sciensano

levenslang gezond

ON THE USE OF FIELD-SUBSTITUTION IN (HEALTH) SURVEYS

Stefaan Demarest

WD Epidemiologie en Volksgezondheid

Dienst levensstijl en chronische ziekten

Team 'Health Surveys'

Seminarie Statistiek Vlaanderen

21 juni 2022

.be

Context (1)

- basic interferential paradigm assumes…
  - 100 % response rate (gross-sample = net-sample)
  - absence of non response

100 % response rates have never been achieved. In global terms, response rates are dropping (depending on the survey topic, mode of data-collection, target population, burden of participation,…)

sciensano

.be

Context (2): response rates ESS 2002-2018

	2002 (round 1)	2018 (round 9)
Albania	-	55.6
Austria	60.4	50.8
Belgium	59.2	57.6
Bulgaria	-	69.4
Croatia	-	43.2
Cyprus	-	53.4
Czechia	43.3	67.4
Denmark	67.6	48.8
Estonia	-	62.7
Finland	73.2	51.8
France	43.1	48.1
Germany	55.7	27.6
Greece	80	-
Hungary	69.9	40.7
Iceland	-	40.5
Ireland	64.5	62
Israel	71	-
Italy	43.7	51.9
Latvia	-	38.9
Lithuania	-	59.2
Montenegro	-	62.3
Luxembourg	43.9	-
Netherlands	67.9	49.6
Norway	65	43.3
Poland	73.2	60.4
Portugal	68.8	34.9
Serbia	-	57.9
Slovakia	-	39.6
Slovenia	70.5	64.1
Spain	53.2	53.8
Sweden	69.5	39
Switzerland	33.5	51.8
United Kingdom	55.5	41

Data source: https://www.europeansocialsurvey.org/data/ - consulted June 15 2022

sciensano

.be

Context (3)

- (Increasing) non-response yields:
  - the reduction of the net-sample size and, consequently, drops in the precision of the estimates
  - possible bias in the estimates, when non-respondents differ from respondents in the characteristics measured

Obtaining a net-sample with a size as close a possible to the predefined sample size and a composition as close as possible to the gross-sample is worth striving for:

- Preventive strategy
- Corrective strategy

sciensano

.be

Intro: what is field substitution?

- Vehovar V. Field substitution and unit nonresponse (Journal of Official Statistics 1999;15:335-50)

  "Field substitution occurs when a nonresponding unit is replaced by a substitute (reserve) unit during the field work stage of the survey process"

- **Taxonomy** of substitution (dimensions) (Lynn P. The use of substitution in surveys. The Survey Statistician 2004;49:14-6):

  - Decision to substitute: by the interviewers/by the office
  - Selection of the substitute: by the interviewers/by the office
  - Method to select substitutes: random versus matched

- Method of substitution highly criticized in literature: introduces additional bias, impacts efforts to contact cases, suggests high participation rates, prolongs fieldwork,…

Introducing the Belgian Health Interview Survey (BHIS)

- BHIS organized in 1997, 2001, 2004, 2008, 2013, 2018
- Commissioned by all ministeries responsible for public health at federal, regional and communal level
- Sample expressed in terms of **NET** sample (3,500/4200 **individual** participants in the Flemish Region, 3,000 in the Brussels Capital Region, 3,500 in the Walloon Region) – possibility for provincial oversampling and oversampling specific age-groups (expressed in additional net-sample figures)
- By means of a household survey (max 4 members/household eligible for interview)
- Interviews to be spread throughout one calender-year (4 trimesters, ¼ of interviews per trimester)
- F2F mode + SAQ for household members 15+
- Content BHIS: health status, lifestyle, medical consumption, prevention,...

