STATEC

Économie & Statistiques
Working papers

N°146 07/2026

From an Official Face-to-Face Survey to an Online Probability-based Panel: The Drivers to Opt-in*

Auteurs :
Cesare A.F. Riillo, Chiara Peroni.[1]

Résumé / Abstract

FR
Les panels en ligne sont de plus en plus courants, mais le recrutement des participants repose souvent sur des mécanismes d'adhésion volontaire (« opt-in ») susceptibles d'entraîner un biais de sélection. Cet article propose une procédure permettant de constituer un panel en ligne probabiliste en utilisant une enquête officielle en face à face comme base de recrutement, en prenant comme étude de cas l'enquête EU-SILC au Luxembourg. Le comportement d'adhésion volontaire est analysé à la lumière de la théorie de l'échange social (« social exchange theory ») et de la théorie de l'influence et de la saillance (« leverage-saliency theory »), en considérant la participation au panel comme une décision de type « coût-bénéfice » motivée par l'engagement des répondants, leur accès au numérique et d'autres caractéristiques individuelles. Un ensemble riche de variables d'enquête et de paradonnées est utilisé pour modéliser ce comportement. Le modèle logit estimé présente des performances prédictives satisfaisantes et met en évidence le rôle des facteurs liés à l'engagement en tant que prédicteurs de l'adhésion. Les pondérations du panel sont calculées en combinant l'inverse des probabilités d'adhésion estimées avec les pondérations de EU-SILC ; il est démontré qu'elles améliorent la représentativité du panel obtenu. Les enquêtes officielles peuvent soutenir le développement de panels en ligne fondés sur la probabilité, offrant ainsi aux instituts de statistique une procédure de collecte de données numériques sans compromettre la qualité des données.

EN
Online panels are becoming increasingly common, yet recruitment of panellists frequently relies on voluntary opt-in mechanisms that can lead to self-selection bias. This paper proposes a procedure for constructing a probability-based online panel using an official face-to-face survey as the recruitment frame, taking EU-SILC in Luxembourg as a case study. Opt-in behaviour is analysed using insights from Social Exchange Theory and Leverage–Saliency Theory, treating panel participation as a cost–benefit decision driven by respondent engagement, digital access, and other individual characteristics. A rich set of survey variables and paradata is used to model opt-in behaviour. The estimated logit model displays satisfactory predictive performance and highlights the role of engagement-related factors as predictors of opt-in. Panel weights are constructed by combining the inverse of estimated opt-in probabilities with the EU-SILC weights and are shown to improve the representativeness of the resulting panel. Official surveys can support the development of probability-based online panels, offering statistical offices a procedure for digital data collection without compromising data quality.

1 STATEC Research, Institut national de la statistique et des études économiques du Grand-Duché de Luxembourg.

* The author(s) gratefully acknowledge the support of STATEC, the National Statistical Office of Luxembourg. This work is conducted in the framework of the APPRECIATE project, conducted with the financial support from the Luxembourg's National Research Fund (grant number FNR-14844092). The opinions and views expressed in this paper are those of the authors and do not reflect in any way those of the institutions they are affiliated with or funding institutions.

From an Official Face-to-Face Survey to an Online Probability-based Panel: The Drivers to Opt-in *

Cesare A.F. Riillo, Chiara Peroni†

July 30, 2026

Abstract

Online panels are becoming increasingly common, yet recruitment of panellists frequently relies on voluntary opt-in mechanisms that can lead to selection bias. This paper proposes a procedure for constructing a probability-based online panel using an official face-to-face survey as the recruitment frame, taking EU-SILC in Luxembourg as a case study. Opt-in behaviour is analysed using insights from Social Exchange Theory and Leverage–Saliency Theory, treating panel participation as a cost–benefit decision driven by respondent engagement, digital access, and other individual characteristics. A rich set of survey variables and paradata is used to model opt-in behaviour. The estimated logit model displays satisfactory predictive performance and highlights the role of engagement-related factors as predictors of opt-in. Panel weights are constructed by combining the inverse of estimated opt-in probabilities with the EU-SILC weights and are shown to improve the representativeness of the resulting panel. Official surveys can support the development of probability-based online panels, offering statistical offices a procedure for digital data collection without compromising data quality.

*The author(s) gratefully acknowledge the support of STATEC, the National Statistical Office of Luxembourg. This work is conducted in the framework of the APPRECIATE project, conducted with the financial support from the Luxembourg's National Research Fund (grant number FNR-14844092). The opinions and views expressed in this paper are those of the authors and do not reflect in any way those of the institutions they are affiliated with or funding institutions.

†STATEC Research, Institut national de la statistique et des études économiques du Grand-Duché du Luxembourg.

1 Introduction

Online surveys are a popular method of data collection due to their convenience, cost-effectiveness, and ability to reach a large and diverse population. However, most online panels rely on voluntary opt-in recruitment schemes, such as advertisements on internet portals or newsletters, which makes the probability of recruiting specific respondents unknown. (Indeed, these panels are referred to as non probability-based panels.) As a consequence, estimates of quantities of interest derived from online panels are prone to coverage and selection bias, limiting the extent to which survey results can be generalized to the target population (e.g., Bethlehem, 2009; Cornesse et al., 2020).[1]

This study proposes a procedure for building a probability-based online panel from an official general population survey that mitigates selection bias.

Specifically, we show that the high-quality sampling design and representativeness of EU-SILC provide a solid foundation for constructing an online panel that retains a probability-based interpretation. The panel object of this study is constructed by recruiting respondents from the Luxembourg Survey on Income and Living Conditions (EU-SILC), an official household survey conducted face-to-face by the National Institute of Statistics and Economic Studies of Luxembourg (STATEC). Each panel member is assigned a non-zero probability to enroll in the online panel (i.e. panel inclusion probability) by combining the known probability to participate in the EU-SILC survey with an estimated 'opt-in' probability based on the respondent's characteristics observed in EU-SILC.

Conventional approaches to reducing selection bias in non-probability online panels rely on calibrating the sample to known population margins, typically using demographic characteristics such as age, sex, education, and region. These variables, however, do not directly capture respondents' propensity to participate in an online survey. By recruiting panel members from an official probability survey, we are able to exploit additional auxiliary information unavailable in conven-

[1] Coverage bias arises when some members of the target population have no chance of being included in the sampling frame (e.g., individuals without internet access in an online survey). Selection bias occurs when the mechanism determining participation leads to systematic differences between respondents and the target population, even among those covered by the sampling frame.

tional online panels, including paradata describing respondents' engagement during the face-to-face interview and information on their access to digital resources. We show that these variables are strong predictors of opt-in behaviour and can substantially improve the estimation of panel inclusion probabilities.

We proceed as follows. First, we analyse the determinants of enrolment into the online panel. Using data from EU-SILC respondents, we estimate the probability of opting-in using a logistic regression model. Candidate predictors are initially selected through a linear LASSO procedure and then refined in light of theoretical considerations and prior evidence from the survey methodology literature. We assume that the opt-in decision reflects a trade-off between perceived participation costs and benefits, as well as respondents' engagement with the surveying institution. This interpretation is based on *Social Exchange Theory* (Dillman, 1978), which views survey participation as a reciprocal exchange, and *Leverage–Saliency Theory* (Groves et al., 2000), which emphasizes the relative importance respondents attach to different survey attributes. We exploit auxiliary information collected during the survey (paradata), such as respondents' engagement during the face-to-face interview, and their access to digital resources. We find that these factors are strong predictors of opt-in decisions. Second, we design a weighting scheme. We estimate 'ex-post' *panel inclusion* probabilities as the product of respondents' known probability of inclusion in the EU-SILC sample and their predicted probability of opting-in from the logit model estimation. Each recruited panellist is weighted by the inverse of the ex-post panel inclusion probability. Finally, we assess the effectiveness of the proposed weighting scheme by comparing panel distributions to known official population statistics. This allows us to evaluate whether the information content of opt-in probabilities can be used to improve the representativeness of the panel.

