Conducting Surveys and Interventions Entirely Online: A Virtual Lab Practitioner's Manual

December, 2020

Nandan Rao (UAB, BGSE)
Dante Donati (UPF, BGSE, IPEG)
Victor Orozco (DIME, The World Bank)[1]

THE WORLD BANK IBRD • IDA | WORLD BANK GROUP Development Economics

i2i DIME

upf. Universitat Pompeu Fabra Barcelona

UAB Universitat Autònoma de Barcelona

[1] This research is part of the entertainment-education program of the World Bank's Development Impact Evaluation Department (DIME). The views expressed herein are those of the authors and do not necessarily reflect the views of the World Bank.

Executive Summary

Online and social media campaigns reach billions of people every day. While commercial companies have built up extensive expertise in using these tools to recruit and build relationships with customers, researchers and policymakers have been slower to take full advantage of these new digital tools. The growth of mobile access in developing countries is opening up many opportunities to use social media tools to pursue development objectives: from surveys and impact evaluations to digital delivery of behavior-change campaigns.

Virtual Lab is an open-source set of tools developed by the World Bank's Development Impact Department (DIME) for online recruitment, intervention, and surveying via digital advertising and social-media platforms. It was built specifically for researchers and policymakers. It can be used to perform impact evaluations, inform the targeting of large-scale digital ad campaigns, collect and visualize cross-sectional survey data in real time, or build high-frequency longitudinal panel surveys.

An online study can only work if it performs inference on groups that can at least be found online, even if they are underrepresented, and only if such inference can be performed entirely from those individuals who are found online, even if they are hard to find. If this is the case, then collecting data via digital advertising and social-media platforms can provide a set of advantages:

1. Costs are naturally low and all solutions are trivially scalable, even beyond country borders.

2. Data is available in real time, for policymaking or to adjust the study design.

3. Recruiting a sample that is representative across a fixed set of measurable variables can be performed efficiently with targeted advertising.

4. If the intervention of interest is itself delivered online (e.g., an online advertising campaign), then performing an impact evaluation on the same platform can be extremely efficient.

5. Targeted advertising allows the recruitment of entirely bespoke and potentially hard-to-reach subpopulations that may be of interest to the policymakers or researchers.

This manual provides an introduction to the survey theory underpinning Virtual Lab, lays out detailed guidelines for designing online studies within the platform, shares costs data and lessons learned from recent trials, and lays out promising areas for future research.

Virtual Lab can be self-hosted on private or public cloud infrastructure. All code is open source. Further details about this platform, including the code to independently run it, can be found at https://vlab.digital/.

I. Introduction

Online and social media campaigns reach billions of people every day. They do so through global ad networks, such as those run by Google and Facebook, that provide increasingly efficient tools to target and engage a significant portion of internet users. While commercial companies have built up extensive expertise in using these tools to recruit and build relationships with customers, researchers are just now starting to use these new digital tools (Grow et al. 2020).

With smartphone ownership and internet access rates rapidly growing in developing countries, the reach of global ad networks is widespread in these countries. In India, for example, more than 35% of the population regularly accessed the internet through a smartphone and 70 percent of young people between 18 and 34 years old used Facebook in 2018 (Statista 2020). This opens up many opportunities to use social media tools to pursue development objectives: from surveys and impact evaluations to digital delivery of behavior-change campaigns and development programs (e.g., training of frontline workers).

Developed by the entertainment-education program of the World Bank's Development Impact Department (DIME), Virtual Lab is an open-source set of tools for online recruitment, intervention, and surveying via digital advertising and social-media platforms. It was built specifically for researchers and policymakers. It can be used to perform impact evaluations, inform the targeting of large-scale digital ad campaigns, collect and visualize cross-sectional survey data in real time, or build high-frequency longitudinal panel surveys.

It should be noted than an online study can only work if it performs inference on groups that can at least be found online, even if they are underrepresented, and only if such inference can be performed entirely from those individuals who are found online, even if they are hard to find[2]. Gathering survey data, whether as part of an experiment or opinion polling, necessarily involves two distinct steps:

1. Recruiting respondents.

2. Asking questions.

Virtual Lab is an integrated platform for performing both steps. Recruiting respondents is performed via targeted digital advertising. The novelty of the platform, and the idea more generally, rests in using the segmentation capabilities and ad-placement APIs[3] of advertising platforms to stratify and optimize recruitment for statistically efficient analysis. By integrating survey responses into the recruiting platform, ad placement can be optimized based on the answers coming in from respondents in real time. For example, if a particular subpopulation (stratum) is underrepresented or has higher attrition or nonresponse then the budget for ads

[2] Collecting data online has one natural and strong disadvantage compared to traditional methods: not everyone is online. If a study must include individuals who do not have access to the internet, or must perform inference on groups entirely absent from the internet, then it cannot be performed online.

[3] An API, or application programming interface, is a feature that allows one software application to exchange data with or control another software application, in this case over the internet.

targeting that subpopulation can be automatically increased to ensure proper representation from that subgroup in the final sample.

While most survey platforms could theoretically be used to ask questions, Virtual Lab provides its own chatbot engine designed especially for longitudinal panel studies and information interventions via messenger apps (currently Facebook Messenger). The chatbot integrates payment flows to send incentives to study participants and includes a number of methods for delivering media-based interventions, such as via videos or links to external websites. Chatbot surveys have a couple of natural advantages to other online surveys: (1) respondents already know and use the messenger applications through which the questions are asked and (2) follow up questions are sent in the same chat thread, with users receiving push notifications on their phone, reducing the friction for longitudinal studies.

The primary purpose of this researcher's manual is to introduce potential users to Virtual Lab and help them design studies that take advantage of it. We hope, however, that the ideas included here, and the design of this platform more generally, are useful to the research community at large when considering digital study designs and can inform the future development of similar tools.

