

Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey

André Grow¹  | Daniela Perrotta¹ | Emanuele Del Fava¹ |
Jorge Cimentada¹ | Francesco Rampazzo²  | Sofia Gil-Clavel¹  |
Emilio Zagheni¹ | René D. Flores³ | Ilana Ventura³ | Ingmar Weber⁴ 

¹Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany

²Department of Sociology, Leverhulme Centre for Demographic Science, Nuffield College, Oxford, UK

³Department of Sociology, University of Chicago, Chicago, Illinois, USA

⁴Social Computing Group, Qatar Computing Research Institute, Ar-Rayyan, Qatar

Correspondence

André Grow, Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, 18057 Rostock, Germany.

Email: grow@demogr.mpg.de

Abstract

Social scientists increasingly use Facebook's advertising platform for research, either in the form of conducting digital censuses of the general population, or for recruiting participants for survey research. Both approaches depend on the accuracy of the data that Facebook provides about its users, but little is known about how accurate these data are. We address this gap in a large-scale, cross-national online survey ($N = 137,224$), in which we compare self-reported and Facebook-classified demographic information (sex, age and region of residence). Our results suggest that Facebook's advertising platform can be fruitfully used for conducting social science research if additional steps are taken to assess the accuracy of the characteristics under consideration.

KEY WORDS

digital censuses, Facebook, online surveys, targeted advertising

1 | INTRODUCTION

Facebook's advertising platform provides aggregated information about the characteristics of the network's users (e.g., gender, age and interests) for targeted advertising. An increasing number

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](#) license, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

of scholars advocate for the use of these facilities in social science research, either for conducting digital censuses that aim to measure characteristics of the general population, or for recruiting participants for survey research (e.g., Alburez-Gutierrez et al., 2019; Alexander et al., 2019; Cesare et al., 2018; Pötzschke & Braun, 2017; Ribeiro et al., 2020; Rosenzweig et al., 2020; Schneider & Harknett, 2019; Zagheni et al., 2017). One reason is that traditional probability-based sampling methods, such as address-based sampling and random digit dialling, have proven increasingly costly, while response-rates and coverage have been declining for many segments of the general population (Stern et al., 2014). Against this backdrop, social media and big data—and Facebook's advertising platform in particular—are an attractive supplement for traditional survey research methods. They offer a potentially less expensive and more timely alternative (Amaya et al., 2020) and make it possible to generate samples of geographic or demographic subpopulations that would otherwise be difficult to reach (Zhang et al., 2020).

The feasibility of using Facebook for conducting digital censuses and generating samples of specific sub-populations depends on the accuracy of the data that underlies its advertising platform. Systematic misclassification of individual traits like gender and age could significantly increase costs and bias scholarly research. Researchers pay for clicks on their ads and having to discard participants outside of the intended population may increase the cost per usable questionnaire. Problems may be even worse for digital censuses, as researchers typically have no opportunity to independently validate their subjects' characteristics to quantify error in results. Unfortunately, Facebook neither offers much information on the accuracy of its userbase, nor how it determines the characteristics and interests partially or completely inferred from user behaviour on the network. Additionally, recent leaks have created doubts about the accuracy of the data that Facebook uses in ad targeting, rendering independent research on this topic ever more important (Fou, 2020). In this paper, we provide such research by comparing individuals' self-reported information in an online survey, where respondents are recruited using the Facebook advertising platform, with Facebook's classification of the same people. While information collected via surveys has its own limitations, our study sheds light on the extent to which data from Facebook's advertising platform, often considered a 'black box' (Araujo et al., 2017), can be trusted for research, as well as the extent to which the targeting features can be leveraged.

Our assessment is based on a large-scale, cross-national online survey. The survey was conducted in seven European countries (Belgium, France, Germany, Italy, the Netherlands, Spain and the United Kingdom) and in the United States, to collect information about behaviours and attitudes in response to the COVID-19 pandemic. Recruitment took place daily via targeted Facebook advertising campaigns that were stratified by users' sex, age and sub-national region of residence (such as the "West" of the United States, as defined by the US Census Bureau). In the questionnaire, respondents were asked to report these characteristics themselves. By comparing participants' answers with information about the specific ads through which respondents arrived at the survey, we could indirectly assess the accuracy of Facebook classification of these users. Given that sex, age and region of residence are commonly used stratification variables in social science research, and are known to relate to a large range of attitudes, behaviours and demographic outcomes (Geary, 2020; Lutz et al., 1998; Ribeiro et al., 2020), our work is relevant for many researchers who seek to use Facebook for social science research.

We are not the first to assess the accuracy of Facebook's advertising data (see, e.g., Pötzschke & Braun, 2017; Rosenzweig et al., 2020; Sances, 2021). However, our study goes beyond earlier work on this topic, by (1) taking a cross-national perspective, (2) assessing classification mismatches across the entire Facebook user population in the respective countries and (3) assessing the directionality of mismatches (e.g., did those incorrectly classified as 25–44 years old report to

be younger or older?). Data were collected between 13 March and 12 August 2020, resulting in a total of $N = 137,224$ questionnaires with complete information on respondents' sex, age and region of residence.

In summary we find that across countries, 99% of the survey answers matched Facebook's categorisation on at least two out of the three characteristics of interest. Specifically, the match between Facebook's categorisation and the users' answers was highest for sex (between 98% and 99% matches) and lowest for region of residence (between 91% and 98% matches). Based on these findings, we suggest that Facebook's advertising platform can be fruitfully used for conducting social science research, if additional steps are taken to assess the accuracy of the specific user characteristics that are in the focus of a given study.

2 | USING FACEBOOK'S ADVERTISING PLATFORM IN RESEARCH

Facebook is the largest social media platform, with 2.45 billion monthly active users worldwide, as of fall 2019 (Facebook Inc., 2019). Its business model centres on revenue from online advertising, which is technically implemented through the Facebook Ads Manager (FAM). The FAM allows advertisers to create ad campaigns that can have various goals, such as creating salience for a given service or product, or generating traffic to an external website. Each advertising campaign can target specific user groups, defined by select self-reported demographics (e.g., gender and age), and a set of characteristics that Facebook infers from the users' behaviour on the network (e.g., political orientation). Campaigns have three levels. At highest level, the goals of the campaign are defined (e.g., generating awareness or generating traffic). The second is the *ad set* level, at which the target audience, budget and ad delivery schedule are defined. The third level includes the advertisements themselves, which can consist of visual materials (e.g., images, videos), text, and a URL. Prior to launching a campaign, the FAM provides an estimate of the expected audience size given the selected combination of user characteristics (i.e., the number of daily or monthly active users who are eligible to be shown an ad). This allows advertisers to optimise their definition of target groups (Cesare et al., 2018).

Earlier social science research has used the FAM mostly in one of two ways. A first set of studies employed the FAM audience estimates prior to launching a campaign for obtaining digital censuses of the user population across geographic regions. The resulting information was then used to make inferences about specific social groups and the general population (e.g., Alexander et al., 2019; Kashyap et al., 2020; Rama et al., 2020; Rampazzo et al., 2018; Ribeiro et al., 2020; Zagheni et al., 2017). For example, Zagheni et al. (2017) used audience estimates to assess the share of foreign-born people living in the United States, comparing these numbers with data from the 2014 round of the American Community Survey (ACS). Their results showed that the Facebook audience estimates were qualitatively similar to the number of migrants observed in the ACS, which suggests that the FAM data can be used to study compositional population properties. One benefit of this approach is that the information that the FAM provides is updated continuously and can be collected programmatically through Facebook's application programming interface. This makes it possible to collect population data in a more continuous and timely manner than is possible with traditional censuses or register data (Ribeiro et al., 2020).