This paper contributes to the literature on probability-based online panels first and foremost by proposing a procedure to build a panel that retains a probabilistic interpretation from an official representative survey. It also provides additional contributions. Second, it embeds opt-in modeling within a clear theoretical framework of survey participation. Thirdly, it documents how rich paradata from an official survey can be leveraged to estimate opt-in probabilities. Finally, it shows that inverse-probability weighting based on these estimates improves panel representativeness.

The paper is structured as follows. Section 2 provides a review of the literature on online panels and the drivers of opt-in rates. Section 3 describes the methodology used to recruit the probability-based online panel and the pre-selection of the drivers to opt-in to the online panel. Section 4 presents the results of the study, on the drivers to opt-in decisions and on the effectiveness of panel weights based on panel inclusion probability. Section 5 briefly reports on two surveys that were conducted using the online panel as a sampling frame. Finally, Section 6 discusses the implications of the study and provides recommendations for future research.

2 Related studies

The use of online panels for survey research has grown considerably in recent years, as they offer advantages such as cost-effectiveness, speed, and easy access to a large pool of potential respondents (Callegaro et al., 2015). However, online panels usually rely on respondents who are recruited using voluntary opt-in schemes, such as advertisements on web portals or newsletters (Bethlehem and Biffignandi, 2010). In such cases, the probability of recruiting particular respondents is unknown, and, consequently, the results of the survey are difficult to be generalised to the population (Cornesse et al., 2020; Yeager et al., 2011).

One approach is to construct a probability-based online panel by directly recruiting from the population registry. This approach has proved to be useful to enhance the representativeness of their online panels in various countries, such as the Longitudinal Internet Studies for the Social Sciences in the Netherlands (Scherpenzeel, 2011), and the Understanding Society panel in the United Kingdom (Wood et al., 2014). One important aspect of probability-based online panels is that the sampling frame includes both the online and the offline population (e.g. offering internet access to offline households), but potential biases may remain from the non-random opt-in behaviour of participants (Blom et al., 2015). Some researchers have used targeted sampling and incentives to increase participation rates and improve the representativeness of online panels (Bethlehem and

Biffignandi, 2010).

Another approach to improve the representativeness of online non-probability survey samples when relevant auxiliary information is available from external sources, such as a probability survey sample (Chen et al., 2020)[2]. A range of weighting approaches has been proposed that adjusts survey responses according to the estimated probability of panel participation (i.e., opt-in) (Rivers, 2007; Lee, 2006; MacInnis et al., 2018; Biemer et al., 2017; Dutwin and Buskirk, 2017). From a survey sampling perspective, Beaumont (2020) interprets these as participation probabilities that, when combined with the known sampling inclusion probabilities from a probability survey, define the overall probability of inclusion in the responding sample. [3] This provides a theoretical basis for constructing inverse probability weights in two-phase survey designs. Among these, inverse probability weighting (IPW) is widely used, with respondent weights defined as the inverse of the predicted opt-in probabilities (e.g., Horvitz and Thompson, 1952; Rao, 2021). This approach has been shown to improve the representativeness of online panels by reducing biases in key socio-demographic and attitudinal variables (e.g. Schonlau et al., 2012; Chen et al., 2020). However, even if a recent review of the literature shows that probability-based samples are generally better (Cornesse et al., 2020), the effectiveness of these methods varies and yields mixed results. Weighting can introduce additional biases and uncertainties, particularly when opt-in probabilities are poorly estimated (Cornesse et al., 2020). From a survey sampling perspective, Beaumont (2020) emphasizes that the effectiveness of such weighting procedures depends critically on the quality of the estimated participation probabilities. Furthermore, the representativeness of online panels may be influenced by factors beyond those accounted for in the opt-in probability models, such as response burden (Jin and Kapteyn, 2022; Roberts et al., 2022), attrition and non-response over time (Lugtig, 2014).

A critical aspect of this approach is the accurate prediction of opt-in probabilities. Researchers have employed various statistical techniques, such as logistic regression, to model opt-in probabil-

[2] See Moreno-Betancur et al. (2025) for an overview of non-probability survey adjustment methods and their implementation in R.

[3] Related ideas have also been developed for non-probability samples through the concept of *pseudo-inclusion probabilities* (Valliant, 2020).

ities using available socio-demographic and behavioural information (Vavreck and Rivers, 2008; Chen et al., 2020). Some of the key predictors of opt-in probabilities identified in the literature include age, gender, education, income, political interest, and internet usage (Berrens et al., 2003; Malhotra and Krosnick, 2007). Bosnjak et al. (2013) investigate sample composition biases in the development of a probability-based online access panel in Germany. The results showed substantial differences in age and education and some differences in personality traits. Additionally, attitudinal factors, such as respondents' trust in surveys and willingness to provide personal information, have also been found to be significant predictors of opt-in behaviour (Couper, 2007). Several studies have examined the validity of self-reported data obtained through online surveys(e.g. Goel et al., 2015; Wang et al., 2015). However, the validity of self-reported data may vary depending on the topic and the characteristics of the respondents, and researchers should be cautious in interpreting self-reported data (Cornesse et al., 2020).

In conclusion, there is still ongoing debate about the effectiveness of using opt-in probabilities in weighting procedures to achieve representativeness. However, the literature on probability-based online panels has documented the potential of using existing probability-based surveys as a starting point for recruiting online panel participants. By modelling opt-in probabilities and weighting responses accordingly, researchers can improve the representativeness of online panels, although the effectiveness of this approach may depend on the quality of the opt-in probability estimates and the presence of unobserved factors that influence panel representativeness.

3 Method

This section presents the proposed procedure for constructing a probability-based online panel from an official survey conducted on the general population. The procedure consists of recruiting panel members from EU-SILC, modelling their probability of enrolment using auxiliary survey information, constructing panel inclusion probabilities, and evaluating the resulting inverse-probability weights. The procedure consists of the following five steps:

1. **Construction of the recruitment frame.** Respondents to the Luxembourg Survey on Income and Living Conditions (EU-SILC) are invited to join the online panel, providing a probability-based recruitment frame.

2. **Pre-selection of drivers to opt-in.** Candidate predictors are pre-selected from the rich set of EU-SILC variables using a LASSO procedure and subsequently refined according to survey participation theory, distinguishing engagement, digital access, socio-demographic, and individual characteristics.

3. **Estimation of opt-in probabilities.** A logistic regression model is estimated to predict each respondent's probability of enrolling in the online panel.

4. **Construction of panel weights.** Individual probabilities of inclusion in the online panel are computed by combining the known EU-SILC participation probabilities with the estimated opt-in probabilities. These probabilities are then used to construct inverse-probability weights.

5. **Evaluation of the weighting procedure.** The effectiveness of the proposed weighting scheme is assessed by comparing weighted panel estimates with official population statistics and with the corresponding unweighted estimates.

The remainder of this section describes each step in greater detail.

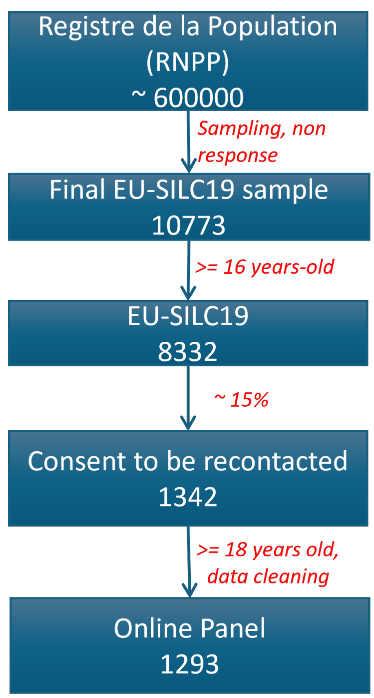
3.1 Construction of the recruitment frame

EU-SILC 2019 is a large official survey which aims to investigate Income and Living Conditions in European countries. The primary objective of EU-SILC is to gather consistent and up-to-date cross-sectional and longitudinal information on living conditions, poverty, income, and social exclusion in European countries.

In Luxembourg, EU-SILC is mandated by law and participation is mandatory. A stratified random sample of resident households is drawn from the population register and invited to participate in EU-SILC for Luxembourg. The interview is conducted face-to-face, and the response rate is quite high for a general population survey (the total response rate is 47% , while the response rate at the first contact is around 30%.). Sampling weights are computed on the basis of a rich set of variables (gender, age, canton of residence, NACE (for who is employed), citizenship and size of the household (extrapolation from the census). One can see (Eurostat., 2019) for more details on the survey.

The pool of respondents to EU-SILC to join the online panel provides a probability-based recruitment frame for the online panel. Figure 1 shows the main steps for constructing the online probability-based panel, starting from the sampling from the target population to the final online panel.

Figure 1: Steps to recruit the online panel



The EU-SILC sample is made of 10773 valid responses and the following question is administered to the 8332 respondents of the EU-SILC 2019 for Luxembourg, who were at least 16 years old in 2018: *STATEC regularly carries out qualitative studies (discussion groups, cognitive interviews, etc.) on various subjects. These studies allow STATEC to understand certain topics better* *and also to test these questionnaires in order to improve their quality. Would you agree to participate in this type of study?*. [4] 1,342 respondents agreed to be contacted for future surveys and voluntarily provided an email address. We exclude 42 respondents who are younger than 18 years of age or whose contact information is missing, duplicated, or otherwise incomplete. The final probability-based online panel comprised 1293 eligible panellists.

A potential concern is that requiring respondents to provide an email address may selectively recruit individuals with better access to, or greater familiarity with, digital technologies. However, this concern is likely to be limited for two reasons. First, 95% of EU-SILC respondents reported having access to the internet at home (which drops to 70% for those aged 70+, see also Figure 3). Second, among respondents who consented to be contacted for future surveys, only 13 did not provide an email address. These findings suggest that requiring an email address is unlikely to introduce substantial additional selection bias, particularly among the working-age population.

3.2 Pre-selection of drivers to opt-in

EU-SILC data consists of several hundred survey variables and paradata that describe the interview process and provide information on the implementation of the survey. Given this high dimensionality, we pre-selected potential predictors of opting-in using a linear LASSO model. LASSO (Least Absolute Shrinkage and Selection Operator) is a penalised regression method for variable selection and regularisation, originally introduced by (Tibshirani, 1996). The method augments the sum of squared errors with a penalty term proportional to the absolute value of the regression coefficients. This penalty induces shrinkage and sets some coefficients exactly to zero, thereby selecting a subset of relevant predictors and reducing the risk of overfitting. We excluded variables with more than 50% item non-response from the candidate set. For the remaining variables, missing values are imputed using the sample mean for continuous variables and the modal category for categor-

[4] The original French wording is: *Le STATEC réalise régulièrement des études qualitatives (groupes de discussion, entretiens cognitifs etc.) portant sur des sujets variés. Ces études permettent au STATEC de mieux comprendre certaines thématiques et aussi de tester ces questionnaires afin d'améliorer leur qualité. Accepteriez-vous de participer à ce type d'études?*.

ical variables. The LASSO models are estimated both with and without the EU-SILC sampling weights; the selected predictors and their predictive performance are broadly similar across the two specifications. Consequently, the weighted specification is retained for the subsequent analysis.

The variables retained through LASSO selection are subsequently reviewed in light of established theoretical perspectives on survey participation, namely *Social Exchange Theory* and *Leverage–Saliency Theory*. *Social Exchange Theory* (Dillman, 1978) emphasises that participation decisions reflect respondents' evaluations of perceived benefits and costs, moderated by factors such as trust in institutions. *Leverage–Saliency Theory* (Groves et al., 2000), while sharing a cost–benefit foundation, further posits that participation is more likely when survey attributes are salient and aligned with respondents' interests. Although these frameworks are originally developed to model participation in a survey, they can be extended to guide the analysis of opting into an online panel. In this context, it is important to note that internet access and digital skills constitute key preconditions for participation in online surveys, as coverage and selection biases may persist even after statistical adjustment (e.g. Dever et al., 2008; Bethlehem, 2009).

3.3 Estimation of probabilities to opt-in

Because the opt-in decision is a binary choice, we estimate the probability of joining the online panel using a weighted logistic regression model (Hosmer et al., 2013; Wooldridge, 2010). Let Y_i denote the opt-in indicator, where $Y_i = 1$ if respondent i agrees to join the panel and $Y_i = 0$ otherwise. The probability of opt-in is modelled as

$$\Pr(Y_i = 1 \mid \mathbf{X}_i) = \frac{\exp(\mathbf{X}_i'\beta)}{1+\exp(\mathbf{X}_i'\beta)}, \tag{1}$$

where $\mathbf{X}_i$ is the vector of explanatory variables selected in the previous step and β is the vector of parameters to be estimated by maximum likelihood. Estimation is performed using the EU-SILC sampling weights to account for the complex survey design and to obtain estimates that are representative of the target population. The fitted model provides, for each respondent, the estimated probability of opting into the online panel,

$$\hat{p}_i = \Pr(Y_i = 1 \mid \mathbf{X}_i), \tag{2}$$

which is used in the subsequent step to construct inverse probability weights.

3.4 Inclusion probabilities and construction of panel weights

The opt-in weights are computed by taking the inverse of inclusion probabilities. These probabilities are calculated using the theorem of Bayes.

Inclusion in the panel requires two sequential events: participation in the 2019 Luxembourg EU-SILC survey (event B), and subsequent opt-in to the online panel (event A). The unconditional probability of inclusion ($P(A)$) is computed by multiplying the conditional probability of opting-in the panel given that the respondent has participated in the EU-SILC ($P(A \mid B)$) by the probability of participating in the EU-SILC ($P(B)$), as follows:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} -> P(A) = \frac{P(A \mid B) P(B)}{P(B \mid A)} \tag{3}$$

The corresponding panel weights are computed using the formula:

$$w_i^{\text{panel}} = \frac{1}{\widehat{P}(A \mid B)\, P(B)} = \frac{w_i^{\text{EU-SILC}}}{\widehat{P}(A \mid B)} \tag{4}$$

Note that $P(A \mid B)$ is the predicted probability of opt-in obtained from the estimation of the logit model described in the previous section 3.3. $P(B)$ is the inverse of the EU-SILC sampling weights. $P(B \mid A)$, the probability of having participated in EU-SILC given that the individual has opted-in to the panel, is equal to 1.