Section II of this manual makes two motivational arguments for online recruitment of respondents. First, that a non-representative sampling frame combined with targeted digital advertising has the potential to be as accurate and efficient as traditional "representative" sampling techniques. In particular, this potential is especially strong whenever a large portion of study invitees refuse to participate or drop out early. Second, that integrating recruitment and data collection in a single platform can open up new and significant opportunities for improved estimation efficiency by adjusting recruitment dynamically based on survey responses. For example, one could learn which variables should be used to stratify at the same time as one is collecting the data.

Section III of this manual discusses chatbot surveying. This is a valuable survey mode in a world of mobile phones with unique strengths compared to traditional web-based surveys. It is especially adapted to sending simple and frequent follow up messages that can be used for constructing panel surveys or experience sampling with repeated observations.

The rest of the manual is devoted to helping researchers build a study online, using Virtual Lab. We suggest some study designs that are a natural fit for the platform (Section IV) and then provide a step-by-step guide for designing a study (Section V). Following that, we share some results, costs, and numbers from previous studies along with a set of mistakes made and lessons learned (Section VI). Finally, we lay out (plentiful) avenues for future research in this space (Section VII).

Open source code ensures full transparency and auditing of security practices. All code is open source and can be self-hosted on any public or private cloud. All public network communication is encrypted by default with TLS[4] and data at rest can be encrypted at the block storage layer provided by the cloud infrastructure. For inquiries and further details about this open source platform, including the code to independently run it, please visit https://vlab.digital/.

[4] The Transport Layer Security protocol provides the encrypted communication that secures all websites accessed via HTTPS. This encryption ensures that data moving over the internet cannot be read, even if it is intercepted.

II. Reference Studies

Throughout this manual, we will refer to three initial studies that the research team has completed with Virtual Lab. While these studies do not take advantage of all the possibilities or use-cases we lay out in this manual, they will provide some concrete examples of some features. The studies will be referred to as “PFI,” “ITALY,” and “MNM” respectively in the rest of this manual. All Studies will be publicly available by early 2021.

PFI - Using Social Media to Change Gender Norms

A collaboration with the Population Foundation of India (PFI), this study is an online randomized control trial (RCT) testing whether two low-dosage 25-minute edutainment web series delivered through Facebook Messenger are effective at changing gender norms and promoting positive behaviors towards violence against women (VAW) in northern India. The team recruited 18-to-24-year-old youths living in New Delhi and six other cities and randomly assigned them to the two treatment conditions: an entertainment drama web series and a documentary web series.

Recruitment was performed on Facebook with a one-week campaign stratified by gender. Approximate gender balance between men and women was achieved at recruitment costs three-times higher for women. As participation incentives, individuals that filled in the baseline and at least one follow up survey were eligible for a raffle to win Samsung Galaxy smartphones or a “selfie” picture with a Bollywood celebrity. Clicking on the ad directed respondents to Facebook Messenger where the survey was administered by the Virtual Lab chatbot. After filling a baseline survey on demographic characteristics and knowledge, attitudes, and beliefs regarding gender norms and violence against women, respondents were sent a series of short videos directly within the chat. Subsequent videos were sent two hours after respondents had watched the previous. Follow up surveys were sent one week after finishing all the videos and again three months after finishing the videos. In follow up surveys, in addition to questions to gauge attitudes, knowledge, and beliefs about gender norms and gender-based violence, users received two calls-to-action: 1) an invitation to visit the websites of various NGOs working on issues relating to gender and 2) an invitation to add a “frame” to their facebook profile with the phrase “end violence against women.”

Both the videos and the links used Virtual Lab’s video-hosting and link-sharing integrations which collect data on A) if the users clicked the link and B) if the users clicked “play” to actually watch the video. This allowed researchers to track compliance and estimate the treatment effect on the treated.

ITALY - Stereotypes and Political Attitudes in the Age of Coronavirus

This project consists of a panel survey conducted in Italy to study how beliefs about the infectiousness of people from different countries, political attitudes towards China and the EU evolve in response to local exposure to COVID-19. Measures of beliefs and attitudes were collected over six waves during the spread of the pandemic, beginning in late February 2020.

Respondents were recruited via Facebook ads targeted to a set of regions in middle and southern Italy which, initially, had little-to-no exposure to the virus. Ads were not stratified and all recruiting finished within a week. Respondents self-reported the postal codes of their homes, which was combined with separate, provincial-level data on the spread of the virus. This was compared to self-reported knowledge about cases in their community.

In each wave, respondents were asked to estimate the spread of the disease among different national groups ("consider all the people in the world of Chinese nationality, according to you, how many out of 1 million are infected by Coronavirus?"). While individual estimates naturally varied widely, the research team was able to compare the repeated estimates from the same individual over time to track how estimates for different nationalities changed throughout the evolution of the global pandemic.

The team also conducted an information experiment that exposed a random subset of respondents to information about cooperation between EU countries in response to the COVID-19 virus. A set of images, with text containing facts about support Italy had received from specific EU countries during the pandemic, were sent within the Virtual Lab chatbot to two treatment groups. One group received information about Germany and Austria while another about France and the Czech Republic. The control group received no images. Two months after the intervention, respondents were asked their opinion on the solidarity that those specific countries had shown to other members of the EU during the last 20 years.

MNM - Targeting and Evaluating Malaria Campaigns Using Social Media

This study is a cluster RCT in India measuring community-level impacts of a social media campaign on malaria incidence, bednet usage, and treatment-seeking behavior. 80 districts in 3 north-Indian states were randomly assigned to treatment and control conditions. Treatment districts received three months of a Facebook ad campaign designed to raise awareness of malaria as well as improve both preventative and treatment-seeking behaviors. The campaign was specifically designed for engagement and social sharing of Facebook posts. Facebook users living in control districts were excluded from the campaign's ads.