A second set of studies have used the FAM's targeted advertising facilities to recruit participants for survey research (e.g., Guillory et al., 2018; Kühne & Zindel, 2020; Pötzschke & Braun, 2017; Rinken et al., 2020; Rosenzweig et al., 2020; Sances, 2021; Schneider &

Harknett, 2019; Zhang et al., 2020). With this approach, researchers define one or more Facebook user groups whose members could be shown an ad that invites them to participate in an online survey. This ad is then displayed, for example, in the users' timelines, and directs users to an external webpage where they can participate in the survey. Given Facebook's reach, this approach is particularly attractive when the goal is to recruit members of sub-populations that account only for a small share of the overall population and that are difficult to identify in existing sampling plans (such as migrants or workers in specific industries) (e.g., Pötzschke & Braun, 2017). More recently, Zhang et al. (2020) have shown that targeted advertisements can also be used to collect representative samples of the general population, if the target groups in the advertising campaign are sufficiently fine grained.

Some of the studies that have used the FAM for survey recruitment have assessed the accuracy of the advertising information in reaching the targeted demographic groups. Pötzschke and Braun (2017), for example, reported for their survey of Polish migrants that about 98% of those who arrived at the survey via a Facebook ad lived in one of the targeted countries in Europe (Austria, Ireland, Switzerland and the United Kingdom) and 96% were indeed Polish migrants. By contrast, Rosenzweig et al. (2020) reported more variation in the observed matches when targeting users in two countries in the global South (Mexico and Kenya). While they reported a nearly 100% agreement between respondents' gender and Facebook's advertising data in Mexico, they only found about 13% matches for educational attainment in Kenya. Similarly, drawing on six studies in the United States, Sances (2021) reported that almost 100% of respondents who were classified by Facebook as 25 years and older also reported to be older than 24, whereas only about 23% of those who were classified as Black reported to be Black.

While insightful, these earlier studies have in common that they applied their recruitment criteria either to a single country or focused on a small subset of the larger population (circumscribed by demographic and social characteristics, and/or by place of residence). In this paper, we add to this body of literature by taking a cross-national perspective in which we assess classification mismatches across the entire Facebook user population in the targeted countries and assessing the directionality of mismatches. This provides additional insights into which users are more likely to be correctly or incorrectly classified. For example, our approach enables us to explore whether members of certain age groups are more likely to be misclassified than members of other age groups, and to explore the age groups to which they are incorrectly assigned.

3 | DATA AND METHODS

3.1 | Survey and Facebook advertising campaigns

This study uses data from the COVID-19 Health Behavior Survey (CHBS) (Del Fava et al., 2020; Grow et al., 2020; Perrotta et al., 2021). The CHBS is an anonymous, cross-national online survey that was conducted in Belgium, France, Germany, Italy, the Netherlands, Spain, the United Kingdom and the United States. Participation was voluntary and not incentivised. Data collection began on 13 March 2020 in Italy, the United Kingdom and the United States. Subsequent countries were added continuously, with Belgium joining last, on 4 April 2020. The data collection ended in all countries on 12 August 2020. The questionnaire had four sections, encompassing questions about respondents' socio-demographic characteristics, health indicators, behaviours and attitudes related to COVID-19 and social contacts. Our focus in this paper is exclusively on respondents' demographic characteristics.



FIGURE 1 Illustration of Facebook advertising campaign used in the United States. *Source:* Figure S1 in Perrotta et al. (2021) [Colour figure can be viewed at wileyonlinelibrary.com]

Participants were recruited via targeted Facebook advertising campaigns. The CHBS ran one ad campaign per country with the goal to generate traffic to the survey's webpage (there was one separate webpage per country). Facebook's ad delivery algorithms aimed to optimise ad delivery to increase the likelihood that users who were shown an ad clicked on it. Each campaign was stratified at the ad set level by users' gender (man or woman), age group (18–24, 25–44, 45–64 and 65+ years), and region of residence (see details in the next subsection), resulting in 24–56 strata per country. This stratification approach ensured a balance in central demographic characteristics of the resulting respondent samples, to which post-stratification techniques could be applied to improve representativeness (Grow et al., 2020; Holt & Smith, 1979; Little, 1993). Figure 1 illustrates the structure of the campaigns for the United States and an example of the advertisements.

3.2 | Inferring Facebook users' classification

Given the stratified nature of the advertising campaigns, we could infer how Facebook had classified the sex, age and region of residence of users from the ad through which they arrived at the survey. For example, a participant who arrived at the survey via an ad that targeted men ages 25–44 in the western United States should have reported a matching age, sex and region of residence in the survey. If his survey answers deviate from the above criteria, this might point to an error in Facebook's user classification, but it might also stem from reporting errors on the side of survey participants, either in the questionnaires or on Facebook. Accordingly, we interpret any difference between participants' answers and their classification by Facebook as bias, regardless of the exact cause of such differences. While Facebook users may see and click on ads that are not targeted at them, we exclude them from our analysis. This can happen, for example, when a Facebook friend of a non-targeted user comments on an ad, which then may appear as organic content in the non-targeted user's timeline.

The FAM allows advertisers to select from two *genders*, 'men' and 'women', which is based on user self-reported information (Facebook Inc., 2020), and which we used for stratifying the advertising campaigns. By contrast, in the CHBS questionnaire, respondents were asked to report their *sex*, with the options 'male' and 'female'. As biological sex and gender are not necessarily equivalent (West & Zimmerman, 1987; Westbrook & Saperstein, 2015), we explore the extent to which these terms overlap in this specific context.

Facebook usage is restricted to individuals aged 13 years and older, and advertisers can use single-year age categories to define their target, which is up to the age of 64 years. Older users are aggregated in the category 65+. In the CHBS questionnaire, respondents were asked to report their age in years, which makes it possible to map their answers onto the four age categories used for stratifying the advertisements. Participation in the CHBS was restricted to individuals age 18 and older, which is the lower age boundary in the advertising campaigns and in the survey data. Facebook employs users' self-reported age in its categorisation (cf. Facebook Inc., 2020; United States Securities and Exchange Commission, 2019).

Facebook offers several means for geographic targeting. For example, advertisers can draw on pre-defined regions, such as the state of California in the United States, or advertisers can define their own regions by selecting a geographic point of reference (defined by latitude and longitude) together with a radius around this point (in miles). User locations are estimated by Facebook based on several pieces of information, such as information from mobile devices, IP address and self-reported information (United States Securities and Exchange Commission, 2019). The CHBS advertising campaigns divided each targeted country into three to seven sub-national regions (here also called macro regions), which were composed of smaller micro regions. The micro regions were based on pre-defined regions offered by Facebook, largely following the NUTS-1 classification in Europe and the census regions in the United States (see Table A1 in Appendix S1 for an overview). The region-related answer categories in the CHBS questionnaire were largely identical to the micro regions that were used in the advertising campaigns.

The only notable exceptions from the region classification approach described above occurred in the United Kingdom and Spain. To minimise the possibility that the large metropolitan area of London dominated the daily recruitment efforts in England, two separate groups of ad sets were created. The first group focused on England while excluding London, whereas the second group only focused on London. This was achieved by defining a custom region centred on London with a radius of 17 miles that was selectively included in or excluded from the ad sets. In the case of Spain, the cities of Ceuta and Melilla in northern Africa were not included in the targeting. The reason is that targeting these cities by defining a radius around a geographic reference point would have led to the inclusion of parts of the surrounding African countries, which were not in the focus of the CHBS. These Spanish cities were therefore not included in the ad targeting, but respondents could select them from the set of answers in the CHBS questionnaire.