To reduce the influence of extreme weights, the panel weights are winsorized at the 95th percentile before calibration. Weight trimming is a standard practice in survey sampling to reduce the variance inflation caused by a small number of extreme weights while retaining all observations (Valliant and Dever, 2018).

3.5 Evaluation of the weighting procedures

To assess the effectiveness of the proposed weighting procedure, we compare the distributions of selected socio-demographic characteristics in the online panel with their corresponding distributions in the general population. The comparison is performed for both the unweighted panel and the panel weighted by the opt-in weights. Representativeness is assessed using two summary measures of discrepancy: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). Let P_j denote the official population proportion for category j and $\hat{P}_j$ the corresponding estimate obtained from the online panel. The two measures are computed as

$$\mathrm{MAE} = \frac{1}{J}\sum_{j=1}^{J}\left|\hat{P}_j - P_j\right|, \tag{5}$$

and

$$\mathrm{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\hat{P}_j - P_j\right)^2}, \tag{6}$$

where J denotes the total number of categories considered across the selected auxiliary variables. Lower values of MAE and RMSE indicate a closer agreement between the panel and the target population and, therefore, greater representativeness. The comparison is carried out using the main demographic characteristics commonly employed in survey calibration, namely age, sex, educational attainment, and area of residence.

4 Results

What follows presents the implementation of the procedure illustrated in the previous section. We focus on the analysis of the determinants of panel participation, and the assessment of the weights constructed from estimated inclusion probabilities.

4.1 Pre-selection of drivers of opting-in

Using EU-SILC sampling weights, the linear LASSO considered 382 candidate predictors. Ten-fold cross-validation selected a penalty parameter of $\lambda = 0.0066$, retaining 103 predictors. The selected model achieved the lowest cross-validated prediction error (0.126) and an out-of-sample R^2 of 0.184. In the training sample, the model attained a mean squared error (MSE) of 0.115 and an R^2 of 0.258, compared with an MSE of 0.128 and an R^2 of 0.176 in the testing sample. The moderate decline in predictive performance indicates good generalization and limited evidence of overfitting.

The 103 predictors retained by the LASSO are subsequently reviewed in light of previous empirical evidence from the survey methodology literature and interpreted according to the theoretical frameworks of *Social Exchange Theory* and *Leverage–Saliency Theory* discussed in the previous section. This process resulted in a parsimonious specification comprising 18 explanatory variables, grouped into four conceptual categories: *engagement and cooperation*, *digital access*, *socio-demographic characteristics*, and *perceived individual characteristics*.

Engagement and cooperation. Several variables reflect respondent engagement during the EU-SILC interview. *Attentiveness*, as rated by the interviewer, indicates conscientious participation and cooperation (Stoop, 2005). *Consent to use administrative data* reflects trust and low privacy concern, while prior *EU-SILC participation* captures continuity in engagement (Laurie et al., 1999, e.g.).

Digital access. *Internet access at home* is a basic prerequisite for online participation and a known predictor of response in web-based surveys (e.g. Dever et al., 2008; Bethlehem, 2009).

Socio-demographic characteristics. Standard background variables—*gender* (e.g. Becker, 2022), *age* (e.g. Blom et al., 2016), *education* (e.g. Bosnjak et al., 2013), *income* (e.g. Angel et al., 2019), *employment status* (e.g. Callegaro et al., 2015), *marital status*, *household position*, and *area of residence* (e.g. Goodman and Gatward, 2008)—are routinely associated with participation propensity. Immigration-related variables—*country of birth*, *nationality*, and *language used during the interview*—are also relevant (Méndez and Font, 2013), especially given the structure of Luxembourg's population [5].

Perceived individual characteristics. Variables such as *self-rated health* and *life satisfaction* are retained by the LASSO selection procedure. Both have been shown to be associated with survey participation, as individuals reporting better health and higher life satisfaction tend to be more likely to engage in surveys (Adams et al., 2019; Chadi, 2012). Perceived individual characteristics differ from socio-demographic characteristics in that they capture more subjective dimensions of individual.

4.2 Estimation of opt-in probabilities

This section presents estimates from the logit model of panel enrolment, and discusses the determinants of "opting-in". This implements step 3 of our procedure, and delivers estimates of opt-in probabilities. (That is, the conditional probability of opting-in the panel given that the respondent has participated in the EU-SILC.)

Table 1 reports the logit estimates together with standard goodness-of-fit statistics. Overall,

[5] Surveying immigrants can be challenging, and given that nearly half of the residents in Luxembourg do not hold Luxembourgish nationality STATEC (2022), this population can be particularly difficult to reach through surveys.

the model displays satisfactory predictive performance. The survey-adjusted Hosmer–Lemeshow test fails to reject the null hypothesis of good fit, indicating no evidence of model misspecification. The pseudo-R^2 is 0.233, which is relatively high for this type of model and comparable to the out-of-sample R^2 obtained from the LASSO.[6] Predictive performance is further assessed using the confusion matrix reported in Table 2. Evaluated on the test sample,[7] the model correctly classifies 82.8% of respondents who opted into the panel (sensitivity) and 67.8% of those who did not (specificity). Overall, these results indicate that the estimated model provides a reasonably accurate representation of respondents' opt-in behaviour.

Figure 2 reports the average marginal effects of the determinants of panel enrolment. Because all explanatory variables are categorical, their effects are directly comparable. The variable most strongly associated with opt-in is the consent to the use of administrative data. *Ceteris paribus*, respondents who provide such consent are 20 percentage points more likely to enroll in the panel than those who do not. (The size of the effect is substantial, given that the unconditional opt-in probability is roughly 15%.)

In addition, respondents who are less attentive during the interview are significantly less likely to enroll in the panel, with their probability of enrolment being 15 percentage points lower than that of highly attentive respondents. Similarly, individuals without internet access at home are 18 percentage points less likely to enroll.

[6] Pseudo-R^2 values between 0.2 and 0.4 are generally considered indicative of good model fit for nonlinear models (McFadden, 2021).

[7] The EU-SILC sample is randomly divided into training and test sets. The training sample is used to estimate the LASSO and logit models, while predictive performance is evaluated on the test sample.

Variables reflecting a cooperative attitude toward online survey participation are strongly associated with opt-in, both in terms of effect size and statistical significance, even after controlling for socio-demographic characteristics such as age and gender. It is worth noting, however, that these variables are typically unavailable as auxiliary information and, as a result, cannot be incorporated into weighting procedures. This suggests that weighting procedures based solely on socio-demographic characteristics may not fully correct for selection bias.

Table 1: Determinants of opting-in: logit coefficients.

Dependent variable: opt-in (Y/N)		
Engagement and cooperation		
Consent to use administrative data: no	ref.	
yes	1.569***	(0.088)
Attentive answers: not seriously	ref.	
Seriously	1.476***	(0.552)
Very seriously	1.764***	(0.550)
EU-SILC 2018 participation: no	ref.	
yes	0.278***	(0.088)
Digital access		
Access to internet at home: yes	ref.	
No - cannot afford it	-3.444***	(1.203)
No other reasons	-3.283***	(0.688)
Socio-demographic characteristics		
Gender: Man	ref.	
Woman	0.0249	(0.100)
Age: 18-30	ref.	
31-40	-0.164	(0.175)
41-50	-0.480***	(0.182)
51-60	-0.857***	(0.192)
61-70	-1.075***	(0.264)
71+	-1.691***	(0.310)
Education: primary	ref.	
Lower secondary	0.840***	(0.216)
Upper secondary	0.986***	(0.203)
Post-secondary non-tertiary	0.936***	(0.347)
Short-cycle tertiary	1.335***	(0.258)
Bachelor	1.510***	(0.232)
Master	1.787***	(0.216)
PhD	2.068***	(0.334)
Don't know/refused	0.107	(0.466)
Household income: Q1	ref.	
Q2	-0.171	(0.120)
Q3	-0.232*	(0.124)
Q4	-0.415***	(0.139)
Employment: Employed	ref.	
Unemployed	0.521	(0.339)
Retired	0.266	(0.326)
Studying	0.670**	(0.336)
Other inactive	0.113	(0.293)
Self-employed: no	ref.	
yes	-0.433*	(0.255)

Seeking job: currently working	ref.	
Yes	0.350	(0.304)
No	-0.0901	(0.283)
Marital status: Never married	ref.	
Married	-0.483***	(0.148)
Separated	-0.183	(0.411)
Widowed	-0.397	(0.315)
Divorced	-0.0872	(0.191)
PACS	-0.730***	(0.252)
Household position: Head	ref.	
Spouse	-0.365***	(0.125)
Partner	-0.368*	(0.220)
Son/daughter	-1.796***	(0.249)
Others	-2.619***	(0.586)
Refused/don't know	0.0652	(0.308)
Area of residence: Capellen	ref.	
Clervaux	-0.904**	(0.368)
Diekirch	-0.0935	(0.237)
Echternach	0.0855	(0.280)
Esch-sur-Alzette	0.301*	(0.163)
Grevenmacher	0.0280	(0.231)
Luxembourg	0.194	(0.170)
Mersch	0.522**	(0.219)
Redange	0.489*	(0.284)
Remich	0.654***	(0.227)
Vianden	-0.252	(0.515)
Wiltz	-0.610*	(0.343)
Country of birth: Luxembourg	ref.	
Born abroad	-0.141	(0.110)
Interview language: Lux.	ref.	
French	-0.0254	(0.111)
English	0.266	(0.210)
Perceived individual characteristics		
Health condition: very good	ref.	
Good	0.0477	(0.108)
Fair	0.165	(0.138)
Bad	-0.255	(0.224)
Very bad	-0.569	(0.468)
Don't know/refused	-2.160***	(0.746)
Life satisfaction: 1-completely unsatisfied	ref.	
2	-0.689	(0.892)
3	0.187	(0.715)
4	-0.264	(0.691)
5	-0.259	(0.601)

6	0.102	(0.601)
7	0.162	(0.592)
8	0.111	(0.592)
9	0.217	(0.597)
10- Completely satisfied	0.351	(0.604)
Constant	-4.182***	(0.805)
Additional Statistics		
Pseudo R^2	0.233	
LL_0	-234458	
LL	-179923	
Hosmer-Lemeshow goodness-of-fit	F(9, 8230)=0.551	
test for survey data	Prob > F =0.5838	
Obs.	8239	
Weighted Obs.	482656	

Standard errors in parentheses

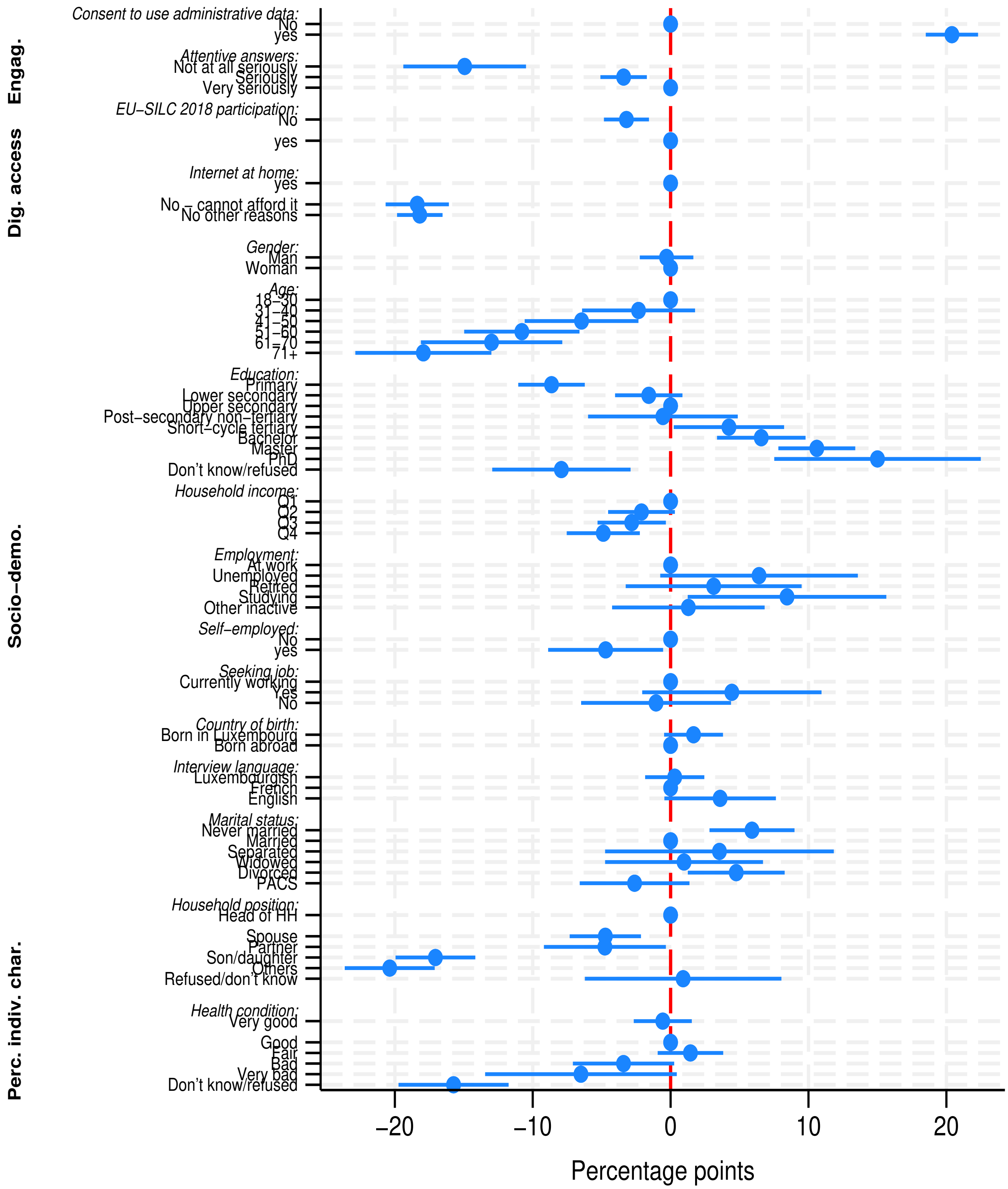
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Confusion Matrix: sensitivity and specificity of the logit model.

	Ratio		
	True Opt-in: Yes	True opt-in: No	Total
	b	b	b
Predicted Opt-in: Yes	82.8	32.2	42.4
Predicted opt-in: No	17.2	67.8	57.6
Total	100	100	100

Prediction of opt-in is 'yes' if model prediction is above the observed sample average of true opt-in

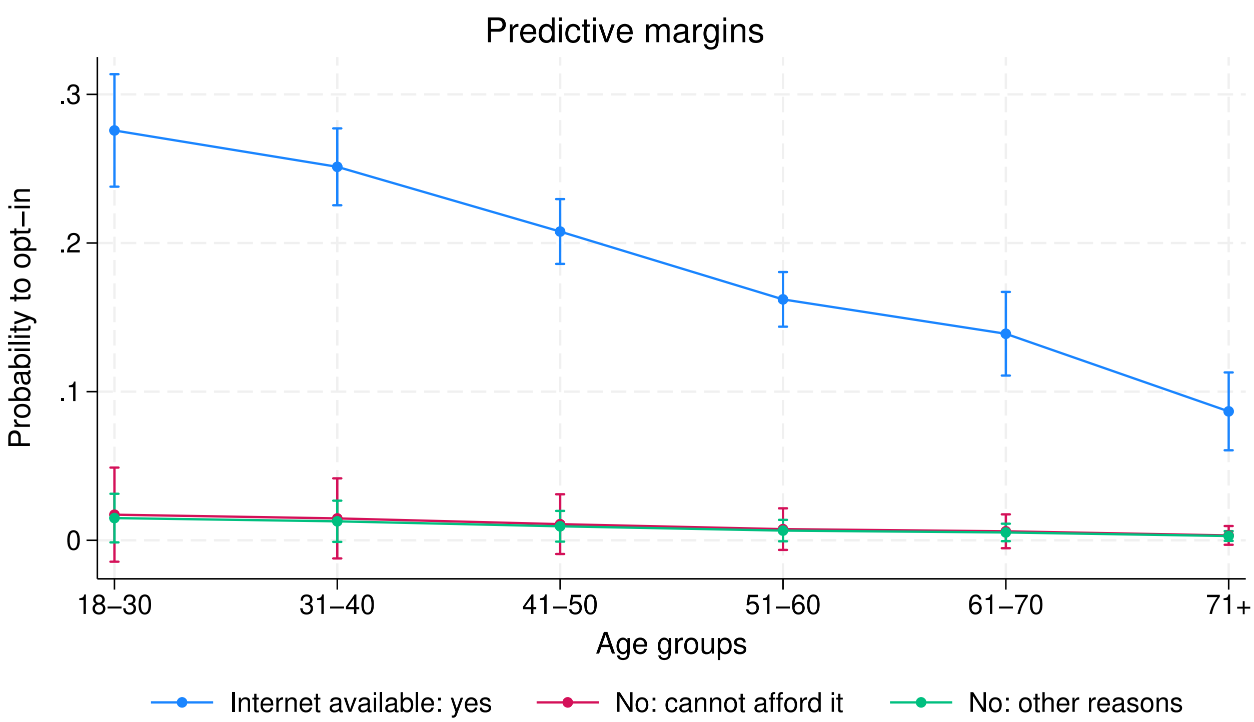
Figure 2: Determinants of the probability of opting-in: average marginal effects.



Note: lines represent 90% confidence intervals calculated with heteroscedasticity-robust standard errors. The estimates are based on regression, including life satisfaction and area of residence. The most frequent category is the baseline of categorical variables. All of the drivers of opt-in are categorical variables, allowing for a direct comparison of their respective effects. Dependent variable: opt-in to the online panel.

Among conventional socio-demographic variables, age and education play a prominent role in explaining the probability of opting-in. Respondents who are younger and have higher levels of education are more likely to opt-in. Figure 3 illustrates how estimated opt-in probabilities vary with age and access to the internet. Within each age group, respondents with home internet access are more likely to opt-in. Among respondents with internet access, however, estimated probabilities decrease with age, ranging from 27% for those aged 18-30 to about 9% for those aged over 70. Regarding other variables, respondents reporting worse health conditions, and those who are not the head of their household and have lower income, have lower probabilities of opting-in.

Figure 3: Probability of opting-in by age and internet at home.

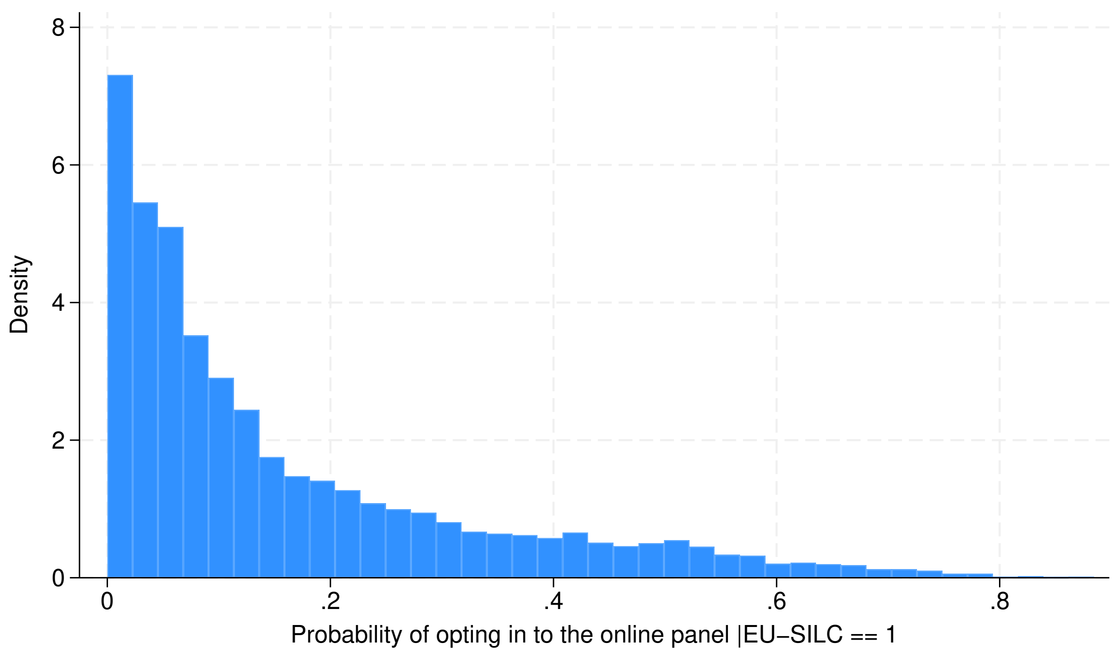


4.3 Opt-in probabilities and weights: descriptive statistics

What follows reports on the computation of the panel inclusion probabilities and the corresponding panel weights. We first describe the distribution of the estimated opt-in probabilities obtained from the logit model. Then, we evaluate the distribution of the resulting panel weights.[8]

Figure 4 displays the distribution of the estimated probabilities of opting-in, conditional on participation in the EU-SILC survey. The distribution is right-skewed, indicating that most respondents have a low predicted probability of enrolment. However, the long right tail reveals the presence of a minority of respondents with relatively high enrolment probabilities. This represents a non-negligible share of the sample.

Figure 4: Probability of opt-in: sample distribution.



Weights are constructed from inclusion probabilities computed from the theorem of Bayes as described in Section 3.4. For diagnostic purposes, Table 3 summarizes the distribution of the normalized panel weights. By construction, the weights have a mean of one. The distribution is moderately right-skewed, with a median of 0.68 and an interquartile range between 0.39 and

[8]Note that 'opt-in' probabilities denote those enrolment probabilities computed from the logit model. The 'inclusion probabilities' are those computed using the theorem of Bayes. Weights are the inverse of the latter probabilities.

1.27, indicating that most respondents receive relatively modest weighting adjustments. To limit the influence of highly influential observations, the weights are winsorized at the 95th percentile. Consequently, the maximum normalized weight is 3.60, implying that no respondent receives a weight exceeding 3.6 times the average weight.

Table 3: Diagnostics of the normalized panel weights.

	Value
Mean	1.000
Std. Dev.	0.918
Coefficient of variation	0.918
Minimum	0.004
25th percentile	0.385
Median	0.684
75th percentile	1.273
95th percentile	3.600
99th percentile	3.600
Maximum	3.600
Kish design effect	1.843
Effective sample size	701.722

The table also provides information on efficiency losses derived from our sampling scheme, in comparison to a pure random sample.[9] The coefficient of variation of the normalized weights is 0.92, corresponding to a Kish weighting design effect of 1.84. This indicates that the weighting increases the variance of survey estimates by approximately 84% compared to equal weighting. In other words, the effective sample size is reduced from 1,293 to approximately 702 respondents. Although weighting entails a moderate loss of efficiency, the design effect remains below 2, suggesting that the weights are sufficiently stable for reliable statistical inference.

For estimation purposes, the normalized weights are subsequently rescaled so that their sum equals the target population of 506,569 adults residing in Luxembourg.

[9]For a random sample, observations have equal weights, equal to 1/n where n is the number of observations.

4.4 Evaluation of the weighting procedures

To evaluate the effectiveness of the opt-in weights, we compare the distribution of socio-demographic characteristics in the online panel (weighted and unweighted) to those of the population using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Table 4 shows that both MAE and RMSE are consistently smaller when we apply the opt-in weights, indicating an overall improvement in the resemblance of the online panel to the population. The opt-in weights enhance the representativeness of the panel to the population in terms of four demographic variables: age, gender, education, and area of residence. In terms of age, the opt-in weighted data aligns more closely with population statistics compared to the unweighted data, as evidenced by the lower MAE and RMSE values. This pattern is particularly pronounced for individuals aged 71 years and older. Although individuals aged 71 years and older represent 11.5% of the general population, they account for only 3.9% of the unweighted sample. After applying the opt-in weights, their proportion increases to 6.5%. Similar patterns emerge for educational attainment. Individuals with lower secondary education are substantially under-represented in the unweighted sample, accounting for only 14.6% of respondents compared with 26.5% in the reference population. Applying the opt-in weights increases their estimated share to 21.0%, bringing the weighted distribution considerably closer to the population. Regarding gender and area of residence, the opt-in weights also result in improvements in both MAE and RMSE, although the impact of the opt-in weights is less pronounced compared to age and education. As noted in previous studies, the application of global adjustment procedures similar to the one used in this study may improve some estimates while having little or no effect on others. (e.g. Cornesse et al., 2020).

Overall, our findings suggest that accounting for inclusion probabilities improves the representativeness of online survey panels by over-weighting panellists with a lower propensity to opt-in.

Table 4: Comparison of online panel to general population characteristics: weighted vs. unweighted.

	Unweighted	Opt-in weights	Population
AGE			
18–30	20.5	18.9	21.7
31–40	25.8	22.1	19.4
41–50	20.9	18.7	18.4
51–60	17.6	19.2	17.3
61–70	11.4	14.5	11.7
71+	3.9	6.5	11.5
MAE	3.0	2.6	
RMSE	4.2	2.9	
GENDER			
Man	47.6	48.6	49.9
Woman	52.4	51.4	50.1
MAE	2.3	1.3	
RMSE	2.3	1.3	
EDUCATION			
Lower secondary	14.6	21.0	26.5
Upper secondary	35.6	39.1	34.1
Bachelor	18.4	15.4	17.6
Master or PhD	31.4	24.5	21.8
MAE	6.0	3.8	
RMSE	7.7	4.1	
AREA OF RESIDENCE			
Center	37.2	36.3	35.8
South	35.3	37.1	37.3
North	11.7	13.1	15.2
East	15.9	13.5	11.7
MAE	2.8	1.1	
RMSE	3.0	1.4	

Source: STATEC population statistics 2019 (gender, age and area of residence in the population column) and Labour Force Survey 2019 (education in the population column).

Note: MAE is the mean absolute error; RMSE is the root mean squared error.

5 Implementation of the online panel

It is worth noting that the probability-based online panel described in this article served as the sampling frame for two online surveys. These surveys, which investigated people's response to certain health policies implemented during the COVID-19 pandemic, allowed us to assess the panel's operational feasibility.

In the first survey, 84 invitation emails were undelivered because of invalid email addresses or full inboxes, leaving 1,216 successfully delivered invitations. Among these, 904 recipients accessed the survey, 812 started the questionnaire, and 730 completed it, corresponding to a response rate of approximately 60%. Considering that participation is entirely voluntary and the survey conducted exclusively online, this response rate is encouraging and compares favourably with those achieved by other general population surveys in Luxembourg (Riillo et al., 2020).

We conducted a second online survey among the 655 respondents who explicitly agreed to be re-contacted at the end of the first survey. We collected a total of 522 completed questionnaires, corresponding to a response rate of around 70% (Riillo and Peroni, 2021).

These surveys demonstrated the feasibility of recruiting a probability-based online panel from an official household survey, and highlighted key operational challenges, notably maintaining up-to-date contact information and mitigating panel attrition through regular replenishment and weighting adjustments.

The working papers Riillo et al. (2020) and Riillo and Peroni (2021) document the implementation of the survey and related empirical results.

6 Conclusions

This study proposes a procedure for the construction of a probability-based online panel from an official general population survey, the Survey on Income and Living Conditions (EU-SILC) conducted in Luxembourg. The goal is to mitigate the selection bias that characterises online panels. The study investigates the determinants of panel enrolment (”opting-in”), and assesses whether weighting based on estimated probabilities of enrolment improves the representativeness of the panel.

The proposed procedure is as follows. First, panellists are recruited from EU-SILC respondents that agreed to participate in future online surveys. Then, non-parametric and theory-driven variable selection approaches are used to pre-select potential determinants of opting-in. Next, a logistic regression model is estimated to predict the probabilities of panel enrolment. These predicted probabilities are then combined with known EU-SILC enrolment probabilities to derive ex-post inclusion probabilities and construct panel weights. The effectiveness of these weights is subsequently assessed by comparing weighted and unweighted estimates against official population benchmarks for age, gender, educational attainment, and area of residence.

