contested as a sampling technique (3, 6, 10, 14, 15), it has been shown to be an acceptable approach to maintain the sample composition but only if uses a rich sampling

frame that enables a relevant matching and if interviewers do not impact the decision to substitute non respondents (1, 14, 16) In the BHIS clusters of four matched households are created. After randomizing the initial order of the households within a cluster, the households positioned at the first place in the cluster are invited to participate in the BHIS. In case of non-participation, the first substitute household is invited to participate. As long as no household in the cluster participates, the substitution process continues. In case none of the households in the cluster participates, the cluster is substituted with another cluster and the substitution process continues. Further details on the substitution process in the BHIS can be found elsewhere (17, 18).

Applying field substitution, as described above, ensures that the requested net sample in terms of both size and socio-demographic composition is reached. It remains uncertain if the socio-economic composition of the net sample reflects adequately the composition of the initial sample. Several studies have demonstrated that participation in health surveys is influenced by the socio-economic characteristics of the units invited for participation. The general conclusion is that sampling units with a lower socio-economic position (measured either by the educational level, income level and/or professional level) are harder to reach and more reluctant to participate in the survey (19–24). Given selective non-participation and the difference in health by socioeconomic position, generalization of the survey estimates on population level can be biased. The finding that units of a lower socio-economic position are underrepresented in a health survey net sample is mostly based on auxiliary (administrative) data sources like register information on health care utilization in hospitals and primary care (20, 26), personnel registers (27) or population censuses (28).

The NR used as the sampling frame for the BHIS does not provide any useful information on the socio-economic position of the civilians (19, 29, 30). Only by comparing aggregated data of the BHIS 2001 with similar data derived from the Belgian Census 2001, Lorant et al. were able to observe that low educated and not working individuals were less likely to participate in the BHIS 2001 when they had a poor health status. As a consequence, in the BHIS the association between poor health status and low SES is likely to be underestimated (31). After one-to-one linking of the BHIS 2001 and the Census 2001 data, it was concluded that, amongst contactable households, the participation rate was significantly lower in households with a lower educational level (19).

To assess the impact of field substitution on the socio-economic composition of the net sample, it is essential to know the socio-economic profile of households initially selected for participation in the BHIS, and of all activated substitute households. In this study, para-data related to the substitution applied in the BHIS 2013 is linked with data on the educational level (highest educational attainment) of the household's reference person derived from the Administrative Census 2011. To gain power, a similar linkage was done with the BHIS 2001 para-data and data derived from the Census 2001. The research question is as follows: does the substitution process impact the educational composition of the BHIS net sample?

Data And Methods

For the BHIS 2013 the province of Luxembourg financed an oversampling of 600 individuals. Consequently, the total net sample size to be obtained was set at 10,600 individuals. For the BHIS 2001, several provinces financed an oversampling, which resulted in a required net sample of 12,050 participating individuals. As the BHIS is a household survey, household reference persons from the NR were selected. A multistage stratified sampling design including several sampling techniques; stratification, clustering, systematic sampling and simple random sampling, was used. Municipalities served as Primary Sampling Units (PSUs) and were selected through a systematic sampling method with a selection probability proportional to their size. In each selected municipality one or more groups of 50 individuals had to be interviewed throughout the calendar year, divided into 4 trimesters (that is; per trimester around 12–13 individuals in every group had to be interviewed). A systematic sampling method was used to select households, the Secondary Sampling Units (SSUs). The number of selected households in one group corresponds with 50 individuals. Finally, at most four individuals – the Tertiary Sampling Units (TSUs) – were selected for the interviews within each household: by default, the reference person and his/her partner (if any) and two (or three) remaining household members using random selection (29).

Field substitution was applied at the level of the SSUs: for each household, three substitute households matched on statistical sector, age group of the households' reference person (administrative contact of the household) and household size were selected in order to create clusters of four households. Once the clusters were created, an ad-random scrambling was applied in order to identify the initial household to be contacted and the potential first, second and third substitute households. For every cluster, an unmatched substitute cluster was identified with the potential fourth till seventh substitute household. Also within a substitute cluster, households had similar characteristics, but there was no match with the characteristics of the households in the original cluster. While in the BHIS 2013 cluster substitution was unconditional (once a cluster was exhausted, the first household of the substitute cluster was activated), in the BHIS 2001 cluster substitution was conditional to the number of participating individuals which were part of the initial cluster of households.