3.3 | Sample selection

Data were collected between 13 March and 12 August 2020. Over this period, 144,034 individuals completed the CHBS questionnaire, but we only considered the subset of respondents who arrived at the survey's page by clicking on an ad that was targeted at them and who reported their sex, age and region of residence in the survey. For consistency, in the Spanish data we also excluded respondents who reported to live in the cities Ceuta and Melilla in northern Africa (<1% of the sample for Spain). The final sample consisted of 137,224 individuals (95% of the original sample;

TABLE 1 Number of respondents per country and their distribution across sex and age as self-reported in the COVID-19 Health Behavior Survey questionnaire

Country	N	Sex (%)		Age (%)			
		Female	Male	18–24	25–44	45–64	65+
Belgium	12,657	65	35	14	29	36	21
France	13,430	69	31	16	29	35	20
Germany	25,707	59	41	17	37	32	15
Italy	15,651	67	33	16	39	31	14
Netherlands	11,280	64	36	11	22	40	27
Spain	13,345	69	31	6	35	43	16
United Kingdom	14,216	65	35	7	21	42	30
United States	30,938	63	37	8	24	36	32
Total	137,224	64	36	12	30	36	22

about 1% of the original sample reported to live in a country that was not in focus of the respective advertising campaign). Tables 1 and A2 in Appendix S1 show the distribution of respondents across countries, sex, age and regions. Compared to their respective national populations, female and older individuals were over-represented in the survey. As discussed in Grow et al. (2020), this bias can be addressed with post-stratification weighting to make the data more representative of the respective national populations (see also Perrotta et al., 2021), but in the analysis reported here, we use unweighted data, as we are not aiming to make statistical inferences about national populations.

3.4 | Analytical approach

We used standard classification-evaluation metrics to assess the quality of Facebook's user classification, namely classification *accuracy*, *precision*, *recall* and the F_1 score (Tharwat, 2020). All four measures are calculated based on a so-called confusion matrix, that cross-tabulates the actual category to which an object belongs (self-reported sex, age and region of residence) and the class to which it has been assigned by a prediction model (Facebook's user classification). Table 2 provides an example of such a confusion matrix, assuming that there is one characteristic with three categories. Each cell reports the number of respondents (n_{ij}) who were observed for each combination of actual (i) and predicted category (j). Cells along the main diagonal (n_{11} , n_{22} and n_{33}) report the numbers of respondents who were correctly classified, whereas all other cells represent incorrect classifications. Note that there is one such matrix for each characteristic and country.

Given this matrix, *accuracy* is defined as the fraction of respondents who were categorised correctly. This measure is calculated as

$$\text{accuracy} = \frac{n_{11} + n_{22} + n_{33}}{\sum_i \sum_j n_{ij}}. \quad (1)$$

Hence, the larger *accuracy*, the more likely that, for a given demographic characteristic, the answer of a randomly selected respondent matches with Facebook's user classification.

TABLE 2 Example of confusion matrix

		Predicted		
		Cat. 1	Cat. 2	Cat. 3
Actual	Cat. 1	n_{11}	n_{12}	n_{13}
	Cat. 2	n_{21}	n_{22}	n_{23}
	Cat. 3	n_{31}	n_{32}	n_{33}

The *accuracy* measure provides a general assessment of the overall quality of the classification, but it has two shortcomings. First, it does not consider that the distribution of correct and incorrect classifications may differ between different categories of the same characteristic (e.g., in the case of sex, there might be more correct classifications for male than for female respondents). Second, if the number of observations across categories is imbalanced, the results tend to be biased towards the dominant category (e.g., if there were more male than female respondents in the sample, the correct and incorrect classifications of male respondents may dominate the results) (Chawla, 2010). The measures *precision* and *recall* address these issues by looking at each category separately. In more detail, *precision* is calculated as the fraction of the predictions for a given category j that were correct. This measure is calculated as

$$\text{precision}_j = \frac{n_{jj}}{\sum_i n_{ij}}. \quad (2)$$

By contrast, *recall* is the fraction of actual instances of category i that were predicted correctly. It is calculated for a given class i as

$$\text{recall}_i = \frac{n_{ii}}{\sum_j n_{ij}}. \quad (3)$$

Hence, *precision* indicates how many of the observations that were predicted to belong to category i actually belonged to category i (e.g., how many of the individuals who were predicted to be male actually reported to be male?), whereas *recall* indicates how many of the observations that actually belonged to category i were correctly predicted to belong to this category (e.g., how many of the individuals who reported to be male were correctly predicted to be male?).

The measures *precision* and *recall* assess different aspects of the confusion matrix, but they do not provide an overall assessment of the classification per category. The F_1 score (F_1 from here on) provides such an assessment, and is calculated as the harmonic mean of *precision* and *recall* for a given category i as

$$F_{1,i} = 2 \frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}. \quad (4)$$

Hence, $F_{1,i}$ will be close to one when both precision_i and recall_i are close to one, but $F_{1,i}$ will be lower when precision_i and/or recall_i are lower. In the discussion of our results, we focus on F_1 as a summary measure, and refer to *precision* and *recall* if there are marked differences between them for a given characteristic.

Assessing the uncertainty that surrounds the above outcome measures is complicated by the fact that the (Facebook user) population that survey respondents come from is unknown

to us. We therefore use percentile bootstrapping (Efron & Tibshirani, 1993), which is a common non-parametric approach to calculate confidence intervals for parameter estimates from unknown populations. With this approach, for each country c , we take 10,000 resamples (with replacement) of size N_c from the original sample (where N is the size of the original sample) and calculate the outcome measures for each new sample. The 95% confidence interval corresponds to the area between the 2.5 and 97.5 percentiles of the resulting bootstrap coefficients. Generally, as reported in the tables below, the 95% confidence intervals are quite narrow.

4 | RESULTS

4.1 | Accuracy

Table 3 reports the shares of respondents who were classified correctly between Facebook and the CHBS on zero, one, two or three characteristics. Across countries, between 86% and 93% of respondents were correctly classified on all three characteristics: sex, age and region of residence. The share of completely correct classifications was lowest in Belgium and France, and highest in the Netherlands. Among those respondents who did not have a perfect match on all three characteristics, typically only one characteristic was incorrect, and very few respondents had only one or no matching characteristics (<2%).

Table 4 reports *accuracy* values across countries and characteristics. Classification accuracy was highest for sex, ranging from 0.980 in France and the Netherlands to 0.987 in Italy and the United States. This means that between 98% and 99% of all classifications were correct. For age, classification accuracy was somewhat lower, ranging from 0.925 in France to 0.963 in the Netherlands. Classification accuracy was lowest for region of residence, and there was somewhat more variation across countries, with values ranging from 0.909 in Belgium to 0.981 in the United States.

TABLE 3 Share of respondents for which zero, one, two or three of their reported characteristics (sex, age and region of residence) matched with Facebook's classification

Country	Correct characteristics (%)			
	0	1	2	3
Netherlands	<1	1	6	93
Italy	<1	1	7	92
United States	<1	1	7	92
Germany	<1	1	8	91
Spain	<1	1	9	90
United Kingdom	<1	1	13	87
Belgium	<1	1	13	86
France	<1	1	12	86

Note: Cells show row percentages; rows sorted by column '3'.