The social media campaign was designed and run by Malaria No More, an international NGO aiming to eliminate deaths by malaria, in association with Upswell, a consulting firm specializing in running social media campaigns for social good. The campaign, as well as the impact evaluation, was funded by Facebook's Campaigns for a Healthier World initiative.

Virtual Lab was used to recruit a panel of respondents and interview them via chatbot on Facebook Messenger. Respondents were sent messages every two weeks where they were asked to report incidences of fever or malaria as well as report on their daily behavior regarding the usage of bednets or other preventative tools. They were incentivized with two tranches of mobile credit top ups: one third of the total credit was sent after the third follow up and the rest after the endline survey. The mobile credit delivery was handled within the chatbot by Virtual Lab's Reloadly payment integration.

A panel of survey respondents was recruited via Facebook ads. An initial sample of 250 respondents per cluster was recruited, where clusters were defined by administrative "districts." Clusters varied widely in per-respondent acquisition cost. Virtual Lab was used to generate separate ad sets for each cluster and continuously optimize the daily ad spend to efficiently recruit respondents from all clusters (lowering spend in "cheap" clusters made them even cheaper, increasing spending in "expensive clusters" ensured quicker recruitment of those respondents).

Recruitment was also adaptively optimized by Virtual Lab to increase the representation of individuals living in kutcha (mud, tin, and/or straw) dwellings. These respondents reported a higher incidence of malaria than other groups and were additionally under-represented when following a naive recruitment strategy. With optimization, the number of districts with more than 20% of respondents living in kutcha dwellings increased from 29 to 65 while the number with less than 10% decreased from 13 to 0. Cost-per-respondent increased under this optimization, leading to an estimated ~20% increase in total recruitment costs compared to the naive recruitment strategy.

III. Recruiting with Online Ads

Non-probability Sampling and Poststratification

The traditional approach of recruiting survey participants starts with probability sampling: the entire population is in the sampling frame and N individuals are selected with known probabilities. This can be contrasted with non-probability sampling, where the selection probability of individuals is unknown to the researcher. Recruiting via digital advertising, as done by the Virtual Lab platform, is a non-probability sampling technique because you cannot calculate the probability of being shown an ad and surveyed for everyone in your population.

Even in probability sampling, and even if the initial frame is perfect and selection probabilities are known, non-random nonresponse implies that survey responses are not themselves representative and thus not unbiased out-of-the-box. For this reason it can be helpful to decompose survey error into (among others) frame error and nonresponse error[5]. Both, however, create the same problem: your sample is biased and responses must be reweighted in order to create an unbiased estimate of a population parameter (Kolenikov 2016).

Reweighting methods can be separated into three categories:

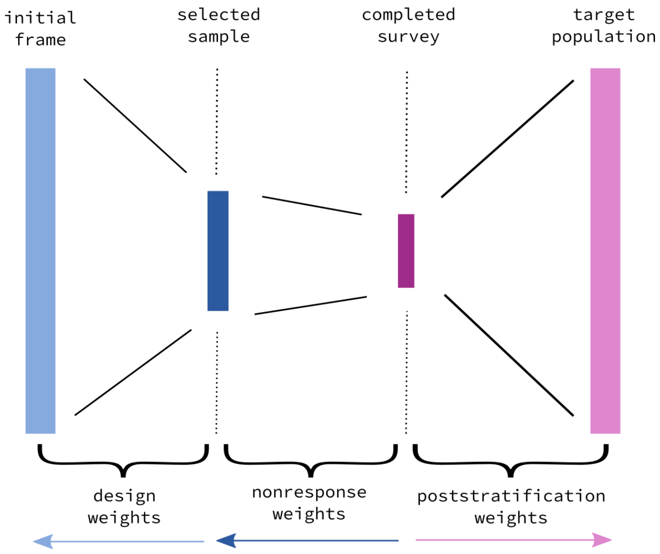
1. Design weights. If the selection probability of each individual in the sampling frame is known, responses can be weighted by the inverse probability of selection. No external data or modeling assumptions are needed.

2. Nonresponse weights. An estimated probability of nonresponse can be calculated within the survey, given the observed variables recorded in the survey. No external data is required, but the researcher must make modeling assumptions to estimate the nonresponse probability as a function of observable covariates available in the initial sampling frame.

3. Poststratification weights. If external data on the population is available (e.g., census data), response weights for a subpopulation can simply be set to reflect their total

[5] This has been discussed heavily in the Total Survey Error framework. See Biemer (2010).

proportion of the population. Techniques such as raking[6] can be used if only marginal distributions are available in the population (Deville and Särndal 1992; Battaglia, Hoaglin, and Frankel 2009) and techniques such as multilevel regression and poststratification (MRP)[7] can be used if the number of variables, and hence subpopulations, is large (Gelman and Little 1997). These techniques do not require known selection probabilities, only non-zero representation of all subpopulations of interest in the final survey results (and by extension, in the original sampling frame). They do require modeling assumptions over the observable covariates available both in the survey data and in the external data.



Non-probability Sampling and Poststratification. *Design weights and nonresponse weights can together be used to recover an estimate of a parameter in the population represented by the initial sampling frame. Alternatively, poststratification weights can be used to build an estimate of a parameter in the target population. If the target population is different from that of the initial sampling frame, or design weights not available (as in nonprobability sampling), then using some form of poststratification weighting is needed to estimate a population parameter from sample data.*

[6] Raking, also known as iterative proportional fitting, consists of iteratively looping over each variable and adjusting the weights of individual observations such that the weighted marginal probability of the variable in the sample approximates that of the population. The process goes like this: given a set of variables, start with one of them and reweight your observations so that the weighted distribution of your sample matches the marginal distribution of this variable in your target population. Repeat for the next variable, and then the next, all the way down the line for all your variables. By the time you adjust for the last variable, your first one is likely off again, so you iterate the process until convergence. While the technique provides no theoretical guarantee on the closeness of the joint density, it has been shown to work well in practice.

[7] Multilevel regression and poststratification consists of using a Bayesian hierarchical model which has the advantage of being able to estimate cell-specific parameters for populations stratified on many variables, even when the sample size in individual cells is small or non-existent. For example, consider a survey in the US where you would like to include the variables race, gender, and state. You might not have many observations of each combination of race/gender within each state (or even just many observations within each state period). However, if you model states as belonging to higher-level groups (regions), you can estimate the region-level parameters and then regularize the state-level parameters to the corresponding regional-level parameters. This allows the individual state-level parameters to deviate from the region only when there is enough state-level data to justify the deviation. Given estimates for every cell in your stratification, poststratification weighting consists in directly applying the target population weights for the cell.

Design weights together with nonresponse weights are sufficient to recover an estimate of the parameter -of interest in the population that makes up the initial sampling frame. Alternatively, poststratification weights with external data are sufficient to estimate a parameter of interest in the population represented by the external data. If the initial sampling frame is the same as the "external data" about your target population (i.e. both are an official government census), then the two techniques are identical.

In practice, however, many surveys start with a sampling frame that is relatively representative but not very covariate-rich. This is the case with random-digit-dialing (RDD), the most common technique used by companies and organizations conducting phone surveys and a staple of opinion polling for many years. Nonresponse rates for RDD surveys have been steadily rising for decades, reaching 91% in the 2010s (Keeter et al. 2017; Shirani-Mehr et al. 2018). Because phone numbers do not come with demographic variables that allow for the creation of a nonresponse model, it is standard practice for these surveys to employ poststratification weighting to estimate their population parameter (Gelman and Little 1997).

What, then, is the value of a representative initial frame if nonresponse is so significant and poststratification techniques needed anyways? Can nonprobability sampling techniques with poststratification replace traditional probability-based sampling measures? If so, in what settings is this feasible and advisable?

These questions have been posed by a number of researchers of late. Wang et al. (2015) applied multilevel regression and poststratification (MRP) to data from an opt-in poll that was made available on the XBox gaming platform to estimate vote share among the two primary candidates for the 2012 US presidential elections. They compare their predictions with polling averages from pollster.com and find that their predictions track the polling averages very closely and indeed even produce better results than the polls in the days running up to the election. Goel, Obeng, and Rothschild (2015) applied MRP to survey data from a sample of 1000 Amazon Mechanical Turk (AMT) workers and from 1,000 respondents collected via online survey company Pollfish to calculate mean outcomes for a variety of questions from the General Social Survey (GSS) and similar questions from surveys performed by Pew Research. They report that their calculated outcomes from AMT and Pollfish respondents have a mean-absolute deviation (MAD) from GSS/Pew benchmarks of 7.2 and 7.4 percentage points respectively, where the GSS and Pew differ from each other with a MAD of 8.6. Further research from Pew shows a similar error magnitude between online opt-in surveys with poststratification methods and their traditional phone-based survey results, reporting average deviations of 6 percentage points (Mercer, Lau, and Kennedy 2018).

What do these results imply? The GSS is a rigorous survey in which considerable time and resources are poured into creating an inclusive sampling frame and then getting responses (in-person) from each sampled individual (minimizing nonresponse), coming in at the considerable price tag of $3 per respondent per question (Goel, Obeng, and Rothschild 2015). While there is no ground truth in the opinion questions measured in these surveys, the GSS is widely considered the best we have. Phone-based surveys from firms like Pew Research start with a significantly more representative sampling frame (all phone numbers) than the convenience samples studied (MTurkers or opt-in visitors to a set of websites/apps). Despite this advantage, it does not systematically outperform the convenience sample. This would seem to imply that either A) the magnitude of nonresponse error in those techniques overwhelms any improvement in frame error and/or B) that poststratification techniques successfully made up for a significant portion of frame error, reducing the importance of the initial frame in the resulting total survey error.

Critically, however, poststratification techniques (like nonresponse modeling) require the researcher to select a set of relevant variables on which to stratify. This need to model the outcome as a function of covariates is happily absent in a pure probability-sampling method with minimal nonresponse.

A simple example can help illustrate this point: imagine you are interested in surveying your city's population to know how concerned they are about a particular global pandemic. Consider, additionally, that you collect information on respondents' race and gender, but do not think to collect information on their age. Assume that in reality, older people are more vulnerable to this disease and thus more concerned.

If you have a comprehensive sampling frame and a crack team of door-knockers (think GSS), your resulting sample (given sufficient size) will likely be representative for age and thus your lack of information about respondents' age will not affect the estimate of the average level of concern in your city. Now consider you are running an RDD-based phone survey and it turns out that young people are much less likely to take your call. You did not collect information on age (you did not know it would matter!), so poststratification cannot make up for this nonresponse bias and you end up woefully overestimating the concern in your city. Conversely, consider you run an internet-based convenience survey in which young people are overrepresented in your sampling frame. Without information on the age of each respondent in your collected data, poststratification cannot overcome this bias.