The fieldwork of the BHIS was spread over a whole calendar year and samples were taken each quarter about 6 weeks before the start of the quarter. An online verification of the vital status of the reference person was performed a few days before the start of the data collection. Reference persons with a change of vital status (e.g. died, moved abroad) and their corresponding household were removed from the sample. The order of the remaining households within the clusters was adapted accordingly. Consequently, a very limited number of clusters counted less than four households. For every selected household and for every selected household member a unique identifier was created. An algorithm was developed to assure the conversion from this identifier to the corresponding number in the NR and was entrusted to a Trusted Third Party (TTP); Statistics Belgium (Statbel).

Para-data obtained during the data collection of BHIS were used to assess the practice of substitution throughout the fieldwork phase. For every activated household (that is; for every household that was effectively invited to participate in the BHIS), interviewers had to document the date, hour and mode (by

telephone or at doorstep) of every contact-attempt. At least five contact-attempts, of which at least one at doorstep had to be made before a household could be labelled as non-contactable. In case a household was not contactable or refused to participate, a substitute household was activated by the central administration of the survey. The same contact procedure as for initial activated households is used for substitute households. This para-data enabled a strict follow-up of the fieldwork phase and enabled to assess whether the activation of a substitute household was justified.

Data on the educational level of the household's reference person were derived from the Administrative Census 2011 and the Census 2001. The Census 2001 (officially entitled the "General Socio-Economic Survey 2001") was the last 'traditional' census based on an exhaustive postal survey among households. Participation to the census was compulsory and resulted in a participation rate at household level of 96.5%. Questions on the educational level had to be completed for every household member aged 15 years and older.

The Census 2011 was based on linked administrative databases and covers, among other, data on the educational level (highest diploma) provided by the Belgian communities responsible for the organization of education. For what concerns the highest level of education, the Census 2011 was an update of the data collected in the context of the Census 2001. For those who obtained a (registered) diploma in the period 2001–2011 that was higher than the one declared during the Census 2001, the highest educational level was adopted.

A first assessment of the data completeness revealed relatively high levels of item-missingness for educational level in the Census databases (e.g. for 16.1% of all reference persons sampled for the BHIS2013 information on the educational level was missing in the Census 2011). Since complete data is an absolute prerequisite to assess the substitution process, regression based multiple imputations (m = 5) procedures were applied, presuming missingness at random (MAR). Variables added to the model were gender, age group and household size. An analysis of cluster homogeneity/heterogeneity in terms of educational level showed a very high level of cluster heterogeneity; only in 13.2% of all clusters the reference persons of the four households had an identical educational level.

After having received the permission of the Belgian Privacy Commission, the BHIS 2013 sample data were one-to-one linked with the BHIS 2013 para-data and the data on the educational level, derived from the Census 2011. The BHIS 2001 sample data were equivalently linked with the BHIS 2001 para-data and the data of the Census 2001. The reference persons highest achieved educational level, stored in the Census databases in 6 categories according to the International Classification of Education (ISCED) (32), was regrouped in three categories: low educational level (no diploma, primary education (ISCED 1) and lower secondary education (ISCED 2)), middle educational level (higher secondary education (ISCED 3) and post-secondary non-higher education (ISCED 4)) and high educational level (bachelor and master (ISCED 5) and doctorate (ISCED 6)).

All households effectively invited to participate in respectively the BHIS 2013 and the BHIS 2001 were re-ordered in terms of the original clusters, that is; in terms of initially selected households, first substitutes,

second substitutes, etc. For each substitution wave, the response rate according to the educational level of the households' reference person was calculated. Given their limited number of cases, the fourth till seventh substitutes were grouped. Differences in response rates were assessed by the Delta method using the TEST statement in SAS© PROC MIANALYZE. A sensitivity analysis, taking only households for which the educational level of the households' reference person was known into account, was applied to test the robustness against departures from the MAR assumption.