TABLE 4 Accuracy (95% CI) for sex, age and region by country

Country	Accuracy		
	Sex	Age	Region
Belgium	0.982 (0.980, 0.984)	0.959 (0.956, 0.963)	0.909 (0.904, 0.914)
France	0.980 (0.977, 0.982)	0.925 (0.921, 0.930)	0.944 (0.941, 0.948)
Germany	0.984 (0.982, 0.985)	0.948 (0.945, 0.950)	0.970 (0.968, 0.972)
Italy	0.987 (0.986, 0.989)	0.951 (0.948, 0.955)	0.972 (0.970, 0.975)
Netherlands	0.980 (0.978, 0.983)	0.963 (0.960, 0.967)	0.984 (0.982, 0.986)
Spain	0.985 (0.982, 0.987)	0.934 (0.930, 0.938)	0.972 (0.970, 0.975)
United Kingdom	0.986 (0.984, 0.988)	0.941 (0.938, 0.945)	0.929 (0.925, 0.933)
United States	0.987 (0.985, 0.988)	0.942 (0.940, 0.945)	0.981 (0.980, 0.983)

4.2 | Precision, recall, and F_1 for sex categories

Table 5 shows *precision*, *recall* and F_1 measures by sex across countries. Generally, the combined measure of F_1 was high for both male and female respondents, but it was consistently higher for female than for male respondents by a margin of about 0.008 to 0.014 points across countries. *Precision* was typically higher among female respondents, whereas *recall* was higher among male respondents. Hence, classification of respondents as women by Facebook were more likely to match with respondents' answers on their sex than classifications as male (*precision*), whereas those who reported to be male were more likely to be classified correctly than those who reported to be female (*recall*). However, while consistent across countries, these differences were relatively small.

4.3 | Precision, recall and F_1 for age categories

Compared to sex, we found more variability by country in the match between Facebook's classification of age and respondents' answers. As Table 6 shows, the overall classification quality (F_1) was highest for the age category 25–44 years (average $F_1 = 0.958$ across countries), and lowest for the age categories 18–24 years and 65+ years (average $F_1 = 0.925$ and $F_1 = 0.929$, respectively). We observed the lowest single value of F_1 for the category 18–24 years in the United Kingdom ($F_1 = 0.855$), and the highest value for the category 25–44 years in Belgium ($F_1 = 0.971$). There were also systematic differences in *precision* and *recall* across age groups. Those who were classified as 18–24 years had a comparatively lower likelihood of reporting membership to this age group (average *precision* = 0.871 across countries), whereas they were more likely to be correctly classified when they reported being 18–24 years old (average *recall* = 0.987 across countries). The opposite was the case for the age category 45–65 years (average *precision* = 0.977 and *recall* = 0.917, respectively). For example, in the United Kingdom, about 75% of those who were classified as 18–24 years old reported an age in this range, whereas about 99% of those who reported to be 18–24 years old were also classified as such. By contrast, about 98% of those who were classified as 45–64 years old reported an age in this range, whereas only about 92% of those who reported being

TABLE 5 *Precision, recall* and F_1 (95% CI) for the different categories of sex by country

Country	Precision		Recall		F_1	
	Female	Male	Female	Male	Female	Male
Belgium	0.994 (0.993, 0.996)	0.960 (0.955, 0.966)	0.978 (0.975, 0.981)	0.989 (0.986, 0.992)	0.986 (0.984, 0.988)	0.975 (0.971, 0.978)
France	0.994 (0.992, 0.995)	0.950 (0.943, 0.957)	0.977 (0.974, 0.980)	0.986 (0.982, 0.989)	0.985 (0.983, 0.987)	0.968 (0.964, 0.971)
Germany	0.992 (0.991, 0.994)	0.972 (0.969, 0.975)	0.980 (0.978, 0.982)	0.989 (0.987, 0.991)	0.986 (0.985, 0.987)	0.980 (0.978, 0.982)
Italy	0.995 (0.994, 0.997)	0.972 (0.968, 0.976)	0.986 (0.983, 0.988)	0.991 (0.988, 0.994)	0.991 (0.989, 0.992)	0.981 (0.979, 0.984)
Netherlands	0.993 (0.991, 0.995)	0.958 (0.952, 0.964)	0.976 (0.972, 0.979)	0.988 (0.984, 0.991)	0.984 (0.982, 0.986)	0.973 (0.969, 0.976)
Spain	0.989 (0.987, 0.991)	0.974 (0.969, 0.979)	0.988 (0.986, 0.990)	0.977 (0.972, 0.981)	0.989 (0.987, 0.990)	0.975 (0.972, 0.979)
United Kingdom	0.996 (0.995, 0.998)	0.968 (0.963, 0.973)	0.982 (0.979, 0.985)	0.994 (0.991, 0.996)	0.989 (0.988, 0.991)	0.981 (0.978, 0.983)
United States	0.994 (0.993, 0.995)	0.974 (0.972, 0.977)	0.985 (0.983, 0.986)	0.990 (0.988, 0.992)	0.989 (0.988, 0.990)	0.982 (0.980, 0.984)

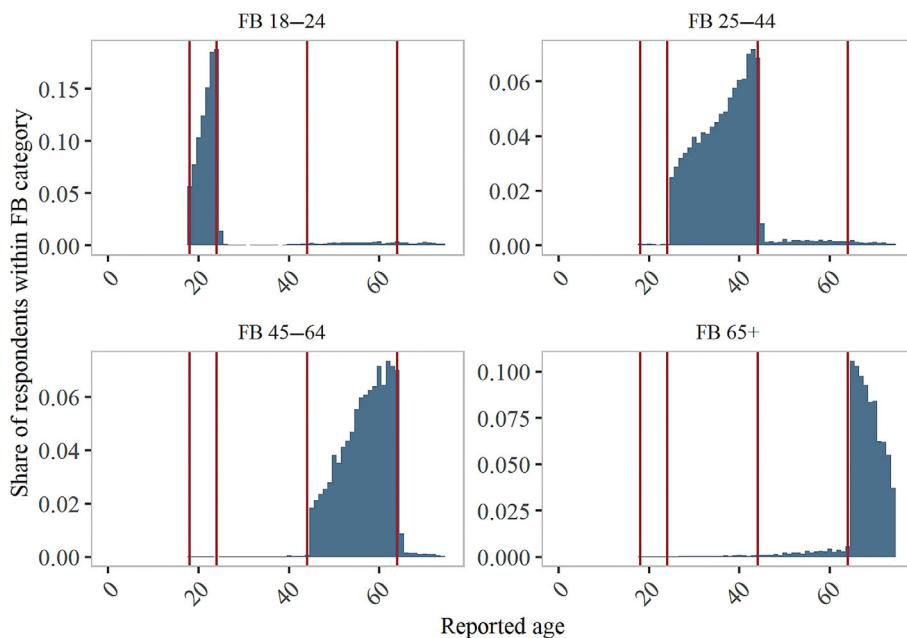


FIGURE 2 Share of respondents who reported a given age by Facebook's (FB) age classification. The red, vertical lines indicate age-group boundaries. Plot has been truncated at age 75 [Colour figure can be viewed at wileyonlinelibrary.com]

45–64 years old were also classified as such. For the other age groups, the differences between *precision* and *recall* were weaker and less systematic across countries.

Because respondents reported their age in single years, we also assess the directionality of misclassifications. Figure 2 shows respondents' reported age and the age category to which Facebook assigned respondents (data pooled from all countries). The red vertical lines demarcate the boundaries of the different age groups. Congruent with the fact that *precision* was typically above 90% across countries and age groups, the mass of the age distributions fell within the boundaries of the respective age groups to which respondents had been assigned by Facebook. Yet, within these age groups, there was a marked skew towards the upper boundary, except for the oldest age group (65+ years), which had a skew towards the lower boundary.