sciensano

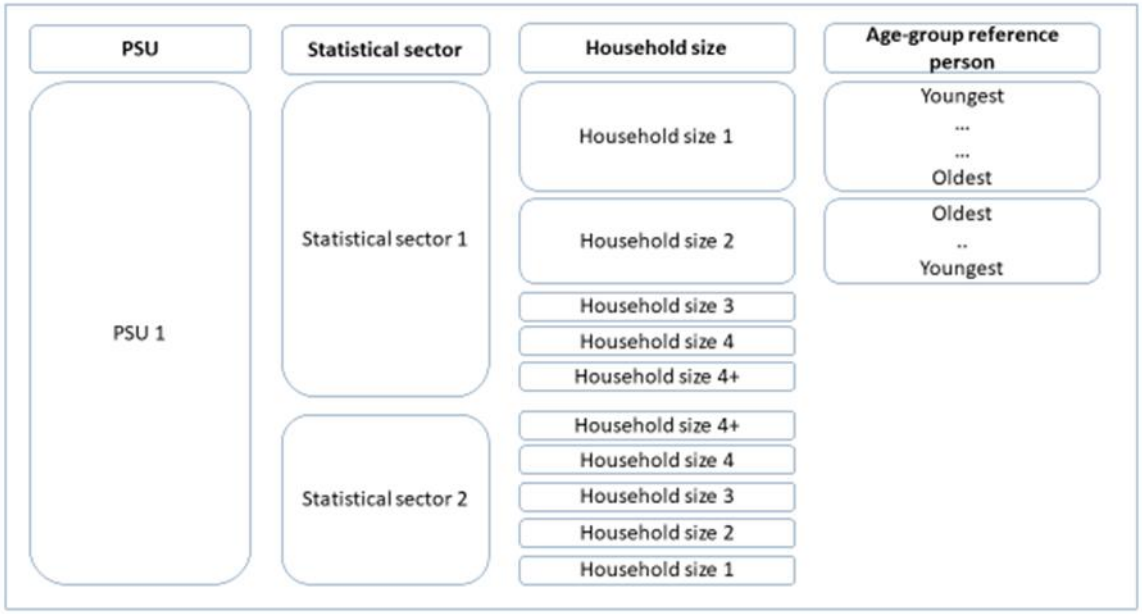
.be

Field substitution in BHIS – an overview

- **Why** field substitution?
  - Uncertainty about response-rates
  - Focused on achieving net-sample size/region, both in size as in composition
  - Interviews to be spread over one calendar-year (seasonal impact on diseases)

- **How** field substitution applied?
  - Sampling frame: National Register
  - Selection of municipalities
  - 12 independent samples (3 regions * 4 trimesters),
  - For each sample: ordering households based on statistical sector, age-group of the reference person, size of the household (number of members) in a selection of municipalities.
  - Stepwise selection of households (step-size based on mean household size in the municipality/2)
  - Each time a household is selected: selection of the three consecutive households (clusters of four households matched on selection criteria)
  - Result: twice as many cluster as strictly needed
  - Vertical (clusters) and horizontal (households within clusters) scrambling
  - Activation of half of the clusters/households at the start of a trimester
  - Interviewers ONLY paid for participating households (60% taxable)

Ordering households within municipalities selected for participation



Stepwise selection of households (example)

Num_Men	Group/Tri	SEC102001	rankstat	HHSELECT	npers	AgeCat
1191012	119	A4PA	4	660	4	50-54
1191088	119	A101	6	1784	4	45-49
1191073	119	A31-	10	2908	4	35-39
1191147	119	A01-	14	4032	1	45-49
1191051	119	A30-	15	6281	1	55-59
1191127	119	A30-	15	5157	4	30-34
1191134	119	A332	21	7405	3	60-64
1191028	119	A71-	24	8529	1	40-44
1191093	119	A00-	29	9653	1	65-69
1191067	119	A700	30	10777	4	50-54
1191047	119	A03-	32	13025	1	75+
1191111	119	A03-	32	11901	3	75+
1191031	119	A60-	36	14149	2	50-54
1191105	119	A042	38	15273	1	30-34

Creating clusters of households (example)

Num_Men	Group/Tri	SEC102001	rankstat	HHnrInGroup				npers	AgeCat
				I_POS 1	I_POS 2	I_POS 3	I_POS 4		
1191012	119	A4PA	4	660	**661**	**662**	**663**	4	50-54
1191088	119	A101	6	1784	**1785**	**1786**	**1787**	4	45-49
1191073	119	A31-	10	2908	**2909**	**2910**	**2911**	4	35-39
1191147	119	A01-	14	4032	**4033**	**4034**	**4035**	1	45-49
1191051	119	A30-	15	6281	**6282**	**6283**	**6284**	1	55-59
1191127	119	A30-	15	5157	**5158**	**5159**	**5160**	4	30-34
1191134	119	A332	21	7405	**7406**	**7407**	**7408**	3	60-64
1191028	119	A71-	24	8529	**8530**	**8531**	**8532**	1	40-44
1191093	119	A00-	29	9653	**9654**	**9655**	**9656**	1	65-69
1191067	119	A700	30	10777	**10778**	**10779**	**10780**	4	50-54
1191047	119	A03-	32	13025	**13026**	**13027**	**13028**	1	75+
1191111	119	A03-	32	11901	**11902**	**11903**	**11904**	3	75+
1191031	119	A60-	36	14149	**14150**	**14151**	**14151**	2	50-54
1191105	119	A042	38	15273	**15274**	**15275**	**15276**	1	30-34