Results

Table 1 presents an overview of the final status of households invited for participating in the BHIS 2013 and the BHIS 2001. In order to obtain the predefined number of 10,600 successful individual participants to BHIS 2013 9,662 households were activated, of which 5,048 participated (52.3%). In the BHIS 2001 in total 11,231 households were activated of which 5,520 participated (49.1%). The bulk of all participating households are initial activated households: 55.4% for the BHIS 2013 and 60.1% for the BHIS 2001. For both the BHIS 2013 and the BHIS 2001, the participation rate was the highest in the initial activated households (BHIS 2013: 55.5%, BHIS 2001: 53.5%), and dropped systematically throughout the substitution process. For the third substitute households (the last households of the initial clusters), the participation rate was respectively 44.9% for the BHIS 2013 and 40.0% for the BHIS 2001.

Table 1
Overview of the final status of households invited for participating in BHIS 2013 and BHIS 2001, according to substitution wave

	BHIS 2013		BHIS 2001	
Initial activated households	n	%	n	%
Participants	2,795	55.5	3,339	53.5
Refusals	1,871	37.1	1,871	30.0
Non-contactables	236	4.7	558	8.9
Other non-participants(*)	134	2.7	478	7.6
Total activated	5,036	100	6,246	100
1st substitute households				
Participants	1,125	50.2	1,270	46.4
Refusals	892	39.8	894	32.7
Non-contactables	142	6.3	291	10.6
Other non-participants	83	3.4	281	10.3
Total activated	2,242	100	2,736	100
2nd substitute households				
Participants	513	45.8	530	41.4
Refusals	491	43.8	447	35.0
Non-contactables	73	6.5	163	12.7
Other non-participants	44	3.9	139	10.9
Total activated	1,121	100	1,279	100
3th substitute households				
Participants	273	44.9	246	40.0
Refusals	271	44.6	217	35.2
Non-contactables	40	6.6	93	15.1
Other non-participants	24	3.9	60	9.7
Total activated	608	100	616	100
4th – 7th substitute households				

(*) Households with an incorrect address, households not living at the indicated address, households not eligible for interview (prisons, psychiatric institutions, monasteries)

	BHIS 2013		BHIS 2001	
Participants	342	52.2	135	38.1
Refusals	255	38.9	115	32.5
Non-contactables	43	6.6	45	12.7
Other non-participants	15	2.3	59	16.7
Total activated	655	100	354	100
Total activated households				
Participants	5,048	52.3	5,520	49,1
Refusals	3,780	39.1	3,544	31,6
Non-contactables	534	5.5	1,150	10,2
Other non-participants	300	3.1	1,017	9,1
Total activated	9,662	100	11,231	100
(*) Households with an incorrect address, households not living at the indicated address, households not eligible for interview (prisons, psychiatric institutions, monasteries)				

Tables 2 provides an overview of the substitution process by the educational level of the households as applied in in the BHIS 2013. An additional file provides an overview of the substitution process by educational level for the BHIS 2001 [see Additional file 1].

Table 2
Number activated households and participating households, by educational level, substitution wave and participation rate, linked AC2011 – BHIS2013 data

