Combined mobile-phone and social-media sampling for web survey on social effects of COVID-19 in Spain

Sebastian Rinken, Juan-Antonio Domínguez, Manuel Trujillo, Regina Lafuente, Rafaela Sotomayor and Rafael Serrano-del-Rosal[1]

Institute for Advanced Social Studies (IESA-CSIC)
Córdoba, Spain

Published in *Survey Research Methods* 14(2): 165-170 (2020)
(https://doi.org/10.18148/srm/2020.v14i2.7733)

To minimize coverage and participation biases, this web survey adopted two complementary sampling strategies: (a) SMS invitations to random-number mobile phones, and (b) Internet advertisements targeting specific sociodemographic groups. This combination of aleatory selection from a rather comprehensive sampling frame, on one hand, and responsive definition of target profiles, on the other, largely succeeded in adjusting the sample to relevant population parameters, albeit with the noteworthy exception of education level.

Keywords: mixed sampling design; mobile-phone sampling; Facebook Ads sampling; CAWI

1. Introduction

This paper reports on ESPACOV ("Estudio Social sobre la Pandemia de COVID-19"), a web survey conducted in Spain in week four of the strict lockdown imposed on March 14th, 2020, in reaction to one of the world's worst COVID outbreaks at that time. ESPACOV was initiated and run by the Institute for Advanced Social Studies (IESA), a unit of Spain's Research Council (CSIC). The aim of collecting timely data on the pandemic's social dimension, including assessments of actual and potential non-pharmaceutical Interventions (NPIs), entailed an experiment with innovative sampling procedures. Lockdown-induced constraints meant that all survey-related tasks had to be accomplished by IESA staff from their home-offices. Staff had access to CAWI tools but not CATI equipment, a problematic limitation because Internet surveys are subject to coverage and participation biases that would ideally be corrected by complementary off-line data (de Leeuw, 2018). More problematic still, a complete sampling frame for our target population (residents aged 18+ years) was unavailable.

To maximize representativeness despite such unpromising context conditions, ESPACOV employed a mixed sampling design inspired by the responsive approach (Groves & Heeringa, 2006), combining two complementary sampling strategies with distinct properties and monitoring the streams of process data to achieve less biased estimates. Subsample A was generated by random-number text-messaging (SMS) invitations to mobile phones. Aleatory selection from this rather comprehensive frame emulates probability sampling as best as possible, given the circumstances. However, apart from potentially disappointing response rates, we had to anticipate under-coverage of some sociodemographic groups (Bosnjak, Neubarth, Couper, Bandilla, & Kaczmirek, 2008; Mavletova & Couper, 2014). Subsample B

[1] We thank our colleagues Lourdes Biedma and Isabel García for their contribution to the ESPACOV study.

was devised to improve the overall sample's fit with population parameters by targeting Internet ads at specific profiles.[2]

2. Sample design

2.1. Subsample A

Availing ourselves of a smartphone penetration rate of about 79%, exceeding 85% except among elderly people[3], the sampling frame comprises all mobile numbers assigned for residential use by Spain's telecoms regulator[4]. This frame includes 93,230,000 possible numbers, 47.1% of which were active as of Autumn 2019[5]. The primary sample was drawn by generating random mobile numbers whose volume was proportionate to the ranges assigned to each prefix; these mobiles were sent SMS invitations with links to the questionnaire. Based on prior projects, we expected a response rate of around 3%; we prepared to send 67,000 SMS to obtain an effective subsample of about 800 completed questionnaires.

Table 1. Penetration rates of various information and communication technologies by sex and age group (Spain, 2019)

	Age group	(a) Mobile phone	(b) Internet	(c) Mobile internet	(d) SMN
Men	18-29	99.7%	99.6%	99.0%	90.3%
	30-44	97.8%	97.2%	93.8%	65.4%
	45-64	96.9%	91.5%	84.7%	45.0%
	65+	78.5%	50.0%	39.1%	13.6%
Women	18-29	99.3%	98.4%	97.3%	88.3%
	30-44	98.6%	98.2%	96.2%	75.2%
	45-64	97.5%	92.3%	86.6%	51.7%
	65+	74.7%	43.5%	36.8%	13.9%
Total		93.0%	84.0%	79.1%	52.6%

Source: National Statistical Institute (INE-TIC) (see footnote 3).

2.2. Subsample B

Provisions were made to recruit a second subsample via paid advertisements on Google Ads and Facebook/Instagram. The former's penetration rate approaches 100% of Internet users, while the second comprises Spain's two principal social media networks (SMNs), with Instagram's audience noticeably younger than Facebook's[6]. Predictably subject to a combination of coverage bias (especially for SMNs, cf. Table 1, item d) and self-selection bias

[2] An additional, respondent-driven subsample C shall serve for studying cumulative selection bias; it will not be considered in this paper.