While the importance of "age" might be obvious in modeling citizens' concern regarding a pandemic, the inclusion or exclusion of other variables might not be so obvious a priori (e.g., political affiliation or social media diet). Thus, any technique relying on poststratification implicitly brings a modeling and, in particular, a variable-selection problem. Indeed, Mercer, Lau, and Kennedy (2018) show in a comparison of poststratification techniques that what matters most for removing the bias of nonprobability samples is not the exact technique used, but rather the variables chosen. This should be a strong cautionary tale to anyone who thinks they have a representative sample but has taken the variables required for representation as given and not chosen them based on the specific outcome. In the following section we will discuss how integrating recruitment and survey responses can allow for the variable selection problem to be formally solved in an online fashion, potentially leading to significant increases in outcome estimation accuracy.

An additional disadvantage of a non-inclusive sampling frame is that, while poststratification weighting can make up for certain populations being underrepresented, it can do nothing for populations entirely absent from the sampling frame. Digital advertising, along with the other online convenience methods compared in the studies reported above, by definition exclude from their initial sampling frame everyone without internet access. This could potentially exclude very poor households entirely, as well as those of certain religious or ethical beliefs. This is a deficit that no amount of poststratification weighting can make up for.

Many new studies have begun to use Facebook specifically as a recruitment tool and apply poststratification to estimate population quantities (Zagheni, Weber, and Gummadi 2017; Perrotta et al. 2020). We are not aware, however, of any study that has systematically compared recruitment via online advertising to traditional probability-based recruitment such as RDD.

We see several advantages that online advertising has as a sampling frame as opposed to other online convenience sampling methods:

1. Large population coverage: Facebook registers 2.7 billion active users (Statista 2020). Google Ads Display Network reports to cover 90% of internet users worldwide across millions of websites (Google 2020).

2. Targeted advertising allows researchers to intentionally pay more to reach under-represented groups. This allows researchers to trade off cost and representativeness in a way that the studied convenience methods (i.e. Xbox live players) do not allow. For example, in the PFI study, gender balance between men and women was achieved by stratifying ads and paying 3 times the cost for women compared to men. Similarly, in the MNM study, balance across regions and dwellings was achieved by paying up to 50 times more for the most expensive strata compared to the cheapest.

3. Real time communication. Digital advertisers expose APIs, which allows software run by the researcher to communicate with the advertising platform in real time and adjust ad placement. While on the one hand this makes it convenient to create hundreds of audiences for the stratified recruitment, the potential extends further. We explore this feature more in the following section, but it's worth highlighting that it is novel, not present in traditional sampling frames or considered in any of the research on this topic that we are aware of.

The combination of the last two points (targets ads controlled via API) is very powerful: it implies that we can build our own ad-optimization engine to optimize the goals of researchers, policymakers, and the public good. This does not come out-of-the-box from ad platforms, whose built-in ad optimization routines are designed to maximize value under the assumption that diversity of audience (customers) is not in-and-of-itself valuable. While this may be the case in retail, it is not the case for research. The value (information gain) of an individual decreases with the number of similar individuals we already have. This is why Virtual Lab has its own ad optimization engine that uses the available tools to optimize for heterogeneity rather than homogeneity.

These additional advantages of digital advertising over the convenience-based methods studied give solid reasons to believe that results could potentially be even better for digital advertising. Future research is needed to test that hypothesis.

Integrating Recruitment and Surveying

Virtual Lab is a platform for recruitment and surveying via digital advertising. As discussed above, this technique presents new opportunities but also challenges that are worth discussion and research. Many important possibilities are opened up by integrating digital recruiting (with all the micro-targeting and dynamic-optimization power of modern digital advertising) with survey answers in a single platform. In particular, the question becomes: what can we gain if we have the ability to adjust recruitment in real time based on initial survey responses that come in?

We will consider several increasingly complex (and realistic) scenarios and see how, in each case, a platform which integrates recruitment with surveying (and can adaptively adjust recruitment) can achieve more efficient results than a traditional process.

Scenario A: Adjusting for attrition

Consider you have a set of simple, demographic variables on which to stratify your population and a target number of respondents for each subpopulation. For example, consider that you want responses from 500 respondents aged 65 and older and 500 aged 65 and younger. How many should you recruit from each group? If attrition rates are not equal among your subpopulations, and not known ahead-of-time, this question can be hard to answer.

If, however, your recruitment platform knows who finishes your survey, it can continue recruiting from each subpopulation until it gets exactly the numbers you want. Virtual Lab's recruitment engine, for example, can be set up to continue recruiting from each subgroup until you have 500 who finish your survey (however you define "finish").

Scenario B: Adjusting for hard-to-find subgroups

Consider you have a set of variables on which you want to stratify your population, but they are not restricted to traditional demographic variables and instead include variables that you ask in your survey. For example, in MNM, we wanted a sample stratified by the respondent's dwelling type (cement dwelling vs. non-cement dwelling). This is not a demographic variable available for sampling from any traditional sampling frame nor was it a variable available for targeting in Facebook's ad platform.

Traditionally, the only way to stratify on these variables would be to "over-recruit" and hope that none of your subpopulations of interest are overly rare. When you run the analysis, you can either lament that your estimates are a bit unstable due to an unlucky low number or lament that you paid more than you needed to estimate your effect. In an integrated recruitment/survey environment, however, recruitment can continue until exactly the point where individual subpopulation targets are fulfilled, even if the subpopulations are defined by survey variables such as "dwelling type."

Taking it one step further, Virtual Lab uses advanced targeting techniques of digital advertising platforms to target ads to groups of users even if they are not defined by traditional demographic variables available for explicit targeting via the platform. Specifically, ad platforms provide two features that allow this platform to optimize for populations that can't be explicitly defined by demographics: A) custom optimization events and B) "lookalike audiences" (also called "similar audiences"). In both cases, Virtual Lab sends the ad platform a list of users who, after answering the survey, are revealed to belong to a certain subpopulation and tells the ad platform to target ads at "people like this." The ad platform then uses all the (massive) set of private variables at their disposal to determine how to find similar people and target ads to them.

In the MNM study, for example, Facebook didn't know for certain we were interested in people living in kutcha dwellings, nor could it even know for certain if people lived in kutcha dwellings, but by providing it with a continuously growing list of users that fit that criterion, we were able to continuously reduce the cost-to-acquire for this subgroup throughout the recruitment process.

Scenario C: Adjusting for outcome variance

Consider now that you have a set of variables on which to stratify your population but do not know how many finished surveys you need from each subpopulation. If you're interested in the expectation of a population value (e.g., average household income), the optimal sample size to allocate to each subpopulation can be described by the Neyman Allocation (Neyman 1934; Groves et al. 2010):

$$n_h = n\frac{W_h S_h}{\sum_h W_h S_h}$$

Where two factors influence how many individuals you sample from each subpopulation: n_h/n, the share of your population that belongs to that subpopulation and S_h, the variance of their outcome. For example, if the outcome is average household income and the strata are defined by profession, you may only need to sample a few "students" while you may need to sample many "managers." In the general case, however, this variance may not be known ahead-of-time.

Once again, if you have a platform with dynamic recruitment, you can adjust the sample size per subpopulation in an adaptive fashion, based on sample estimates of the variance, spending more to recruit respondents from subpopulations with high outcome variance and spending less on more homogenous subpopulations.

Scenario D: Adjusting the subgroups themselves

Consider now that you do not actually know the correct set of variables by which to stratify your population. In general, these should be variables with strong associations to the outcome, but that might not be known a priori. Without integrating recruitment and surveying, the best variables for stratification can only be determined a posteriori and even then only with a bit of luck: to consider a new variable for stratification you need to have strong representation across all possible values of that variable.

Consider the example of election polling. Despite the industry being well-funded and the techniques well-polished, new variables are constantly being added at the end of every election cycle and the lack of new variables constantly being blamed for previous prediction failures and polling biases. If the process is integrated, however, both the outcome values and the ideal feature representation can be learned simultaneously in one adaptive survey. We believe this is an exciting area of research for both statistical theory and survey platforms to explore.

IV. Asking questions via Chatbot

There are many ways in which one can deliver a survey: face-to-face, phone calls, web forms (Qualtrics), etc. "Chat" is a relatively new form of surveying. Respondents receive questions as text messages in a messaging application on their smartphone and respond by writing their responses as text messages in return.

By using the chatbot capabilities provided by many popular messaging apps (Whatsapp, Facebook Messenger, Telegram, Viber, etc.), one can deliver survey questions and collect responses within the apps that many people already use to communicate every day.

Very little research exists showing the mode effects of chat (also known as "messenger" or "chatbot" survey designs) as a format for questionnaires and this is an important question for further research. Toepoel et al. (2020) do present, however, some initial results comparing the two modes. They show that there are indeed differences, including that users provide shorter answers to open-ended questions when provided with a chat design as opposed to a web survey design. There are, however, some key limitations to their study, for example that the majority of their users were on desktop and the messenger application was a custom-built interface that users had never used before.

However, despite the fact that it is a new survey mode, there are key advantages of chat that lead us to believe it can be a powerful data-gathering tool:

1. The entire process is asynchronous. Unlike all other methods, users do not stop everything for a contiguous block of uninterrupted time to fill the survey. Instead, they respond to the next question at their own convenience. This could, of course, be considered a disadvantage as well if you don't want your respondents getting distracted by their daily lives in the middle of answering your questionnaires.

2. The interface is familiar. Web surveys inherently require respondents to learn a new digital interface (how do I move to the next question, how do I enter my response, etc.). In a chat survey, on the other hand, respondents use an interface they are already familiar with.

3. Seamless follow ups. Sending a follow up, whether it's 1 question or 20, after two hours or two months, is experienced by the respondent as "just another" message in a thread which can be read and responded to with minimal friction. Web surveys, on the other hand, have no natural way to follow up with respondents and must resort to external tools such as email. In the ITALY study example, attrition over 6 distinct waves and 3 months was only 50%, with more than 90% of respondents voluntarily completing each wave from a single push notification, without any additional reminders or encouragement.

4. Notifications. Sending messages can be used to "push" notifications to respondents based on external events. For example, when incentive payments have been processed, respondents can be notified directly in the chat ("you've been paid!" or "your payment could not be processed, please provide a different number"). In MNM, the implementation of this payment processing inside the bot dropped the failed payment rate from 5% to 0.2%, by communicating errors directly to respondents and allowing them to provide alternative mobile numbers for top up payments.

5. Mobile first. Web surveys are digital imitations of paper surveys. Modern survey software has re-boxed that imitation to work well on small screens like mobile devices. Text messaging, on the other hand, is a mobile idea. Chat surveys replicate a text-message conversation, not a paper survey, and thus provide a mobile-native experience.

Thus, while much empirical research is still needed to understand the effects of these differences, the affordances of the design itself allow us to draw some initial hypotheses as to when chat, as a survey mode, might be advantageous. Table 1 shows a side-by-side comparison of the two modes and their relative advantages across different features.

Table 1. Mode comparison

	Chat Survey	Web Survey
Follow ups	Seamlessly integrated. Can be many, potentially very frequent and with few questions.	Must use external process (email). Best for few follow ups, preferably with many questions at once.
Open-ended responses	Better for relatively short answers.	Short or long answers could both work.
Mobile vs. desktop	Great on mobile, desktop, and allows platform switching.	Favors desktop. Cannot switch platforms mid-survey.
Integrated payment	Yes (in Virtual Lab's chatbot)	Yes (depending on platform)
Users share media (photos, audio)	Seamless and natural	Possible
Users can share content with friends or contacts	Seamless and natural	Possible

V. Virtual Lab Study Archetypes

We will consider several study design archetypes/patterns that we think are particularly well-suited to Virtual Lab, along with examples of how they could be (or were) designed. These should be considered as rough sketches of potential research designs meant to encourage brainstorming, rather than detailed blueprints to be followed directly.[8]

Individual-level randomized control trials

In this type of study, participants are randomized into treatment groups at the individual level and outcomes are measured via survey responses or recordable online actions (i.e. clicking links or agreeing to share content). The treatment itself could be administered via:

1. A digital ad campaign, where individuals in the study are "retargeted" in the same ad platform used for recruitment and shown ads. The ads are shown to the individuals "in the wild," anywhere in the ad network (i.e. in their Facebook feed or while browsing websites), and targeted directly to treatment-group respondents.

2. A within-survey intervention, where the treatment is delivered within the survey itself. For example, the survey chatbot might send respondents a video (or several!) to watch, as in the PFI study or might send users images, as in the ITALY study.

[8] Where appropriate, it is important to submit research protocols to institutional review boards for ethical approval, as well as collect informed consent from all research participants.

3. A real-world intervention. For example, respondents can provide their address and researchers can send bednets to their houses to encourage them to sleep under bed nets in order to prevent mosquito-borne diseases.

Clustered randomized control trials

In this type of study, participants are randomized at cluster-level, where clusters might, for example, consist of geographic regions or communities. Outcomes, again, are measured via survey responses or recordable online actions. The treatment is administered to the entire cluster and this can take many forms: for example, any advertising campaign targeted by region or region-specific government policies.

An example of this type of study is MNM. The treatment was an ad campaign meant to promote preventative behaviors and proper treatment and one of the primary outcomes of interest was malaria incidence. Because there are potentially large geographic spillovers from engaging in malaria-preventative behavior (malaria spreads from person to person) and high correlation of malaria-incidence within regions, it made sense to randomize at cluster level (in this case, the administrative "district"). Virtual Lab was used to generate control and treatment regions in the Facebook ad platform that could then be used by the advertising team to target their ads everywhere except for control districts.

Survey-response targeted ad campaign

Rather than (or in addition to) using Virtual Lab to perform impact evaluation, it can be used to directly improve the targeting of the main ad campaign itself.

In this pattern, a survey is run in parallel to the main ad campaign. The responses to the survey questions divide respondents into ideal audience and not-ideal audience categories. This information is returned to the ad platform which generates a model to predict new individuals who would likely form part of the ideal audience if asked the survey questions (this is done with the models traditionally used to predict "high-value customers" for retail advertisers). This prediction model is then used to target the main ad campaign at those more likely to be the ideal audience.

Consider as an example that you are running a vaccination campaign. Ideally, you only want to advertise to those who have not yet been vaccinated. A survey can be used to gather data on who has been vaccinated and the ad platform creates a model to predict those who have not yet been vaccinated. With that, you can target your ads specifically at those likely to not yet have been vaccinated!

This could be performed continuously so that the audience is updated in real time. This has the potential to drastically improve the cost-effectiveness of the campaign, as those who actually see the ads are more likely to be those who might actually benefit from it.

Quick population surveys with real time data visibility (dashboard)

The ability to instantly deploy ads, quickly recruit respondents, and immediately visualize response data in a dashboard can be extremely useful for any policymaker or researcher who needs to gather data on new outcomes of interest.

This can have advantages over going to a pre-existing pool of respondents in the following scenarios:

1. When the variables that define "representativeness," or subpopulations of interest, for the new outcome are different then those around which previous respondent pools were created. In this case, creating a new respondent pool, stratified around a different set of variables, can be done extremely quickly with digital advertising.

2. When the subpopulations of interest are unknown, as is likely the case if the outcome is new or of recent interest, an integrated recruitment/data-collection platform like Virtual Lab could theoretically jointly learn the important stratification covariates along with the collected outcome data.

Consider, for example, a policymaker wanting to understand the social impacts of new social-distancing measures to counter Covid-19. Populations of particular interest might be those who have pre-existing health conditions as well as those who are employed in heavily impacted sectors. Such a sample likely does not exist in any survey company's portfolio, but could be created with the targeting tools of digital advertising.

Population panels with high-frequency waves and real time data visibility (dashboard)

Building high-frequency data based on low-touch interactions is a perfect fit for Virtual Lab and the chatbot survey engine. Consider a panel of respondents that provide an answer to a repeated question on a weekly, or even daily, basis ("Do you have a fever?" or "How worried are you about ___?").

As an example, in MNM, respondents were asked several questions bi-weekly during the malaria season, including "did you or anyone in your family have malaria in the past two weeks?" With hundreds of responses each day, researchers were able to track in real time the evolution of the malaria self-reports across multiple geographic regions.

It should be noted that the choice to use a panel vs. a repeated one-off survey design will depend on the outcome of interest, but in both cases the Virtual Lab platform allows for quickly deploying a recruiting and monitoring solution.

VI. Steps for Designing a Virtual Lab Study

1. Select population parameter(s)

Digital advertising provides a lot of flexibility and power in sample recruitment. This means, however, that it is especially important to be explicit about the population and population parameter that you wish to estimate with the sample you recruit.

In individual-level experiments, your population may be the sample itself and the treatment effect estimated for the sample only. Many behavioral lab experiments are (for better or worse) designed this way. If your experiment seeks to test something that could be a policy, however, you might want to estimate your effect of the policy on the population that would be affected by the policy.