Educ. level HH	# of activated HH (95% CI)	Share activated HH (95% CI)	# of participating HH (95% CI)	Share participating HH (95% CI)	HH participating rate (95% CI)	Diff. low educ. HH	p value for difference
Initial selected households							
Low	2,129 (2,096–2,163)	42.3 (40.8–43.7)	1,099 (1,083–1,115)	39.3 (37.5–41.2)	51.6 (49.3–53.9)	-	-
Middle	1,461 (1,420–1,501)	29.0 (27.6–30.4)	804 (782–827)	28.8 (27.0–30.6)	55.1 (52.4–57.7)	+ 3,5	0.0541 (*)
High	1,446 (1,396–1,496)	28.7 (27.3–30.2)	892 (862–922)	31.9 (30.0–33.8)	61.7 (58.8–64.5)	+ 10,1	< .0001 (**)
1st substitute households							
Low	980 (963–997)	43.7 (41.6–45.8)	455 (426–483)	40.4 (37.0–43.9)	46.4 (42.8–50.1)		
Middle	655 (632–677)	29.2 (27.2–31.2)	327 (313–341)	29.1 (26.2–31.9)	50.0(45.9–54.0)	+ 3.6	0.2325
High	607 (589–625)	27.1 (25.2–29.0)	343 (324–362)	30.5 (27.5–33.5)	56.5 (52.2–60.7)	+ 10.1	0.0020
2nd substitute households							
Low	477 (454–500)	42.6 (39.3–45.8)	190 (174–207)	37.1 (32.2–42.0)	39.9 (35.2–44.8)		
Middle	327 (313–341)	29.2 (26.3–32.0)	147 (139–156)	28.7 (24.6–32.8)	45.0 (39.2–51.0)	+ 5.1	0.2217
High	317 (301–333)	28.3 (25.4–31.1)	175 (166–184)	34.2 (29.9–38.5)	55.3 (49.7–60.8)	+ 15.4	< .0001
3rd substitute households							

* under the hypothesis that the difference between response rates low - middle educated households is equal to zero

** under the hypothesis that the difference between response rates low-high educated households is equal to zero

Educ. level HH	# of activated HH (95% CI)	Share activated HH (95% CI)	# of participating HH (95% CI)	Share participating HH (95% CI)	HH participating rate (95% CI)	Diff. low educ. HH	p value for difference
Low	260 (232–289)	42.8 (37.4–48.2)	111 (103–119)	40.7 (34.6–46.7)	42.7 (36.3–49.3)		
Middle	182 (175–189)	29.9 (26.2–33.6)	85 (75–95)	31.2 (25.1–37.3)	46.9 (38.2–55.8)	+ 4.2	0.4906
High	166 (142–190)	27.3 (22.6–32.1)	77 (64–90)	28.1 (21.7–34.6)	46.3 (38.7–54.0)	+ 3.6	0.4765
4th – 7th substitute households							
Low	264 (249–279)	40.4 (36.2–44.5)	128 (118–137)	37.3 (31.8–42.9)	48.3 (41.6–55.0)		
Middle	187 (181–192)	28.5 (25.0–32.0)	102 (97–107)	29.8 (24.9–34.7)	54.6 (47.3–61.8)	+ 6.3	0.2288
High	204 (192–216)	31.1 (27.3–34.9)	112 (102–122)	32.9 (27.4–38.3)	55.2 (47.6–62.5)	+ 6.9	0.2222
Total activated households							
Low	4,111 (4,046–4,176)	42.5 (41.4–43.7)	1,983 (1,933–2,032)	39.3 (37.7–40.8)	48.2 (46.4–50.1)		
Middle	2,811 (2,757–2,865)	29.1 (28.1–30.1)	1,466 (1,436–1,496)	29.0 (27.7–30.4)	52.2 (50.1–54.2)	+ 4.0	0.0152
High	2,740 (2,662–2,818)	28.4 (27.3–29.5)	1,600 (1,552–1,647)	31.7 (30.2–33.2)	58.4 (56.3–60.5)	+ 10.2	< .0001
* under the hypothesis that the difference between response rates low - middle educated households is equal to zero							
** under the hypothesis that the difference between response rates low-high educated households is equal to zero							

For the BHIS 2013, initially 5,036 households were activated at the start of the data collection phase, of which 42.3% were classified in the low educated group, 29.0% in the middle educated group and 28.7% in the highest educated group. In total 2,795 initial activated households participated (55.5%) (Table 1). In

the lowest educated group, the participation rate is significantly lower (51.6%, 95% CI: 50.9% – 53.7%) than the rate found in the highest educated group (61.7%, 95% CI: 60.0% – 62.7%) ($p < 0.0001$).

In total 2,242 households served as first substitute households. The educational composition of the group of first substitute households resembled to a large extent the one of the initial selected households and did not at all account for the differential participation according to educational level that was found for the initially selected households; 43.7% are classified as low educated, 29.2% as middle educated and 27.1% as high educated. In total 1,125 first substitute households participated, which resulted in a participation rate of 50.2% - a drop of 5.3% in comparison with the corresponding rate in initially selected households. Also for the first substitute households, the participation rate is lower in low educated households (46.4%) than in high educated households (56.5%) ($p = 0.0020$).