Horizontal/vertical scrambling of households (example)

Group/Tri	SEC10200	rankstat	AS_POS1	AS_POS2	AS_POS3	AS_POS4	npers	AgeCat
119	A4PA	4	663	660	662	661	4	50-54
119	A60-	36	14149	14151	14150	14152	2	50-54
119	A30-	15	6281	6284	6282	6283	1	55-59
119	A31-	10	2910	2911	2908	2909	4	30-34
119	A00-	29	9654	9656	9653	9655	1	65-69
119	A03-	32	11901	11904	11903	11902	3	75+
119	A332	21	7408	7407	7406	7405	3	50-54
119	A71-	24	8531	8532	8530	8529	1	40-44
119	A03-	32	13028	13027	13025	13026	1	75+
119	A700	30	10778	10779	10777	10780	4	50-54
119	A101	6	1786	1785	1787	1784	4	45-49
119	A042	38	15273	15275	15274	15276	1	30-34
119	A30-	15	5160	5158	5157	5159	4	30-34
119	A01-	14	4033	4034	4032	4035	1	45-49

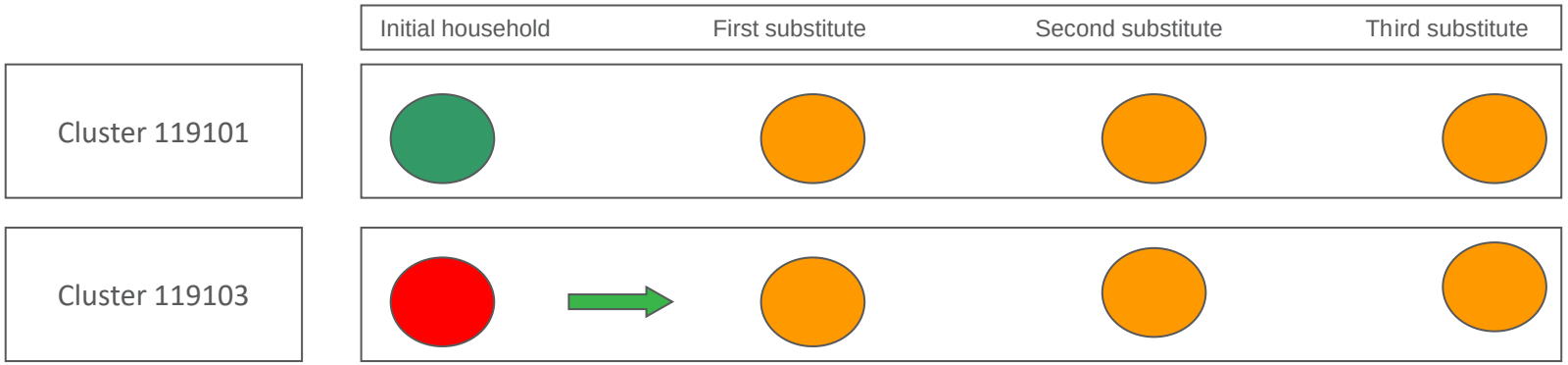
Identifying initial/substitute clusters (example)

INITIAL CLUSTERS								
Group/Tri	SEC102001	rankstat	Activation order of households				npers	AgeCat
119	A4PA	4	663	660	662	661	4	50-54
BHIS HOUSEHOLD ID			**1191011**	**1191012**	**1191013**	**1191014**		
119	A60-	36	14149	14151	14150	14152	2	50-54
BHIS HOUSEHOLD ID			**1191031**	**1191032**	**1191033**	**1191034**		
119	A30-	15	6281	6284	6282	6283	1	55-59
BHIS HOUSEHOLD ID			**1191051**	**1191052**	**1191053**	**1191054**		
119	A31-	10	2910	2911	2908	2909	4	30-34
BHIS HOUSEHOLD ID			**1191071**	**1191072**	**1191073**	**1191074**		
119	A00-	29	9654	9656	9653	9655	1	65-69
BHIS HOUSEHOLD ID			**1191091**	**1191092**	**1191093**	**1191094**		
119	A03-	32	11901	11904	11903	11902	3	75+
BHIS HOUSEHOLD ID			**1191111**	**1191112**	**1191113**	**1191114**		
119	A332	21	7408	7407	7406	7405	3	50-54
BHIS HOUSEHOLD ID			**1191131**	**1191132**	**1191133**	**1191134**		
SUBSTITUTE CLUSTERS								
Group/Tri	SEC102001	rankstat	Activation order of households				npers	AgeCat
119	A71-	24	8531	8532	8530	8529	1	40-44
BHIS HOUSEHOLD ID			**1191025**	**1191026**	**1191027**	**1191028**		
119	A03-	32	13028	13027	13025	13026	1	75+
BHIS HOUSEHOLD ID			**1191045**	**1191046**	**1191047**	**1191048**		
119	A700	30	10778	10779	10777	10780	4	50-54
BHIS HOUSEHOLD ID			**1191065**	**1191066**	**1191067**	**1191068**		
119	A101	6	1786	1785	1787	1784	4	45-49
BHIS HOUSEHOLD ID			**1191085**	**1191086**	**1191087**	**1191088**		
119	A042	38	15273	15275	15274	15276	1	30-34
BHIS HOUSEHOLD ID			**1191105**	**1191106**	**1191107**	**1191108**		
119	A30-	15	5160	5158	5157	5159	4	30-34
BHIS HOUSEHOLD ID			**1191125**	**1191126**	**1191127**	**1191128**		
119	A01-	14	4033	4034	4032	4035	1	45-49
BHIS HOUSEHOLD ID			**1191145**	**1191146**	**1191147**	**1191148**		