[3] The National Statistical Institute's survey on equipment and use of information and communication technologies (INE-TIC, www.ine.es) estimates that in 2019, 79% of Spain's population (aged 18+ years) used mobile Internet outside their home or office, and hence smartphones; cf. Table 1, item (c). Mobile phone penetration rates are higher still (Table 1, item a), achieving acceptable coverage even among elderly people; however, URL links sent via SMS to non-smart mobiles have to be transferred manually to a web browser, a threshold that makes survey participation less likely.

[4] https://numeracionyoperadores.cnmc.es/numeracion.

[5] http://data.cnmc.es/. We found no data on multiple phone ownership, a distorting variable.

[6] https://iabspain.es/estudio/estudio-anual-de-redes-sociales-2019/

associated with ultra-low ratios between ad displays and completed questionnaires (Stern, Bilgen, McClain, & Hunscher, 2017), subsample B was devised for time-lagged (consecutive) launch with a view to complementing subsample A. Both platforms allow to direct ads at specific profiles, but do not inform about procedural details.

3. Fieldwork instrument and management

The questionnaire was developed and programmed during weeks two and three of the lockdown; most items regard assessments of NPIs. Questionnaire extension, item semantics, and response formats were optimized for self-administration with a range of mobile devices. All procedures were designed to guarantee anonymity without requiring authorization under Europe's demanding data protection norms: no personal data were collected at any stage. Since only "exempted" cookies, non-parametric links, and user-unspecific passwords were employed, we relied on a questionnaire item to classify respondents by subsample. SMS invitations and Internet ads mentioned the survey's sponsor, subject-matter, and anonymity guarantee (exact wording differed due to platform requirements). All respondents converged on the same questionnaire (encuestacovid19.net).

SMS were sent starting April 4th at a rate of between 800 and 1200 messages per hour, 12 hours per day, to cover time slots convenient for different recipient profiles and protect our server against excess traffic[7]. Arguably due to factors such as the topic's outstanding interest, lockdown-induced extra margins of spare time, respondent-friendly questionnaire design, and trust in privacy-protection, the response rate almost doubled our baseline expectations. Thus, recruitment for subsample A was discontinued on April 7th with 51,046 SMS sent at a cost of 1,838€, 49.2% of which (25,126) were delivered; proportions with range assignments were maintained. The effective size of subsample A amounts to N=1,379 (5.6% of delivered SMS). However, this success was tempered by larger-than-expected underrepresentation of some age segments and, especially, of people with low levels of formal education (cf. Tables 2 and 3, "subsample A").

Since last-minute issues impeded the timely use of Google Ads, we had to rely exclusively on Facebook/Instagram, a SMN platform which places ads in associated apps and websites, for subsample B. After a test-run without target restrictions had evidenced patterns similar to those in subsample A, we first excluded people with university education and/or aged 45 to 64 years, then focused entirely on people with basic education only and either younger than 30 years, or elder than 65. From April 8th through 11th the ad was displayed 1,337,856 times to a total of 1,187,580 people, and was clicked on 3,974 times by 3,752 different individuals, 994 of whom completed the questionnaire (cost: 512€). While the overall response rate is ultra-low (0.08%), as anticipated, the click-to-response rate amounts to a respectable 26%. A vast majority (85%) of clicks occurred on smartphones.

4. Obtained sample

Tables 2 and 3 compare unweighted sample distributions by age group and sex and by education level, respectively, with expected distributions and population data. Expected distributions are estimated with penetration rates of mobile internet use (subsample A) and social media networking (subsample B) (cf. Table 1), assuming zero participation bias (i.e.,

[7] This risk was tangible due to the option of respondent-driven recruitment (cf. footnote 2), which generated around 18.000 additional questionnaires.

equal response rates across all sociodemographic features). Regarding subsample A, any differences between expected and obtained values depict plain selection bias. Regarding subsample B, things are more complicated: in Table 2, the "expected" column reflects the age and sex distribution without profile targeting (i.e., diverging from the general population in line with SMN coverage), whereas in Table 3, it estimates the distribution by education level based on the obtained age and sex distribution. Thus, in Table 2, comparison between both subsample B columns illustrates a mix of self-selection bias and the effect of targeted sampling, while in Table 3, that comparison reveals both a failure of profile targeting as far as education level is concerned, and added selection bias: rather than being substantially lower (as it should have, given our targeting instructions), the share of tertiary education is even higher than the subsample's actual age and sex distribution would suggest.

Table 2. Expected and obtained sample distribution by sex and age group, and comparison with population parameters

		Subsample A		Subsample B		**Sample (A+B)**	Population
	Age	Expected	Obtained	Expected	Obtained		
Men	18-29	9.9%	6.4%	13.6%	9.7%	**7.7%**	7.6%
	30-44	15.8%	18.3%	16.6%	9.3%	**14.5%**	13.5%
	45-64	18.7%	20.9%	15.0%	11.3%	**16.9%**	17.3%
	65+	4.8%	6.2%	2.5%	16.1%	**10.3%**	10.1%
Women	18-29	9.2%	7.9%	12.5%	10.6%	**9.0%**	7.3%
	30-44	16.2%	18.0%	19.0%	13.7%	**16.2%**	13.3%
	45-64	19.6%	18.9%	17.6%	17.9%	**18.5%**	17.6%
	65+	5.8%	3.5%	3.3%	11.6%	**6.9%**	13.3%
TOTAL		100%	100%	100%	100%	**100%**	100%

Sources: Serrano-del-Rosal et al., 2020; INE-TIC (cf. footnote 3); INE (Padrón).

Table 3. Expected and obtained sample distribution by education level, and comparison with population parameters

	Subsample A		Subsample B		**Sample**	
Education	Expected	Obtained	Expected	Obtained	**(A+B)**	Population
Obligatory or less	33.8%	14.7%	33.7%	16.2%	**15.3%**	44.0%
Secondary (non-oblig.)	26.7%	29.1%	26.7%	33.8%	**31.0%**	23.2%
Tertiary	39.4%	55.7%	39.6%	49.6%	**53.2%**	32.8%
TOTAL	100%	99.5%	100%	99.6%	**99.5%**	100%

Sources: Serrano-del-Rosal et al., 2020; INE-TIC (cf. footnote 3).
Notes: "Obligatory or less" comprises all situations up to completing the compulsory part of secondary education. Column totals short of 100% are due to item non-response.

Regarding age and sex, subsample A's obtained distribution diverges little from expectations, indicating scant selection bias; still, older people (65+years) are subject to sizable coverage bias, as predicted. Subsample B corrects the under-representation of older men quite successfully, less so for women. However, regarding formal education, our sampling strategy did not work well. Due to a combination of coverage and participation bias, people with tertiary education are vastly over-represented, and less educated people vastly under-represented in subsample A. Subsample B fails to correct this problem - we can only speculate that this might be due, at least partially, to incorrect ad platform records on the education level of Facebook/Instagram users. The ensuing distortion further accentuates in the 65+ age group, where the sample share of people with tertiary education (54.6%) almost quadruples the population value.

The sample's geographical distribution (urban/rural; regions) is generally satisfactory, except for some underrepresentation (perhaps due to language issues) of Spain's largely Catalan-speaking East (regions of Cataluña, Balearic Islands, and Valencia).

Raw data were weighted by iterative (raking) adjustments regarding municipality size, region (aggregated as NUTS-1), age group, sex, and education level. First results were published on April 24th on the CSIC website. On May 14th, the dataset was made available online (Serrano-del-Rosal et al., 2020).

5. Discussion

To maximize representativeness amidst a population lockdown that restricted data collection to CAWI, the ESPACOV survey employed a mixed sampling design that combines two complementary sampling strategies with distinct properties. Emulating the logic of probability sampling as best as possible in this difficult context, subsample A was generated by random-number SMS invitations to mobile phones. This procedure achieved very satisfactory response rates, with minor qualifications for some sociodemographic groups; however, it proved subject to substantive coverage bias with regard to age and education level. Subsample B was devised to improve the overall sample's alignment with population parameters by targeting paid advertisements on Facebook/Instagram at specific respondent profiles. This strategy was quite successful regarding the age distribution, but failed with regard to education level, possibly because of incorrect user data on the ad platform's part. In retrospect, excluding elderly people would have improved formal data quality; however, given this group's susceptibility to COVID-19, that limitation would have been questionable on substantive grounds. Overall, the survey's two-pronged sampling procedure provided less biased data than either of its two component approaches could have done on a stand-alone basis. With cash expenses of about 1€ per completed questionnaire (1.33€ for subsample A & 0.52€ for subsample B), ESPACOV's cost-benefit profile is probably beyond reproach.

6. References

Bosnjak, M./ Neubarth, W./ Couper, M. P./ Bandilla, W./ Kaczmirek, L. (2008). "Prenotification in Web-Based Access Panel Surveys: The Influence of Mobile Text Messaging Versus E-Mail on Response Rates and Sample Composition". *Social Science Computer Review*, *26*(2), pp. 213-223.

de Leeuw, E. D. (2018). "Mixed-Mode: Past, Present, and Future". *Survey Research Methods*, *12*(2), pp. 75-89.

Groves, R. M./ Heeringa, S. G. (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of the Royal Statistical Society: Series A*, *169*(3), pp.439–457.

Mavletova, A./ Couper, M.P. (2014). "Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-mail Invitations". *Journal of Survey Statistics and Methodology*, *2*(4), pp. 498-518.

Serrano-del-Rosal, R./ Biedma-Velázquez, L./ Domínguez, J.-A./ García-Rodríguez, I./ Lafuente, R./ Sotomayor, R./ Trujillo, M./ Rinken, S. (2020), "Estudio Social sobre la Pandemia del COVID-19 (ESPACOV)", data file (v1), 2020. DIGITAL.CSIC (http://dx.doi.org/10.20350/digitalCSIC/12517).

Stern, M./ Bilgen, I./ McClain, C./ Hunscher, B. (2017). "Effective Sampling from Social Media Sites and Search Engines for Web Surveys: Demographic and Data Quality Differences in Surveys of Google and Facebook Users". *Social Science Computer Review, 35*(6), pp. 713-732.