The educational profiles of the activated second and third substitute households resembled both the initial selected households and the first substitute households. The overall participation rate drops to 45.8% for the second substitute households and further to 44.9% for the third substitute households. For the second substitute households, the participation rate of low educated households remains significantly lower (39.9%) than in high educated households (55.3%) ($p < 0.0001$). For the third substitute households no significant differences in participation according the educational level could be found.

Given the small numbers, the results for the fourth to seventh substitute households are grouped. It must be noted that these households belong to substitute clusters and are not matched with the households of the initial clusters. Yet, also here the educational composition of the activated households is similar to the composition of the initial selected households. Due to the relatively low number of households involved in this part of the substitution process, no significant differences in participation rates according to the educational level of the households could be found.

The composition of the net sample in terms of educational level deviates from the initial sample in favor of high-educated households. While the high educated households account for 28.4% (95% CI 27.5% – 29.5%) in the initial sample, their share is significantly higher in the participating households (31.4%, 95% CI 30.1% – 32.7%).

In overall terms, the findings for the BHIS 2013 are also applicable for the BHIS 2001: also in this survey, the participation rates for the high-educated households are significantly higher compared to the rates found in lower educated households and this for every stage of the substitution process. The share of high educated households in the net sample (27.5%, 95% CI 26.3% – 28.7%) is significantly higher than in the activated sample (25.2%, 95% CI 24.4% – 26.1%).

Discussion

The National Register, used as the sampling frame for the BHIS, does not include information on the socio-economic position of the citizens. Hence, the socio-economic position could not be used as a matching variable between initial selected and substitute households. In this study, the educational level of the

households' reference person was used as proxy for the socio-economic position of the household. Linking the BHIS 2013 and BHIS 2001 samples with census data, enabled to assess this educational level. Although households of a same cluster were matched on the selection criteria, they showed to be very heterogeneous in terms of their educational level. Consequently, in the substitution process e.g. non-participating low educated households could be substituted by higher educated households and vice versa. It has been shown that for all stages of the substitution process, the participation rate in lower educated households is significantly lower than in higher educated households and that this mainly due to higher refusal rates, a finding in line with other studies (6, 19, 20, 22, 31).

The analysis showed that participation rates drop systematically throughout the substitution process. Critics of the substitution method might state that, when applying this method, interviewers lower their efforts to obtain participation from a household since they know that this household will be substituted (6, 10, 11). Of course, the time and efforts interviewers dedicate to contact households and to convince households to participate, impact participation. Yet, since in the BHIS substitution of households is a prerogative of the central administration and since interviewers are not paid for non-participating households, it is in their interest to maximize their efforts to obtain participation. Shrinking participation throughout the substitution process might be associated with the criteria used to construct the clusters of households; statistical sector, age group of the reference person and size of the household. Exit meetings with the BHIS interviewers indicated that some residential areas (statistical sectors) are very hard to survey, especially areas with huge building blocks where the first at doorstep contacts with the selected households are done by intercom.

Compared to the educational profile of the initial participating households, no differences can be observed with the educational profile of all participating households. This can partially be explained by the fact that a small majority of all participating households are initial participating households. Notwithstanding (a) the differential participation rates by educational level, with higher participation rates for the higher educated households and (b) overall declining participation rates throughout the substitution process, substitution as such does not affect the educational composition of the net sample. It has been demonstrated that similarity between initial selected households and substitute households in the BHIS was achieved for the criteria used for substitution (age-group of the reference person, household size) and that consequently the predefined net sample size and -composition could be realized (18). Yet, the educational level could not be used as a criterion for substitution, since this information in not available in the Belgian NR. Clusters of households, each of them homogeneous in terms of the criteria used for substitution, showed to be heterogeneous in terms of the educational level of the households' reference person. Due to this heterogeneity, the composition of each substitute wave resembles the educational composition of the initial selected households.