sciensano

.be

Data-collection BHIS (1)



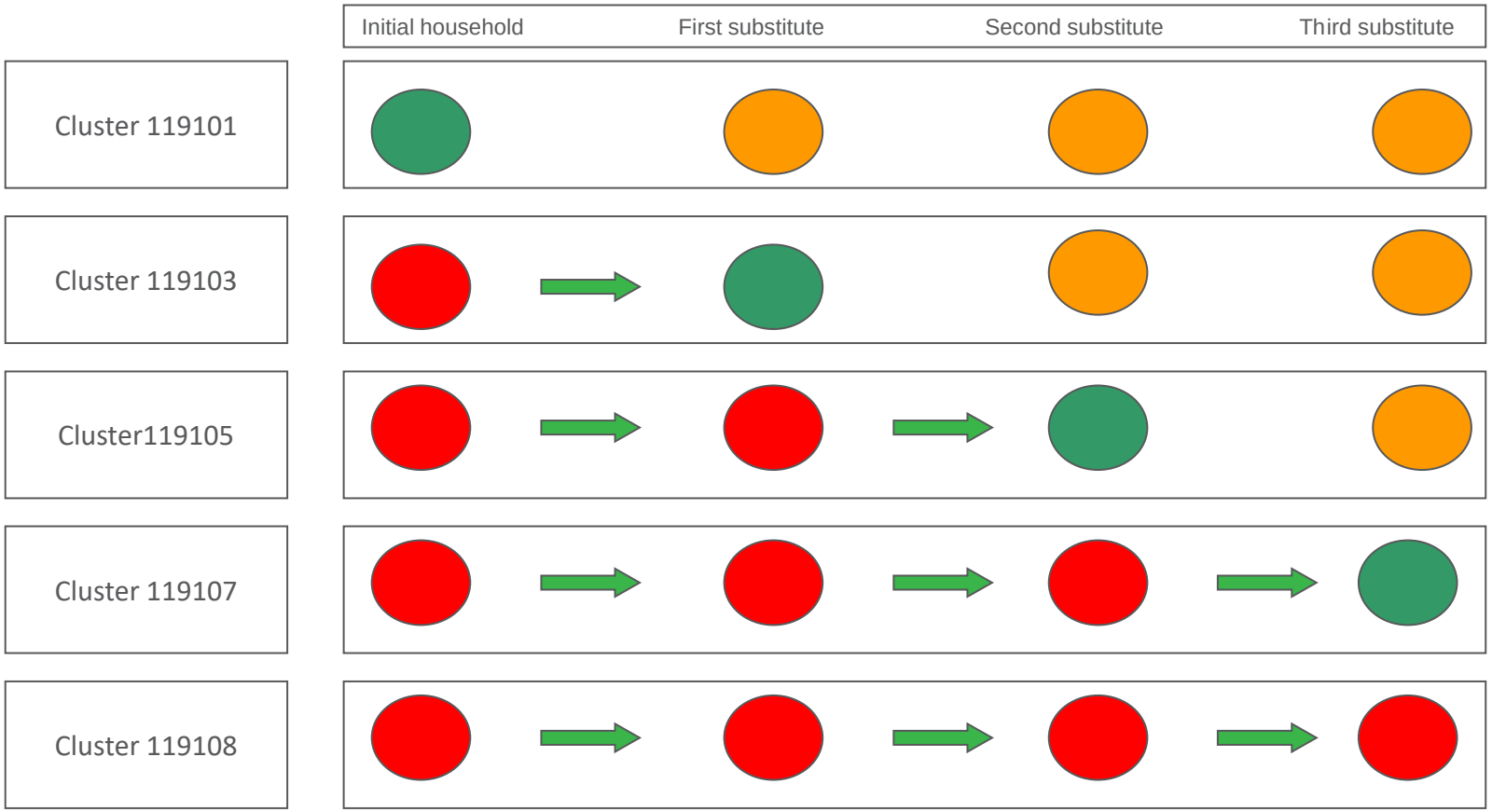
- At least 5 documented contact-attempts, of which at least one at doorstep
- Contact-attempts on different days of the week (including weekends)
- Contact-attempts on different hours of the day
- Final participation status decided upon by the interviewer when satisfying the criteria