4.4 | Precision, recall and F_1 for region of residence

Table A3 in Appendix S1 shows the *precision*, *recall* and F_1 measures for each of the different regions across countries. Overall, we show high classification quality. The F_1 score mostly varied between 0.925 (for the “England” region within the United Kingdom) and 0.993 (for the “Northern Ireland” region also in the United Kingdom). The only outliers were the regions of Brussels in Belgium and London in the United Kingdom, with F_1 scores of 0.787 and 0.791, respectively. The values of *precision* and *recall* were generally high and did not differ systematically across countries. This indicates that across countries, individuals who were classified as living in a given region by Facebook often also reported living in the same region (*precision*), and most respondents

TABLE 6 Precision, recall, and F_1 (95% CI) for the different categories of age by country

Country	Precision			Recall			F_1					
	18–24	25–44	45–64	65+	18–24	25–44	45–64	65+	18–24	25–44	45–64	65+
Belgium	0.923 (0.910, 0.935)	0.960 (0.954, 0.966)	0.978 (0.974, 0.982)	0.952 (0.944, 0.960)	0.994 (0.990, 0.998)	0.982 (0.977, 0.986)	0.937 (0.930, 0.944)	0.945 (0.936, 0.953)	0.957 (0.950, 0.964)	0.971 (0.967, 0.975)	0.957 (0.953, 0.961)	0.949 (0.942, 0.955)
France	0.863 (0.850, 0.877)	0.931 (0.923, 0.939)	0.968 (0.963, 0.973)	0.906 (0.894, 0.916)	0.988 (0.983, 0.992)	0.966 (0.960, 0.972)	0.882 (0.873, 0.891)	0.890 (0.878, 0.902)	0.921 (0.913, 0.929)	0.948 (0.943, 0.953)	0.923 (0.917, 0.929)	0.898 (0.889, 0.906)
Germany	0.931 (0.923, 0.938)	0.962 (0.958, 0.966)	0.976 (0.972, 0.979)	0.881 (0.871, 0.891)	0.991 (0.988, 0.993)	0.961 (0.957, 0.965)	0.908 (0.902, 0.914)	0.950 (0.943, 0.957)	0.960 (0.955, 0.964)	0.962 (0.959, 0.964)	0.941 (0.937, 0.945)	0.914 (0.908, 0.921)
Italy	0.929 (0.919, 0.939)	0.957 (0.952, 0.962)	0.982 (0.978, 0.986)	0.900 (0.888, 0.913)	0.981 (0.975, 0.986)	0.973 (0.969, 0.977)	0.917 (0.909, 0.922)	0.932 (0.922, 0.943)	0.954 (0.948, 0.960)	0.965 (0.961, 0.968)	0.949 (0.944, 0.953)	0.916 (0.908, 0.925)
Netherlands	0.893 (0.877, 0.909)	0.946 (0.938, 0.955)	0.987 (0.984, 0.990)	0.974 (0.969, 0.980)	0.995 (0.991, 0.998)	0.981 (0.976, 0.986)	0.944 (0.937, 0.951)	0.964 (0.957, 0.970)	0.941 (0.932, 0.950)	0.963 (0.958, 0.968)	0.965 (0.961, 0.969)	0.969 (0.964, 0.973)
Spain	0.832 (0.808, 0.834)	0.941 (0.934, 0.947)	0.973 (0.969, 0.977)	0.873 (0.859, 0.886)	0.974 (0.963, 0.984)	0.973 (0.968, 0.977)	0.903 (0.895, 0.910)	0.918 (0.907, 0.930)	0.897 (0.883, 0.911)	0.956 (0.952, 0.961)	0.936 (0.932, 0.961)	0.895 (0.885, 0.905)
United Kingdom	0.753 (0.730, 0.778)	0.922 (0.913, 0.932)	0.981 (0.978, 0.985)	0.957 (0.951, 0.963)	0.987 (0.979, 0.994)	0.972 (0.966, 0.978)	0.922 (0.915, 0.928)	0.937 (0.907, 0.945)	0.855 (0.839, 0.870)	0.947 (0.941, 0.952)	0.951 (0.946, 0.954)	0.947 (0.942, 0.952)
United States	0.846 (0.832, 0.859)	0.916 (0.910, 0.922)	0.968 (0.964, 0.968)	0.964 (0.960, 0.968)	0.988 (0.984, 0.992)	0.981 (0.978, 0.984)	0.924 (0.919, 0.929)	0.922 (0.917, 0.927)	0.911 (0.903, 0.919)	0.948 (0.944, 0.951)	0.945 (0.942, 0.948)	0.943 (0.939, 0.946)

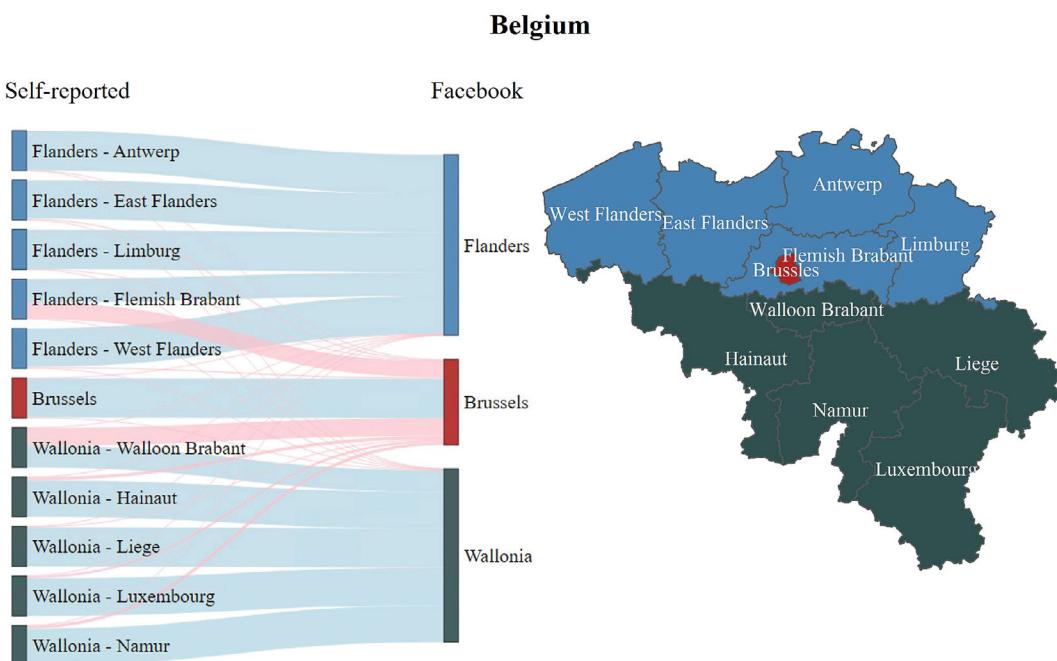


FIGURE 3 Respondents' reported region in comparison with Facebook's categorisation in Belgium. Blue lines indicate correct classifications, red lines indicate incorrect classifications [Colour figure can be viewed at wileyonlinelibrary.com]

who reported living in a given region were also correctly classified by Facebook (*recall*). Again, the only marked exceptions were Brussels and London, for which *precision* tended to be lower than *recall* (0.654 vs. 0.987 for Brussels and 0.662 vs. 0.983 for London, respectively). Hence a large share of respondents who were classified as living in Brussels or London by Facebook reported living in other regions (*precision*), whereas those who reported living in Brussels or London were usually correctly classified by Facebook (*recall*).

The fact that respondents reported detailed regions of residence (micro regions) enables us to assess the directionality of misclassifications. In Figures 3 and 4, we look closer at the misclassifications that occurred in Belgium and the United Kingdom, respectively. Focusing first on Belgium, Figure 3 shows that the low precision for the region of Brussels was largely due to respondents who reported living in the Flemish Brabant and Walloon Brabant regions, but who were classified by Facebook as living in the nearby region of Brussels. Correspondingly, the *recall* values for Flanders and Wallonia in Table A3 were somewhat lower than their *precision* values. Focusing next on the United Kingdom, Figure 4 shows that most misclassifications for the region of London concerned respondents who reported living in the East and the South East of England, which are the two regions that geographically surround London. Notably, a substantial share of respondents who reported living in South West England were misclassified as living in the adjacent region of Wales.

5 | DISCUSSION AND CONCLUSION

In this paper, we examined whether the information that FAM provides about its user database is accurate enough to be used in social science research. We compared the sex, age and region

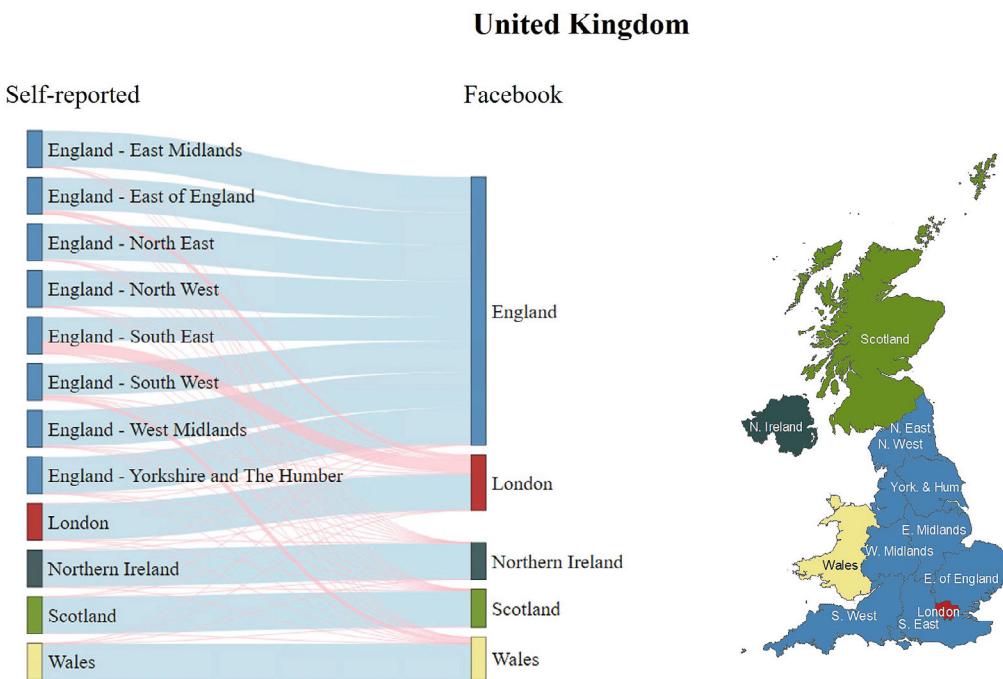


FIGURE 4 Respondents' reported region in comparison with Facebook's categorisation in the United Kingdom. Blue lines indicate correct classifications, red lines indicate incorrect classifications [Colour figure can be viewed at wileyonlinelibrary.com]

of residence that participants of an anonymous online survey reported with Facebook's classification of the same individuals used in its advertising algorithms. We relied on the CHBS, which recruited its participants via targeted ads on Facebook in eight countries. Our results showed that there was a very good, albeit imperfect, match between respondents' self-reported characteristics and Facebook's classification. Across countries, about 86%–93% of respondents' answers matched Facebook's categorisation on all three characteristics that we considered. Misclassifications were most likely to occur for region of residence and least likely to occur for sex.

Why was the error rate for region of residence higher than for sex and age? One possible explanation is that Facebook's gender and age classifications are largely based on self-reported information that is not very likely to change, or to change predictably, over time. By contrast, users' region of residence is partially inferred by Facebook and may change frequently, thereby increasing the chance for erroneous classifications. Interestingly, most of the incorrect region classifications concerned people who reported living in regions that were adjacent to those to which they were incorrectly assigned by Facebook. The largest share of misclassifications concerned respondents who Facebook had classified as living in Brussels and London, but who reported living in the surrounding areas. It seems likely that daily commuting for work from the surrounding suburbs may have contributed to the large number of classification errors that we observed. This result parallels the findings of Sances (2021), who reported that classifications were more likely to be correct in larger regions than in smaller regions in the United States. In the case of London in our study, this trend may have been aggravated by the fact that the targeting was based on geographic radius around the centre of London, rather than its exact borders, as was the case for the

other regions that we considered. Given that the actual shape of London is more complex than a simple circle, this approach may have inadvertently included Facebook users who lived close to the border of London, but not in London itself. A final error source may be residential moves among users that occurred shortly before participating in the survey, which may not have been reflected in Facebook's advertising data yet. Future research could assess the extent to which this may affect the accuracy of the data that Facebook provides, by querying participants about past residential changes.

Regarding age, we observed distinct misclassification patterns across age groups, as well as distinct participation patterns within age groups. Those who were classified as 18–24 years old were least likely to report an age in this interval, whereas those classified as 45–64 years old were most likely to report an age in this interval. Conversely, those who reported being 18–24 years old were most likely to be correctly classified by Facebook, whereas those who reported being 45–64 years old were least likely to be correctly classified. If respondents' survey answers were truthful, there is the possibility that a substantial share of 45- to 64-year-old Facebook users mis-reported their age when registering on the social network. Alternatively, if respondents correctly indicated their age on Facebook, younger survey participants may have reported to be older than they are. Facebook has acknowledged that information on age among younger users may be less accurate (United States Securities and Exchange Commission, 2019), but with our data, we cannot adjudicate between these sources of bias.

Additionally, we observed notable skews in the age distributions within age groups. Younger age groups skewed towards the upper age boundary, whereas the oldest age group (65+ years) skewed towards the lower age boundary. These patterns may result from at least three interacting processes. First, the CHBS is a health-related survey and the ads showed health-related content (see Figure 1 for an example; see Grow et al. (2020) for all images used in the campaigns). Older adults tend to be more interested in health-topics than younger individuals (Pew Research Center, 2015) and COVID-19 tends to have more negative health outcomes for older individuals (Nikolich-Zugich et al., 2020). Hence, within each stratum of the CHBS advertising campaigns, older Facebook users may have been more likely to click on the ads and participate in the survey, thereby leading to a skew in the age distribution within the different strata. Second, Facebook's advertising algorithms are designed to maximise the likelihood that users who are shown an ad click on it. If older users were more likely to engage with the CHBS ads, Facebook's advertising algorithms may have reinforced the resulting skew by preferentially targeting older users. Unfortunately, we cannot determine whether the observed age patterns in our data are (at least partially) the result of Facebook's advertising algorithms. However, our results underscore the importance of stratifying advertising campaigns on important demographic characteristics, such as age, if the goal is to obtain representative samples of the population of Facebook users. Third, the pattern observed in the oldest age group may result from the relatively fewer very old individuals on Facebook (cf. Gil-Clavel & Zagheni, 2019). The skew towards the lower age boundary in the age group 65+ years may simply reflect the age structure in this segment of the Facebook user population.

Some of the observed mismatches in sex may be due to users whose gender identity differs from their biological sex (Facebook uses gender rather than biological sex for targeting ads), who have non-binary gender identities, or whose gender identity has changed over time without updating their gender on Facebook. This may reduce the likelihood that Facebook assigns them to a gender category that aligns with their biological sex. We cannot directly assess this potential source of bias, but our results show that even though Facebook's user categorisation is based on gender, this information can be used reliably to recruit respondents of a specific sex. Note that

transgender, gender fluid, or non-binary respondents may have opted for the category 'prefer not to answer' when asked for their sex. In this case, they would not be included in the analyses presented in this paper.