The proposed opt-in model shows good predictive performance, correctly classifying 82.8% of respondents who opted into the panel and 67.8% of those who did not. Results from the opt-in model show that panel enrolment is driven not only by socio-demographic characteristics - primarily age and education - but also by indicators of respondents' attitudes towards, and ability to participate in, surveys. Consent to the use of administrative data, attentiveness during the interview, and home internet access are among the strongest predictors of opting-in. Such auxiliary information, however, is typically unavailable in conventional panels. Thus, this analysis suggests that weighting adjustments based solely on socio-demographic variables may have limited ability to fully correct for selection bias in online panels, and demonstrates the added value of recruiting panellists from official surveys.

Finally, the study shows that the use of the proposed opt-in weighting scheme improves the representativeness of the online panel. Compared with the unweighted panel, the weighted estimates exhibit a closer correspondence with the target population. The improvements are particularly pronounced for age and educational attainment, while more modest gains are observed for gender and area of residence.

More generally, the findings illustrate the value of official probability surveys as recruitment frames for online panels. Beyond their probabilistic sampling design, these surveys collect rich auxiliary information and paradata that can substantially improve the estimation of participation probabilities. Statistical offices can therefore exploit existing survey infrastructure to develop cost-effective online panels while preserving inferential quality.

Future research could examine the applicability of the proposed procedure to other official surveys in Luxembourg and abroad. Another extension would be to assess the impact of opt-in propensity weighting on substantive survey estimates. In particular, future studies could compare a two-step procedure, in which opt-in probabilities are first used to construct base weights and the resulting weights are then calibrated to population totals, with a single-step procedure that relies solely on calibration to population totals. These extensions remain a topic for future research.

Acknowledgments Thanks are due to Francesco Sarracino for insightful discussion, and to Guillaume Osier, Marc Pauly, and Johann Neumayr for their help and support. The Authors are grateful to Peter Lugtig, Claude Lamboray and other colleagues at STATEC and STATEC Research for advice and useful comments. Views and opinions expressed in this article are those of the author(s) and do not reflect those of STATEC and STATEC Research.

Author Contributions (CRediT)

Cesare A. F. Riillo: Conceptualization, Methodology, Formal analysis, Validation, Drafting – original draft.

Chiara Peroni: Drafting – review, Approval.

References

Adams, M. J., Hill, W. D., Howard, D. M., Dashti, H. S., Davis, K. A. S., Campbell, A., Clarke, T.-K., Deary, I. J., Hayward, C., Porteous, D., Hotopf, M., and McIntosh, A. M. (2019). Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *International Journal of Epidemiology*, 49(2):410–421.

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year?: explaining measurement error in survey income data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(4):1411 – 1437.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics. *Survey Methodology*, 46(1):1–28.

Becker, R. (2022). Gender and survey participation: an event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. *methods, data, analyses*, 16(1):30.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., and Weimer, D. L. (2003). The advent of internet surveys for political research: A comparison of telephone and internet samples. *Political Analysis*, 11(1):1–22.

Bethlehem, J. G. (2009). Selection bias in web surveys. *International Statistical Review*, 77(2):161–188.

Bethlehem, J. G. and Biffignandi, S. (2010). Incentives in web surveys: Methodological issues and a review. *International Journal of Market Research*, 52(4):491–517.

Biemer, P. P., de Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., and West, B. T. (2017). *Total survey error in practice*. John Wiley & Sons.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., and Krieger,

U. (2016). A comparison of four probability-based online and mixed-mode panels in europe. *Social Science Computer Review*, 34(1):8–25.

Blom, A. G., Gathmann, C., and Krieger, U. (2015). Setting up an online panel representative of the general population: The german internet panel. *Field methods*, 27(4):391–408.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., and Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods*, 25(4):339–360.

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (2015). *Online panel research: A data quality perspective*. John Wiley & Sons.

Chadi, A. (2012). I would really love to participate in your survey!: Bias problems in the measurement of well-being. *Economics Bulletin*, 32(4):3111–3119.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532):2011–2021.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., and Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Couper, M. P. (2007). Issues of representation in ehealth research (with a focus on web surveys). *American Journal of Preventive Medicine*, 32(5):S83–S89.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2(2):47–60.

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley-Interscience, New York.

Dutwin, D. and Buskirk, T. D. (2017). Apples to Oranges or Gala versus Golden Delicious?: Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples. *Public Opinion Quarterly*, 81(S1):213–239.

Eurostat. (2019). Quality report luxembourg eu-silc 2018.

Goel, S., Obeng, A., and Rothschild, D. (2015). Non-representative surveys: Fast, cheap, and mostly accurate. *Work Pap*.

Goodman, A. and Gatward, R. (2008). Who are we missing? area deprivation and survey participation. *European journal of epidemiology*, 23:379–387.

Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64(3):299–308.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley, 3 edition.

Jin, H. and Kapteyn, A. (2022). Relationship between past survey burden and response probability to a new survey in a probability-based online panel. *Journal of Official Statistics*, 38(4):1051–1067.

Laurie, H., Smith, R., and Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15(2):269–282.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2):329.

Lugtig, P. (2014). Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *"Sociological Methods & Research"*, 43(4):699–723.

MacInnis, B., Krosnick, J. A., Ho, A. S., and Cho, M.-J. (2018). The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension. *Public Opinion Quarterly*, 82(4):707–744.

Malhotra, N. and Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 anes to internet surveys with nonprobability samples. *Political Analysis*, 15(3):286–323.

McFadden, D. (2021). Quantitative methods for analysing travel behaviour of individuals: some recent developments. In *Behavioural travel modelling*, pages 279–318. Routledge.

Méndez, M. and Font, J. (2013). *Surveying ethnic minorities and immigrant populations: Methodological challenges and research strategies*. Amsterdam University Press.

Moreno-Betancur, M., Beaumont, J.-F., Bocci, C., Haziza, D., et al. (2025). *nonprobsvy: Statistical Inference Methods for Non-Probability Survey Samples*. R package version 0.2.0.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83:242–272.

Riillo, C. A. and Peroni, C. (2021). Acceptability of covid-19 contact-tracing apps in luxembourg: a panel analysis. *Economie et Statistiques*, 127:1–34.

Riillo, C. A., Peroni, C., and Sarracino, F. (2020). Determinants of acceptability of contact tracing apps for covid-19: initial results from luxembourg. *Economie et Statistiques*, 117:1–31.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*, volume 4. American Statistical Association Alexandria, VA.

Roberts, C., Herzing, J. E., Asensio Manjón, M., Abbet, P., and Gatica-Pérez, D. (2022). Response burden and dropout in a probability-based online panel study – a comparison between an app and browser-based design. *Journal of Official Statistics*, 38(4):987–1017.

Scherpenzeel, A. (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/Bulletin de Méthodologie Sociologique*, 109(1):56–61.

Schonlau, M., van Soest, A., Hurd, M., and Kapteyn, A. (2012). Selective attrition in panel data: The case of the health and retirement study. *Journal of Applied Statistics*, 39(5):1101–1114.

STATEC (2022). Luxembourg in figures.

Stoop, I. A. L. (2005). *The Hunt for the Last Respondent: Nonresponse in Sample Surveys*. Social and Cultural Planning Office of the Netherlands (SCP), The Hague.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Valliant, R. and Dever, J. A. (2018). *Survey weights: a step-by-step guide to calculation*, volume 1. Stata Press College Station, TX.

Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18(4):355–366.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Wood, M., Kunz, S., et al. (2014). Cawi in a mixed mode longitudinal design. Technical report, Understanding Society at the Institute for Social and Economic Research.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2 edition.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., and Wang, R. (2011). Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4):709–747.

7 Appendix