- Activation substitute household: next in line in the cluster
- Coordinates of the (substitute) household announced to the interviewer
- Exact the same contact-procedure as with initial selected households

sciensano

.be

Data-collection BHIS (2)



sciensano

.be

Number of participants (scheduled – realized) BHIS by survey year and region

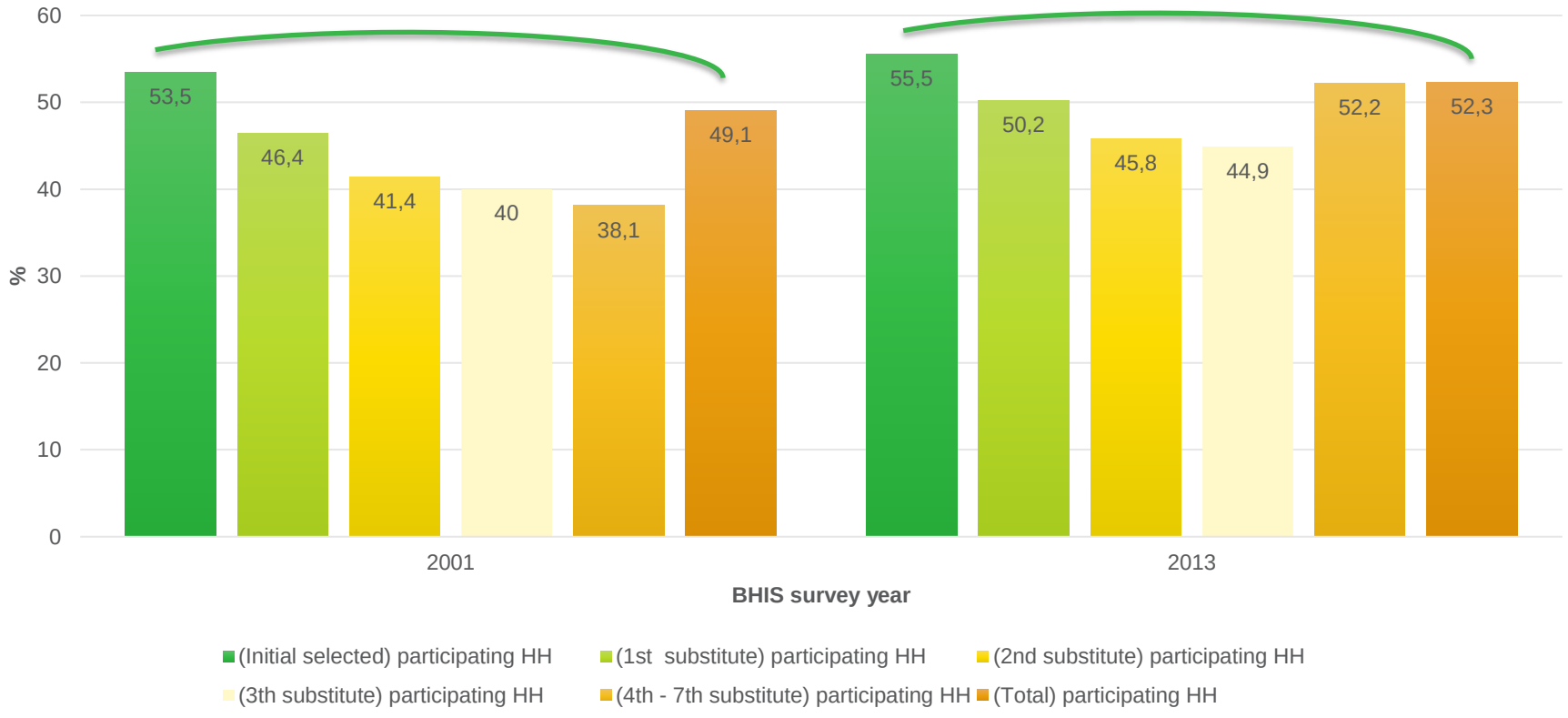
	Flemish Region		Brussels Capital Region		Walloon Region		Belgium	
	Scheduled	Realised	*Scheduled*	Realised	*Scheduled*	Realised	*Scheduled*	Realised
1997	*3500*	3536	*3000*	3051	*3500*	3634	*10000*	10221
2001	*4050*	4100	*3000*	3006	*5000*	5005	*12500*	12111
2004	*4400*	4513	*3350*	3440	*4850*	4992	*12600*	12945
2008	*3950*	3897	*3350*	3351	*3950*	4006	*11250*	11254
2013	*3500*	3512	*3000*	3103	*4100*	4214	*10600*	10829
2018	*4200*	4296	*3000*	3099	*4100*	4216	*11300*	11611

Fraud!

sciensano

.be

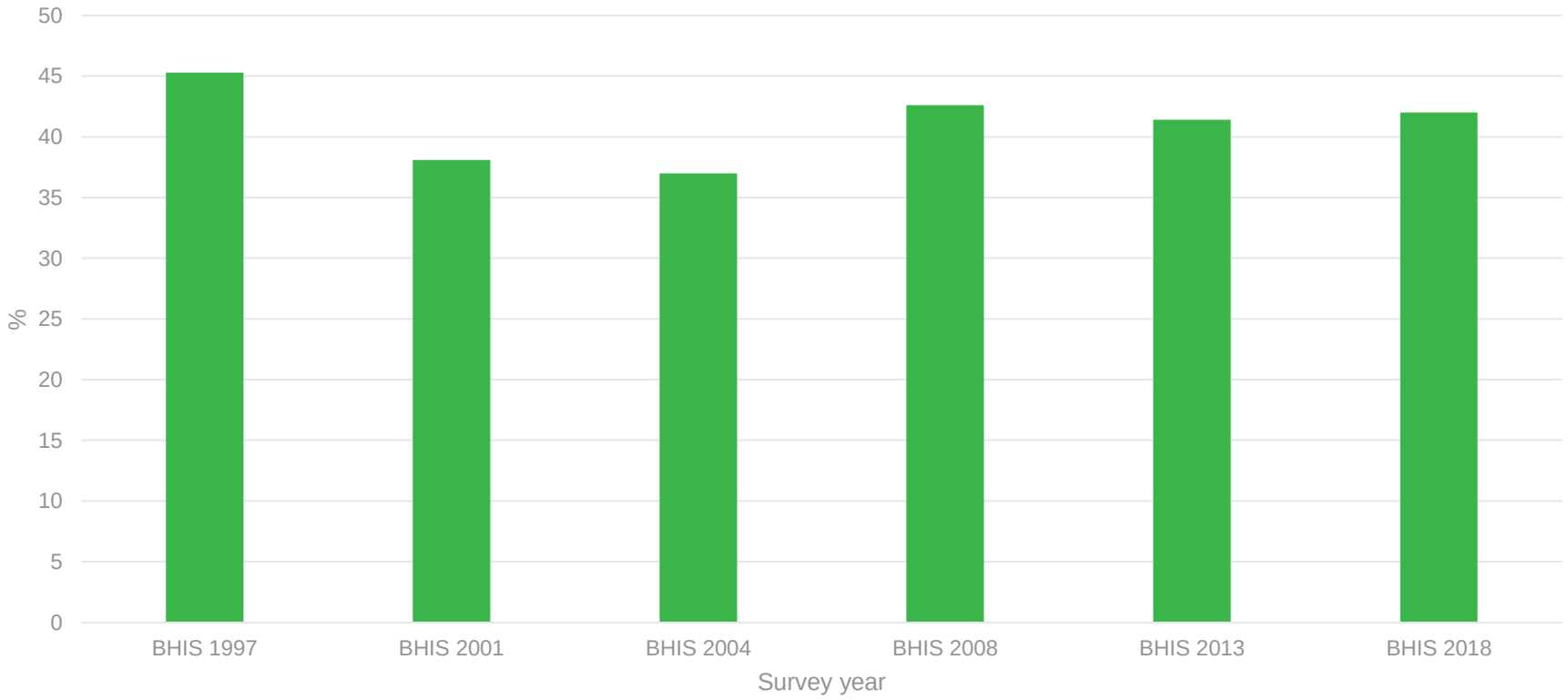
Percentage of participating households according to substitution wave BHIS 2001,2013