What was found for the linked BHIS 2013 – Census 2011 data, was confirmed by an identical analysis of the linked BHIS 2001 – Census 2001 data, although the substitution process applied in BHIS 2013 slightly differed from the BHIS 2001.

The study shows some limitations. The highest diploma achieved by the households' reference person was used as a proxy for the socio-economic position. Of course, socio-economic position is a broader concept than just the educational level. Alternatives, such as the income level or the labor status were either lacking in the census-data (income) or shows very high level of missingness (labor status) and were therefore not used to define the socio-economic position.

Even the quality of the "highest diploma" variable is doubtful since it depends largely on the quality of the gathered information in the context of the Census 2011. The Census 2011 is an administrative update of the Census 2001 data. Citizens for which no information was available in the Census 2001 and that did not obtain a new diploma in the period 2001–2011 remained in the category "highest diploma unknown" in the Census 2011.

To deal with missing data on the "highest diploma", multiple imputation was applied, under the missingness at random assumption. Unfortunately, only few variables derived from the (administrative) census were available and could be included in the imputation model. A sensitivity analysis showed that withholding only households, for which the educational level of the households' reference person was known, did not alter the findings after having applied multiple imputation.

Conclusion

The study shows that sample substitution, as it was applied in the BHIS 2013 and the BHIS 2001, does not impact the socio-economic composition of the net sample. Regardless of the socio-economic differences in participation for every wave of substitution – with invariably lower participation rates for low educated households at each start of a substitution wave - the original socio-economic composition of the sample is restored at the start of each substitution wave.

Abbreviations

BHIS Belgian Health Interview Survey

NR National Register

SES Socio-Economic Status

TTP Trusted Third Party

MAR Missingness At Random

ISCED International Classification of Education

Declarations

Ethics approval

Ethics approval for the Health Interview Survey 2013 was obtained by the Ethical Committee of the Gent University, Belgian registration number: B670201215121.

Consent for publication

Not applicable

Availability of data and materials

Data used for this paper can be obtained at request and after having received the permission of the Belgian Data Protection Authority

Acknowledgements

The authors are grateful to Statistics Belgium for all the preparatory work needed to link BHIS data with Census data.

Funding

This paper presents independent research for which no specific grant from funding agencies in the public, commercial, or not-for-profit sectors was received.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SD was responsible for designing the objectives and approach of the study. YB prepared the study specific key that enabled do link BHIS data with census data. FB prepared the necessary statistical program to apply multiple imputation in order to deal with item missingness for the variable ‘education’. GM, SD, RC, EB, HVO and GVH critically revised the manuscript. All authors read and approved the final manuscript.

References

1. Baldissera S, Ferrante G, Quarchioni E, Minardi V, Possenti V, Carrozzi G, et al. Field substitution of nonresponders can maintain sample size and structure without altering survey estimates. The experience of the Italian behavioral risk factors surveillance system (PASSI). Ann Epidemiol. 2014;24(4):241–5.
2. Li L, Krenzke T, Mohadjer L. Considerations for Selection and Release of Reserve Samples for In-Person Surveys. JSM. 2014;40(1):105–23.
3. Nishimura R. Substitution of Nonresponding Units in Probability Sampling. University of Maryland; 2015.