Our assessment of Facebook's advertising data improves on earlier work on this topic by taking a cross-national perspective, by studying the entire demographic spectrum of Facebook's user base, and by exploring in detail the directionality of observed mismatches. Yet, there are also some caveats. First, our work is not a direct assessment of the accuracy of Facebook's user classification algorithms. Mismatches between Facebook's classification and participants' self-reported characteristics may stem from a genuine misclassification on Facebook's side, but respondents may also have misreported (either on purpose or by accident) their characteristics in the survey. Conversely, users may have misreported their characteristics on Facebook. While our study provides information about the likelihood of misclassifications, and which characteristics are particularly affected by it, it does not provide insights into the causes of these misclassifications.

Furthermore, our results apply to Facebook users who are actively using the social network, who are willing to participate in online surveys, and who have an interest in health-related topics. Additionally, the highly educated were somewhat over-represented in our sample (cf. Perrotta et al., 2021), which is congruent with the observation that more educated people generally are more likely to participate in survey research (Spitzer, 2020). These aspects may be problematic for several reasons. For example, the accuracy of Facebook's classification may be lower among people who are less active on the platform, if those who use Facebook less frequently are also less likely to keep their profile information up to date. We cannot assess this possibility with the CHBS data, but future research could assess this possibility by including questions about social media use in surveys that use Facebook for participant recruitment. Furthermore, people who are less inclined to participate in surveys may generally be more concerned about their privacy, and this may be associated with less accurate reporting of personal characteristics to Facebook. As these individuals were less likely to take part in the CHBS, we may have inadvertently overestimated the accuracy of Facebook's advertising data.

Relatedly, the topic of the CHBS and the special circumstances of the COVID-19 pandemic may have affected participant recruitment, especially during the early months of the pandemic. As Grow et al. (2020) show, the inclination of Facebook users to click on the CHBS ads was particularly high in April and May 2020 but levelled off thereafter. Furthermore, Grow et al. (2020) obtained estimates of the monthly and daily active users of Facebook for each day of the survey period per country, which suggest that during these months users were more active on Facebook than usual. Indeed, at the start of the survey period, the virus was novel, which may have precipitated elevated interest and participation in the survey. At the same time, in many countries, lockdown measures may have increased users' available time for social media survey participation. To the extent that survey participation and Facebook usage are associated with concerns about the accuracy of the information that Facebook has about individuals, as described above, this may have affected our results.

Finally, the strata that the CHBS used for ad targeting were comparatively broad, and smaller categories may be associated with more classification errors. Based on our results for region, and in particular the results for Brussels and London, we expect that researchers interested in smaller spatial resolutions, such as cities or towns, will experience more errors, as it likely will become more difficult for Facebook to classify users correctly. Unfortunately, the CHBS data do not allow us to assess this possibility, so we encourage future

research to assess this possibility systematically. Another avenue for future research is to explore the causes of observed misclassifications. For example, certain ages can be associated with different stigma for different groups, and this may systematically affect the direction of mismatches (see England & McClintock, 2009, for a discussion on sex-specific age stigma).

These caveats notwithstanding, our work has practical implications for scholars who want to use Facebook's Ads Manager in social science research. Our results suggest that the FAM is a valuable and largely reliable tool for research, given that Facebook's user categorisation matched the self-reported central demographic characteristics reported in our survey. At the same time, there were some mismatches, and their number varied between countries and between the different categories of the characteristics that we considered. We therefore suggest that scholars who want to use the FAM conduct (pre-test) surveys among the Facebook sub-population of interest to assess the accuracy of the provided user information. This can help to gauge the costs per usable questionnaire in survey research in advance, and to assess the uncertainty that surrounds point estimates of digital censuses. For example, if the goal is to study Turkish immigrants in Germany, researchers could target this group via Facebook ads and invite them to participate in a short demographic survey, in which their country of birth and immigration status are queried and later compared with Facebook's classification of the same users.

Our work also has implications for big data social science research at large. Big data are increasingly seen as an attractive supplement for survey research, as they offer a potentially 'less expensive, less burdensome, and more timely alternative for producing a variety of statistics' (Amaya et al., 2020, p. 90). At the same time, the use of big data comes with its own methodological challenges. One challenge is assessing bias in big data (Amaya et al., 2020; Baker, 2017; Schober et al., 2016; Sen et al., 2019). Most notably, Amaya et al. (2020) recently suggested assessing big data in a way similar to the Total Survey Error (TSE) framework, calling this new approach the Total Error Framework (TEF). The TSE has been established to quantify bias in survey research, encompassing all research steps from defining the inferential population to questionnaire design and drawing inferences. The TEF applies this approach to big data, considering error sources that may occur, for example, during data identification and extraction. The approach that we have presented here can contribute to both, the TSE and the TEF. In terms of the TSE, the FAM has been likened to sampling frames that are often used in survey research; like other sampling frames, FAM too suffers from systematic under-coverage of certain segments of the population (e.g., those who do not have a Facebook account). Our findings, as well as the approach proposed in this paper, contributes to the assessment of sampling error using FAM that arises when there is discrepancy between Facebook's user classification and respondents' actual characteristics. Similarly, in the case of the TEF, our approach provides insights into the biases that may emerge when researchers use the FAM for conducting digital censuses, as described above. These findings also open the door to further studies in multi-mode and multiple sample-frame survey research, and the possibility to target different segments of the population through different sample-frames and modes according to population coverage in each, including across national borders.

DATA AVAILABILITY STATEMENT

The aggregated data that underlie the calculation of the classification quality measures are available from the corresponding author upon request.

ACKNOWLEDGEMENT

Open Access funding enabled and organized by Projekt DEAL.

ORCID

André Grow  <https://orcid.org/0000-0003-2470-0071>

Francesco Rampazzo  <https://orcid.org/0000-0002-5071-7048>

Sofia Gil-Clavel  <https://orcid.org/0000-0003-4707-849X>

Ingmar Weber  <https://orcid.org/0000-0003-4169-2579>

REFERENCES

Alburez-Gutierrez, D., Aref, S. & Gil-Clavel, S. (2019) Demography in the digital era: New data sources for population research. In: Arbia, G., Peluso, S., Pinna, A. et al. (Eds.) *Book of short papers SIS2019*. Pearson, pp. 22–33. <https://doi.org/10.31235/osf.io/24jp7>

Alexander, M., Polimis, K. & Zagheni, E. (2019) The impact of Hurricane Maria on out-migration from Puerto Rico: evidence from Facebook data. *Population and Development Review*, 45, 617–630. <https://doi.org/10.1111/padr.12289>

Amaya, A., Biemer, P.P. & Kinyon, D. (2020) Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, 8, 89–119. <https://doi.org/10.1093/jssam/smz056>

Araujo, M., Mejova, Y., & Weber, I. (2017) Using Facebook ads audiences for global lifestyle disease surveillance: promises and limitations. *Proceedings of the 2017 ACM on web science conference*, Troy New York: ACM, 25 June 2017, pp. 253–257. <https://doi.org/10.1145/3091478.3091513>.

Baker, R. (2017) Big data: a survey research perspective. In *Proceedings of statistics Canada symposium 2016: Growth in statistical information: challenges and benefits*, pp. 47–69. <https://doi.org/10.1002/9781119041702.ch3>.