sciensano

.be

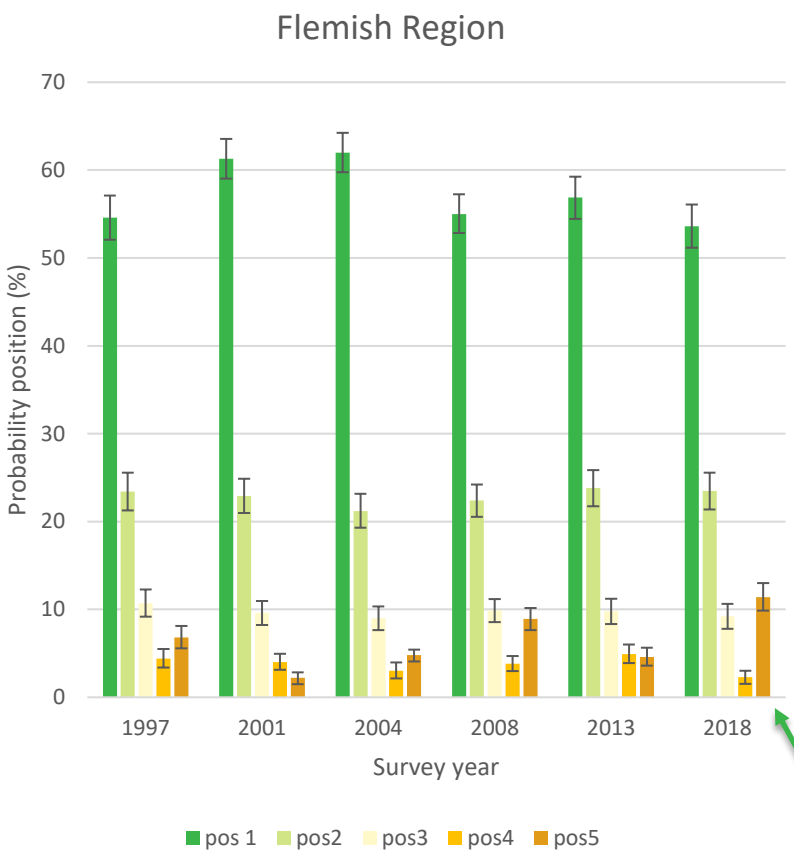
Share of participating substitute households among all participating households according to survey year, BHIS 1997 - 2018

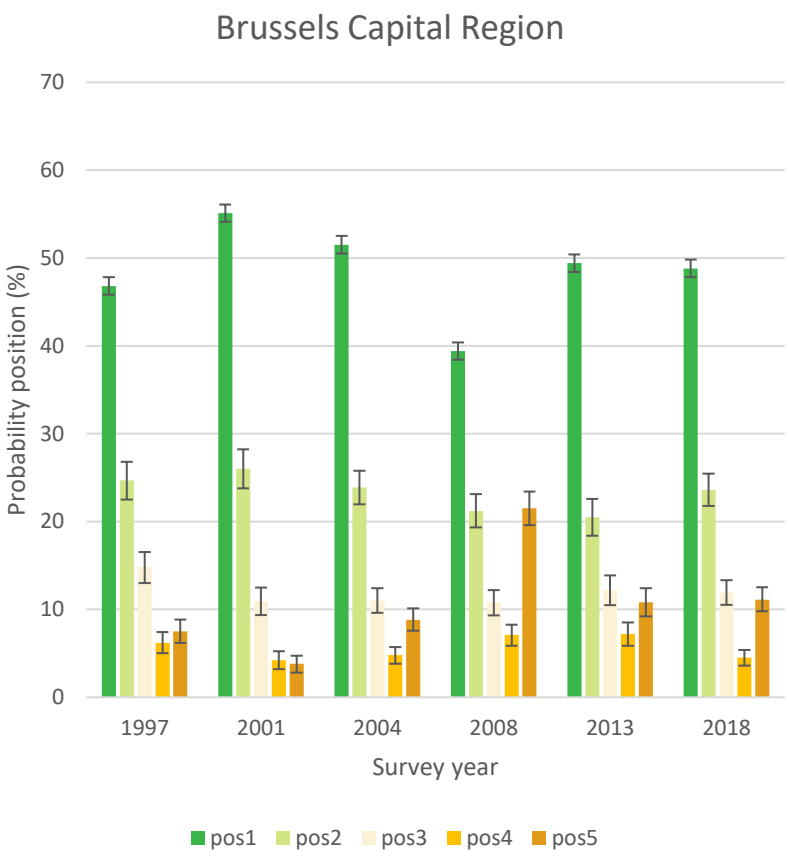


sciensano

.be

Probability of participation according to the place in the cluster, by survey year, Flemish Region versus Brussels Capital Region





Unmatched households

Does field-substitution impact the socio-economic composition of the (net) sample?

- Hypothesis: field substitution *inflates* educational differences in BHIS participation (given lower participation rates for 'low educated households' and substitution by similar households)
- Data on participation status: derived from BHIS2001 – BHIS2013 para-data
- Proxy for socio-economic position of the households: highest diploma obtained by the households' reference person
- Info on highest diploma derived from the Census 2001 and administrative Census 2011 - multiple imputation applied (highest diploma: +/- 16% missing data)
- 'one-to-one' merge BHIS para-data and census data (enabled by common National Register number) for households' reference persons

sciensano

.be

Composition of the 'net' sample (HH level) *prior* to field substitution

Info derived from linked BHIS 2001 – Census 2001 database					
Educational level	Activated HH	Participating HH	Participation rate (%)	Diff. with Low	p value for difference
Low	3,013	1,566	51.7		
Middle	1,700	897	52.8	1.1	0.5009 (*)
High	1,533	886	57.8	6.1	0.0011 (**)

Info derived from linked BHIS 2013 – Census 2011 database					
Educational level	Activated HH	Participating HH	Participation rate (%)	Diff. with Low	p value for difference
Low	2,129	1,099	51.6		
Middle	1,461	804	55.1	3.5	0.0514 (*)
High	1,446	892	61.7	10.1	<0.0001 (**)

sciensano

* 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

.be

Composition of the 'net' sample (HH level) *after* field substitution

Info derived from linked BHIS 2001 – Census 2001 database					
Educational level	Activated HH	Participating HH	Participation rate (%)	Diff. with Low	p value for difference
Low	5,389	2,534	47.0		
Middle	3,008	1.468	48.8	1.8	0.1536 (*)
High	2,834	1,518	53.5	6.5	<0.0001 (**)

Info derived from linked BHIS 2013 – Census 2011 database					
Educational level	Activated HH	Participating HH	Participation rate (%)	Diff. with Low	p value for difference
Low	4,111	1,983	48.2		
Middle	2,811	1,466	52.2	4,0	0.0152 (*)
High	2,740	1,600	58.4	10.2	<0.0001 (**)

sciensano

* 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

.be

Educational composition by substitution wave

BHIS 2001			BHIS 2013		
Initial selected households			Initial selected households		
Low	3,013	48.2	Low	2,129	42.3
Middle	1,700	27.2	Middle	1,461	29.0
High	1,533	24.5	High	1,446	28.7
1st substitute households			1st substitute households		
Low	1,313	48.0	Low	980	43.7
Middle	727	26.6	Middle	655	29.2
High	696	25.4	High	607	27.1
2nd substitute households			2nd substitute households		
Low	620	48.5	Low	477	42.6
Middle	331	25.8	Middle	327	29.2
High	328	25.6	High	317	28.3
3rd substitute households			3rd substitute households		
Low	298	48.4	Low	260	42.8
Middle	159	25.9	Middle	182	29.9
High	159	25.7	High	166	27.3
4th – 7th substitute households			4th – 7th substitute households		
Low	144	40.8	Low	264	40.4
Middle	91	25.7	Middle	187	28.5
High	119	33.5	High	204	31.1
Total activated households			Total activated households		
Low	5,389	48.0	Low	4,111	42.5
Middle	3,008	26.8	Middle	2,811	29.1
High	2,834	25.2	High	2,740	28.4

Conclusion

- Field substitution – as currently applied in BHIS - assures the predefined sample size is obtained, both in size as in composition (selection criteria)
- +/- 40% of all participating households are substitute-households (BHIS 2018: increase of participating households belonging to substitute clusters)
- Field substitution does not affect – in neither sense - the socio-economic (~ educational) composition of the net sample

sciensano

.be

Discussion

- Field-substitution…
  - is redundant if applying preventive strategies is 100% successful…
  - is an acceptable technique in data-collection: consider it when preparing a survey
  - requests a 'rich' sampling frame (matching criteria) and a relative long data-collection phase
  - can also be applied in not-in-person surveys (e.g. online surveys)
  - enables a very tight follow-up of data-collection and enables to adapt data-collection

sciensano

.be

Contact

stefaan.demarest@sciensano.be

Thanks

Sciensano • Juliette Wytsmanstraat 14 • 1050 Brussel • België
T + 32 2 642 51 11 • T presse + 32 2 642 50 01 • info@sciensano.be • **www.sciensano.be**