4. Chapman DW. The impact of substitution on survey estimates. In: Madow WG, Olkin I, Rubin DB, editors. Incomplete data in sample surveys. New York: Academic Press; 1983. p. 45–61.
5. Chapman DW, Roman AM. An investigation of substitution for an RDD survey. In: Proceedings of the Section on Survey Research Methods. American Statistical Association; 1985. p. 269–74.
6. Chapman DW. To substitute or not to substitute that is the question. Surv Stat. 2003;48:32–4.
7. Chiu WF, Yucel RM, Zanutto E, Zaslavsky AM. Using matched substitutes to improve imputations for geographically linked databases. Surv Methodol. 2005;31:65–72.
8. Kish L, Hess I. A “Replacement” Procedure for Reducing the Bias of Nonresponse. Am Stat. 2004 Nov 1;58(4):295–7.
9. Rubin DB, Zanutto E. Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations. Surv Nonresponse. 2002;389–402.
10. Vehovar V. Field substitutions-a neglected option. In: Proceedings of the Survey Research Methods Section. American Statistical Association; 1994. p. 589–94.
11. Vehovar V. Field Substitutions in Slovene Public Opinion Survey. In: Ferligoj A, Kramberger A, editors. Contributions to Methodology and Statistics. 1995. p. 39–66.
12. Vehovar V. Field substitution and unit nonresponse. J Off Stat. 1999;15:335–50.
13. Vehovar V. Field substitutions redefined. Surv Stat. 2003;(48):35–7.
14. Lynn P. The use of substitution in surveys. Surv Stat. 2004;49:14–6.
15. Pickery J, Carton A. Oversampling in relation to differential regional response rates. Surv Res Methods. 2008;2(2):83–92.
16. Nathan G. Substitution for non-response as a means to control sample size. Sankhya C. 1980;42:50–5.
17. Demarest S, Gisle L, Van der Heyden J. Playing hard to get: field substitutions in health surveys. Int J Public Health. 2007;52(3):188–9.
18. Demarest S, Molenberghs G, Van der Heyden J, Gisle L, Van Oyen H, de Waleffe S, et al. Sample substitution can be an acceptable data-collection strategy: the case of the Belgian Health Interview Survey. Int J Public Health. 2017;
19. Demarest S, Van der Heyden J, Charafeddine R, Tafforeau J, Van Oyen H, Van Hal G. Socio-economic differences in participation of households in a Belgian national health survey. Eur J Public Health. 2013;23(6):981–5.
20. Ekholm O, Gundgaard J, Rasmussen NKR, Hansen EH. The effect of health, socio-economic position, and mode of data collection on non-response in health interview surveys. Scand J Public Health. 2010 Sep 17;38(7):699–706.
21. Helasoja V, Prättälä R, Dregval L, Pudule I, Kasmel A. Late response and item nonresponse in the Finbalt Health Monitor survey. Vol. 12, European journal of public health. 2002. 117–123 p.
22. Korkeila K, Suominen S, Ahvenainen J, Ojanlatva A, Rautava P, Helenius H, et al. Non-response and related factors in a nation-wide health survey. Eur J Epidemiol. 2001;17(11):991–9.

23. Lindén-Boström M, Persson C. A selective follow-up study on a public health survey. Eur J Public Health. 2012/01/16. 2013 Feb;23(1):152–7.
24. Mannetje A 't, Eng A, Douwes J, Ellison-Loschmann L, McLean D, Pearce N. Determinants of non-response in an occupational exposure and health survey in New Zealand. Aust N Z J Public Health. 2011 Jun;35(3):256–63.
25. Demarest S, Van Oyen H, Roskam A-J, Cox B, Regidor E, Mackenbach JP, et al. Educational inequalities in leisure-time physical activity in 15 European countries. Eur J Public Health. 2014;24(2).
26. Gray L, McCartney G, White IR, Katikireddi SV, Rutherford L, Gorman E, et al. Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol. BMJ Open. 2013 Jan 1;3(3):e002647.
27. Martikainen P, Laaksonen M, Piha K, Lallukka T. Does survey non-response bias the association between occupational social class and health? Scand J Public Health. 2007;35(2):212–5.
28. Demarest S, Gisle L, Heyden J. Playing hard to get: Field substitutions in health surveys. Int J Public Health. 2007;52(3).
29. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian Health Interview Survey 1997-2008. Arch Public Health. 2013;71(1):24.
30. Van Der Heyden J, Demarest S, Van Herck K, De Bacquer D, Tafforeau J, Van Oyen H. Association between variables used in the field substitution and post-stratification adjustment in the Belgian health interview survey and non-response. Int J Public Health. 2014;59(1).
31. Lorant V, Demarest S, Miermans PJ, Van Oyen H. Survey error in measuring socio-economic risk factors of health status: a comparison of a survey and a census. Int J Epidemiol. 2007 Dec 1;36(6):1292–9.
32. UNESCO. International standard classification of education: ISCED 2011. UNESCO Institute for Statistics Montreal; 2012.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx