Methodological note for AES 2022

Thomas Delclite, Statbel, 23/11/2023

1. Sampling procedure

The targeted net sample size for this survey was 9,130 individuals: 3,310 young adults (18-24 years old) and 5,820 adults (25-69 years old). To balance cost considerations and profile diversity, we opted for a survey employing two data collection methods. Therefore, 75% of the net sample was planned to be obtained through web surveys (CAWI), while the remaining 25% was planned to by collected through face-to-face surveys (CAPI).

Due to field constraints, the initial step involved selecting primary sampling units (PSUs), which are segments of Belgian municipalities containing more than 300 individuals each. Out of a total of 2,237 available PSUs, we randomly chose 319 PSUs.

To enhance regional precision at NUTS1 level, we opted to increase the number of PSUs drawn in smaller regions. To determine the allocation of PSUs to each region, we used the square root of the population size (number of individuals) in each region as a basis. The PSUs were then ordered based on their median fiscal income, and a systematic draw method was employed to make the final PSU selections.

Table 1. Drawn sample size by age and region at CAWI stage

	Number of PSU according to size of region	Number of PSU according to $\sqrt{size}$ of region
Brussels	36	65
Flanders	185	147
Wallonia	98	107
Total	319	319

The sampling approach for the AES survey involved two stages: an initial CAWI stage based on the primary sample, followed by a CAPI stage targeting non-respondents from the CAWI stage. Further details regarding these two stages are provided below.

a. CAWI stage

We began with a well-defined population frame as of May 28, 2022, consisting of individuals aged 18 to 69, residing in private households. This frame encompassed a total of 7,304,511 individuals.

For our CAWI survey, we assumed a response rate of 25% for the young adults and 30% for the adults, based on estimates derived from previous CAWI surveys conducted by Statbel. We foresaw the possibility of individuals residing in the same household within our sample by increasing the sample size slightly and, in the event of this scenario occurring, we randomly selected one individual from each household for the final sample. In such cases, only one individual per household, chosen randomly, was kept in the final sample.

For each PSU, we randomly selected 79 individuals (32 young adults and 47 adults), for a total of 23,357 individuals. The response rate turned out to be significantly lower than anticipated, standing at around 15%

instead of the expected 30%. Consequently, we found it necessary to draw a second sample of individuals while retaining the same primary sampling units (PSUs). In this second sample, we selected 65 individuals per PSU, comprising 26 young adults and 39 adults. As a result, our overall gross sample expanded to encompass 44,851 individuals, and this sample was distributed as follows (table 2). During the CAWI stage, we successfully collected a total of 6,257 responses.

Table 2. Drawn sample size by age and region at CAWI stage

	Youngs Adults	Adults	Total
Brussels	3,746	5,575	9,321
Flanders	8,270	12,480	20,750
Wallonia	5,814	8,966	14,780
Total	17,830	27,021	44,851

b. CAPI stage

Following the CAWI stage, we proceeded to the CAPI step by selecting a sub-sample from the 38,594 individuals who did not respond during the initial CAWI survey. Similar to the CAWI stage, we determined the expected number of net CAPI interviews based on region and age. The CAPI sample for each category of age was randomly drawn.

We established a gross sample of 9,343 individuals based on the expected response rates. These individuals were distributed as follows (table 3). During the CAPI stage, we successfully collected a total of 2,017 responses.

Table 3. Drawn sample size by age and region at CAPI stage

	Youngs Adults	Adults	Total
Brussels	3,201	640	3,841
Flanders	2,290	461	2,751
Wallonia	2,291	460	2,751
	7,782	1,561	9,343

2. Extrapolation

Responses from both the CAWI and CAPI methods are subject to extrapolation. This involves adjusting survey weights to take into account the non-response, followed by a calibration process aimed at enhancing result accuracy. It's important to note that specific calculation steps are required for extrapolation, particularly because the CAPI sample is a subset of the CAWI sample.

When the sample s is drawn, the drawing probabilities π_k and associated sample weight $1/\pi_k$ remain unchanged. An unbiased estimator of total of Y is given by the Horvitz-Thomson estimator :

$$\hat{Y} = \sum_{k \epsilon s} \frac{1}{\pi_k} y_k$$

The sample can be conceptually divided into two distinct segments. The first segment consists of CAWI respondents, who are retained with their original survey weights. The second segment comprises CAWI non-respondents (from which a sub-sample is drawn).

To conceptualize this, we can frame it as a two-phase survey approach. In the initial phase, the standard sample is drawn, and it is assumed that all individuals selected are presented with the question: "Are you willing to participate via CAWI?" Subsequently, the responses are used to create two strata based on individuals' willingness to participate in CAWI.

In the first stratum, we include individuals who respond affirmatively with a "YES." These individuals are chosen to participate in a second phase with a certainty of 1 (meaning they are guaranteed to be interviewed via CAWI), and no non-response correction is applied. Within this stratum, the survey weights used before calibration remain unchanged, with a multiplication factor of 1 introduced for the second phase.

In the second stratum, we consider individuals who respond negatively with a "NO." From this stratum, a sub-sample is drawn with a known probability π_{2k} and the chosen individuals are subsequently interviewed via CAPI. Within this second stratum, we have a third phase of non-response. This phase is treated similarly to previous non-response scenarios but is specifically limited to this second stratum. The non-response correction is exclusively applied to this portion of the sample, resulting in an estimator that takes the following form:

$$\hat{Y} = \sum_{k \epsilon s1} \frac{1}{\pi_k} y_k + \sum_{k \epsilon r2} \frac{1}{\pi_k \pi_{2k} \pi_{rk}} y_k$$

Using this formula, it becomes apparent that the weights associated with the second group of respondents (CAPI) are considerably higher than those associated with the first group (CAWI). This is primarily due to the second group having to support both the additional sampling fraction and account for non-response. We address this by sequentially adjusting their weights: first for the second draw and then for the non-response.

The non-response correction, specifically for the CAPI portion, is executed while considering the following variables:

- REGION
- FL_BELGE (a binary variable indicating Belgian or non-Belgian individuals)
- CD_INC (quantiles of tax income categorized into 5 modalities)
- CD_AGE_DET (age of individuals categorized into 5 modalities)

As an illustration, we had initially projected a weight ratio of 7.1. These weight ratios were established based on assumptions regarding response rates and the potential for under-sampling within the CAPI portion of the survey.

Table 4. Assumptions about response rate and weight ratio

CAWI response rate		CAPI response rate			
30%		40%			
π_k	π_{2k}	π_{rk}	Weight CAWI	Weight CAPI	Weight ratio
0.0025	0.350	0.400	402	2,869	7.1

Ultimately, the CAWI response rates turned out to be significantly lower than anticipated, with rates of only 14%. This disparity had a pronounced impact on the weight ratios assigned to individuals.

Specifically, the weight ratio increases to approximately 15 in Flanders, and dramatically increases to more than 18 in Wallonia and Brussels. These weight ratios reflect the adjustment made to account for the lower-than-expected response rates at CAWI level.

Table 5. Actual responses rate and weight ratio for adults by region

	π_k	π_{2k}	π_{rk}	Weight CAWI	Weight CAPI	Weight ratio
Brussels	0.0075	0.176	0.317	133	2,429	18.26
Flanders	0.0032	0.188	0,365	310	4,538	14.64
Wallonia	0.0043	0.182	0.285	232	4,410	19.01

As demonstrated, the methodology employed results in substantial variations in weights based on the data collection method. A considerable portion of the sample (CAWI respondents) is used to extrapolate a relatively small segment of the target population, while CAPI respondents are extrapolated across a much larger portion of the population. This situation contributes to a high variance for each indicator, mainly because of the expected high variance associated with the CAPI stage. To address this issue, we put forth a method aimed at mitigating this variance.

The total $\hat{Y}$ can therefore be estimated as the weighted sum of two totals $\hat{Y}_1$ and $\hat{Y}_2$ according to the two collection methods:

$$\hat{Y} = \alpha_1 \hat{Y}_1 + \alpha_2 \hat{Y}_2$$

With :

$$\alpha_1 \hat{Y}_1 = \sum_{k \epsilon s1} \frac{1}{\pi_k} y_k$$

$$\alpha_2 \hat{Y}_2 = \sum_{k \epsilon r2} \frac{1}{\pi_k \pi_{2k} \pi_{rk}} y_k$$

$$\alpha_1 + \alpha_2 = 1$$

The total $\hat{Y}$ is unbiased but the variance is high because of the variance of $\hat{Y}_2$. However, by changing the α and assuming independence between $\hat{Y}_1$ et $\hat{Y}_2$, it is possible to decrease the total variance $V(\hat{Y})$ at the cost of an increase in bias.

$$\hat{Y}' = \alpha_1' \hat{Y}_1 + \alpha_2' \hat{Y}_2$$

It is possible to measure the quadratic difference between the initial estimate and an estimate for another pair $(\alpha_1' ; \alpha_2')$ using the following formula:

$$\text{Me} = B^2 + V(\hat{Y}')$$

$$\text{Me} = (\hat{Y}' - \hat{Y})^2 + V(\hat{Y}')$$

Thus, if the decrease in variance is large enough to compensate for the bias generated by the change in α, a new, more accurate estimate can be proposed.

For each stratum in the survey, which in this case corresponds to the regions crossed by sex, a pair of values $(\alpha_1' ; \alpha_2')$ can be calculated. This pair $(\alpha_1' ; \alpha_2')$ must be applied to all the survey indicators, so a small number of indicators must be chosen to determine the optimal pair $(\alpha_1^o ; \alpha_2^o)$ and observe the impact on variance and bias. We applied the optimisation method to 7 main indicators of the GBV survey for this optimisation.

To find the optimal α_1^o value that minimizes MSE for a given indicator, numerical simulations were conducted by testing a range of α_1' values between 0 and 1. In particular, it's worth noting that the α_1^o value was significantly increased for the Youngs Adults in Brussels compared to α_1. This adjustment is driven by two primary reasons:

1. The Brussels region had a very low number of CAPI responses, which naturally resulted in a high variance for the CAPI component of the survey.
2. Responses to the AES indicators between CAWI and CAPI in this region exhibited a relatively small degree of dissimilarity. Therefore, allocating a larger proportion to the CAWI portion of the survey did not introduce significant bias but did lead to a substantial reduction in variance.

Table 6. Actual and optimized α by age and region

	$\boldsymbol{\alpha_1}$		$\boldsymbol{\alpha_1^o}$	
	Youngs Adults	Adults	Youngs Adults	Adults
Brussels	9%	13%	25%	15%
Flanders	15%	17%	20%	20%
Wallonia	11%	12%	15%	15%

After optimizing the parameter α, the next step involved calibrating the margins from a known population frame as of January 1, 2023. This calibration process employed the same set of variables as those used in the non-response correction crossed by region, with this following variables :

This calibration process employed the following variables crossed by region :

- CD_AGE_DET (18-24, 25-34, 35-54,55-69)
- CD_INC * CD_AGE (18-24, 25-69)
  CD_SEX
  EDU_2017_CL (best level of education achieved in 2017, categorized into 3 modalities)

To perform this calibration, we utilized the SAS Calmar2 macro, which employs linear calibration constrained within the bounds of 0.5 and 1.5. To calculate the variance, the entire sample design was integrated, and this was accomplished using the SAS Poulpe macro.