Cesare, N., Lee, H. & McCormick, T. (2018) Promises and pitfalls of using digital traces for demographic research. *Demography*, 55, 1979–1999. <https://doi.org/10.1007/s13524-018-0715-2>

Chawla, N.V. (2010) Data mining for imbalanced datasets: an overview. In: Maimon, O. & Rokach, L. (Eds.) *Data mining and knowledge discovery handbook*. Boston, MA: Springer, pp. 875–886. https://doi.org/10.1007/978-0-387-09823-4_45

Del Fava, E., Cimentada, J., & Zagheni, E. (2020) The differential impact of physical distancing strategies on social contacts relevant for the spread of COVID-19. *medRxiv*. <https://doi.org/10.1101/2020.05.15.20102657>.

Efron, B. & Tibshirani, R.J. (1993) *An introduction to the bootstrap. Monographs on Statistics and Applied Probability*. Springer Science and Business Media.

England, P. & McClintock, E.A. (2009) The gendered double standard of aging in US marriage markets. *Population and Development Review*, 35, 797–816.

Facebook Inc. (2019) *Facebook reports third quarter 2019 results*. Available at: <https://investor.fb.com/investor-news/press-release-details/2019/Facebook-Reports-Third-Quarter-2019-Results/default.aspx> [Accessed 15th April 2020].

Facebook Inc. (2020) *Age and gender*. Available at: <https://business.facebook.com/business/help/717368264947302?id=176276233019487> [Accessed 15th March 2021].

Fou, D. A. (2020) About that leaked employee email saying that facebook targeting was crappy. Available at: <https://www.forbes.com/sites/augustinefou/2021/12/28/about-that-leaked-employee-email-saying-that-facebook-targeting-was-crappy/> [Accessed 28th March 2022].

Geary, D.C. (2020) *Male, female: the evolution of human sex differences*, 3rd edition. Washington, DC: American Psychological Association.

Gil-Clavel, S. & Zagheni, E. (2019) Demographic differentials in Facebook usage around the world. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, pp. 647–650.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S. et al. (2020) Addressing public health emergencies via Facebook surveys: Advantages, challenges, and practical considerations. *Journal of Medical Internet Research*, 22, e20653. <https://doi.org/10.2196/20653>

Guillory, J., Wiant, K.F., Farrelly, M., Fiacco, L., Alam, I., Hoffman, L. et al. (2018) Recruiting hard-to-reach populations for survey research: Using Facebook and Instagram advertisements and in-person intercept in LGBT bars and nightclubs to recruit LGBT young adults. *Journal of Medical Internet Research*, 20, e197. <https://doi.org/10.2196/jmir.9461>

Holt, D. & Smith, T.M. (1979) Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142, 33–46.

Kashyap, R., Fatehkia, M., Tamime, R.A. & Weber, I. (2020) Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demographic Research*, 43, 779–816.

Kühne, S. & Zindel, Z. (2020) Using Facebook and Instagram to recruit web survey participants: A step-by-step guide and application. *Survey Methods: Insights from the Field*, 1–16. <https://doi.org/10.13094/SMIF-2020-00017>

Little, R.J.A. (1993) Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88, 1001–1012. <https://doi.org/10.1080/01621459.1993.10476368>

Lutz, W., Goujon, A. & Doblhammer-Reiter, G. (1998) Demographic dimensions in forecasting: Adding education to age and sex. *Population and Development Review*, 24, 42–58.

Nikolich-Zugich, J., Knox, K.S., Rios, C.T., Natt, B., Bhattacharya, D. & Fain, M.J. (2020) SARS-CoV-2 and COVID-19 in older adults: what we may expect regarding pathogenesis, immune responses, and outcomes. *GeroScience*, 42, 505–514. <https://doi.org/10.1007/s11357-020-00186-0>

Perrotta, D., Grow, A., Rampazzo, F., Cimentada, J., Del Fava, E., Gil-Clavel, S. et al. (2021) Behaviors and attitudes in response to the COVID-19 pandemic: Insights from a cross-national Facebook survey. *EPJ Data Science*, 10, 17.

Pew Research Center. (2015) Public interest in science and health linked to gender, age and personality. *Numbers, Facts and Trends Shaping the World*, 1–25.

Pötzschke, S. & Braun, M. (2017) Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review*, 35, 633–653.

Rama, D., Mejova, Y. & Tizzoni, M. (2020) Facebook ads as a demographic tool to measure the urban-rural divide. *Proceedings of the web conference*, pp. 327–338. Available at: <https://doi.org/10.1145/3366423.3380118> [Accessed 24th February 2021].

Rampazzo, F., Zagheni, E. & Weber, I. (2018) Mater Certa Est, pater Numquam: What can Facebook advertising data tell us about male fertility rates? *Proceedings of the international AAAI conference on web and social media*.

Ribeiro, F.N., Benevenuto, F. & Zagheni, E. (2020) How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. *WebSci'20: 12th ACM conference on web science*.

Rinken, S., Domínguez-Álvarez, J.A., Trujillo, M., Lafuente, R., Sotomayor, R. & Serrano-del-Rosal, R. (2020) Combined mobile-phone and social-media sampling for web survey on social effects of COVID-19 in Spain. *Survey Research Methods*, 14(165–170), 2. <https://doi.org/10.18148/srm/2020.v14i2.7733>

Rosenzweig, L., Bergquist, P., Pham, K.H., Rampazzo, F. & Mildenberger, M. (2020) Survey sampling in the Global South using Facebook advertisements. *SocArXiv*. <https://doi.org/10.31235/osf.io/dka8f>.

Sances, M.W. (2021) Missing the target? Using surveys to validate social media ad targeting. *Political Science Research and Methods*, 9, 215–222. <https://doi.org/10.1017/psrm.2018.68>

Schneider, D. & Harknett, K. (2019) What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, 51, 108–140.

Schober, M.F., Pasek, J., Guggenheim, L., Lampe, C. & Conrad, F.G. (2016) Social media analyses for social measurement. *Public Opinion Quarterly*, 80, 180–211. <https://doi.org/10.1093/poq/nfv048>

Sen, I., Floeck, F., Weller, K., Weiss, B. & Wagner, C. (2019) A total error framework for digital traces of humans. arXiv:1907.08228 [cs.CY]. Available at: <http://arxiv.org/abs/1907.08228> [Accessed 15th March 2021].

Spitzer, S. (2020) Biases in health expectancies due to educational differences in survey participation of older Europeans: It's worth weighting for. *The European Journal of Health Economics*, 21, 573–605. <https://doi.org/10.1007/s10198-019-01152-0>

Stern, M.J., Bilgen, I. & Dillman, D.A. (2014) The state of survey methodology: challenges, dilemmas, and new frontiers in the era of the tailored design. *Field Methods*, 26, 284–301. <https://doi.org/10.1177/1525822X13519561>

Tharwat, A. (2020) Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>

United States Securities and Exchange Commission. (2019) *Annual report on form 10-K*. Facebook Inc. Available at: <https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm> [Accessed 15th March 2021].

West, C. & Zimmerman, D.H. (1987) Doing gender. *Gender & Society*, 1, 125–151. <https://doi.org/10.1177/0891243287001002002>

Westbrook, L. & Saperstein, A. (2015) New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society*, 29, 534–560.

Zagheni, E., Weber, I. & Gummadi, K. (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43, 721–734.

Zhang, B., Mildenberger, M., Howe, P.D., Marlon, J., Rosenthal, S.A. & Leiserowitz, A. (2020) Quota sampling using Facebook advertisements. *Political Science Research and Methods*, 8, 558–564. <https://doi.org/10.1017/psrm.2018.49>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S. et al. (2022) Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S343–S363. Available from: <https://doi.org/10.1111/rssa.12948>