2. Decide how to measure parameters based on users' online interactions

Decide how to measure that parameter from a sample of individuals based on their interactions within a survey, chat application, or website. While the simplest outcome is an answer to a survey question, it can also be possible to collect data on users actions, however. While you cannot track everything a user does on their phone, you can send them to a website that you control, collect data on their actions there, and link those actions back to them. This opens up many possibilities for measured outcomes.

For example, you can direct them to a page with streaming videos and record how much they watched. You can also invite them to click on a link to an external website and record whether or not they clicked on it. Even if you can't directly record the actions, with a bit of leg work and creativity you can potentially still recover the variables. In PFI, for example, the survey chatbot asked users if they would like to add a "frame" to their Facebook profile pictures supporting putting an end to violence against women. If users answered "yes," they were sent a link to a specific frame. While the survey could record users' intentions (the "yes" answer), there was no automated way to check if they did, in the end, change their profile picture. However, the researchers could manually go through the profile pictures of the respondents and record who had added the frame.

3. Pick stratification variables

These can be explicit demographic variables available via the digital advertising platform or they can be variables you will collect yourself in your survey. In general, stratification variables should be discrete (or discretized) variables that exhibit strong dependence with your outcome and/or interaction with your treatment. Ideally, the strata defined by these variables can be used for both recruitment and analysis, allowing you to target recruitment to maximize statistical power.

In the case of a well defined population of interest, you should have measures for your stratification variables (or a subset of them) in the population. For example, if you're interested in population-level outcomes for a country's voting population, voter registration records with information on covariates such as age, gender, location, etc. might be available. Targeted digital advertising implies that many of the problems traditionally left to post-stratification techniques to solve can be worked on significantly through an adaptive recruitment process itself: spending extra advertising money on respondents in under-populated cells and saving money on respondents in would-be-overpopulated cells.

In addition to stratifying recruitment on poststratification variables, you might have some variables for which you do not have population data but still wish to use in your primary analysis, as would be the case when estimating heterogeneous treatment effects, for example. In that case, you would also want to stratify recruitment on those variables.

Geographic variables are easily forgotten in online experiments. If, however, you have any reason to believe your outcome is dependent on urban/rural divides or varies from major cities to smaller cities, then it is likely extremely important to consider stratifying on geography as well.

Create a desired target sample size for each stratum. Virtual Lab will use the marketing API of the digital advertising platform to generate separate ad sets for each individual stratum and optimize their spend and duration to recruit the target number of respondents. Keep in mind: greater stratification naturally implies higher advertising costs, so be sure to estimate your budget accordingly.

Consider your stratification variables in two groups: those which are available via digital advertising directly (demographic data such as age, gender, location, etc., call these "demographic variables") and those which you will collect in your questionnaire ("custom variables"). This platform can use variables from either group to stratify and recruit, however, with custom variables the targeted advertising will get more efficient the larger the sample size. More details on targeted advertising over custom variables are available in the documentation on the Virtual Lab website.

4. Panel? Cross-section? Repeated cross-section?

Digital advertising allows for continuous, reliable, automated recruitment. If you are not interested in panel data per se but want data over time, you can still get repeated "waves." Turning the recruitment process on and off is as easy as clicking a button, which greatly reduces the friction of repeating surveys. Repeated cross-sectional surveys will likely have less attrition and lower costs than the same data collected in panel format.

Chat, however, is extremely well suited to panel data as it allows for natural follow up questions. Unlike in traditional panel survey design, you have extreme flexibility over how you design the timing of your follow ups. Want to send a follow-up question four hours later? Four days? Four months? All are equally possible. This flexibility, combined with the fact that users can, at any time, put down your survey and pick it up again at their convenience later, requires you to think differently about the respondent experience and design your survey accordingly.

5. Select an incentive strategy

You may want to incentivize respondents to answer your questions or participate in your experiment. Two common schemes are A) respondents are entered into a lottery or B) respondents each receive a small individual reward.

In the case of individual rewards, Virtual Lab includes a set of payment-provider integrations and an integrated payment system that allows the chatbot to notify respondents when payments have been processed, inform them of any errors, and advance in the chat only after payment has successfully been processed. This greatly reduces (or eliminates!) the back-office work required to successfully deliver incentives to respondents in a large study. A current list of payment provider integrations can be found in the platform's online documentation on the official website.

Attrition is a major concern for any study, but especially so when the involvement is more demanding, such as in multiple-wave panel data or experiments with time-intensive interventions. Integrating individual rewards into the chat can be a great way to build trust with respondents and can potentially help to reduce attrition. Even a very small reward, successfully delivered early in the survey process, can help legitimize the process and build trust, reducing attrition. Large payments at the end of the study (i.e., after the final follow up) increase the incentive to complete the study, which can similarly reduce attrition.

Table 2 compares lottery vs individual reward incentive schemes in the context of digital advertising.

Table 2. Comparison of incentives

	Lottery	**Individual Reward**
Up-front (advertising) costs	Potentially high if the lottery is not that attractive.	Potentially low if incentive is attractive.
Incentive (reward) costs	Potentially low if lottery prize is cheap	Potentially high if incentive is expensive
Total Costs	Depends	Depends
Building Trust and Attrition	Hard to build trust. Attrition potentially high as a result.	Possible to build trust if the reward can partially be delivered early in the process. This can reduce attrition.
Costs increase/decrease linearly with respondents	Lottery cost is fixed. Advertising costs are paid at the beginning and costs don't change, regardless of how many respondents complete the survey/experiment.	Yes. You pay per respondent, which implies payment is proportional to the number of respondents who complete the survey/experiment.

6. Design and Test

Consider running several pilot surveys to improve respondent experience and reduce attrition. Virtual Lab's dashboards allow you to monitor how respondents finish each wave of your survey and see which questions they tend to get stuck on. Randomize survey length and intensity and see how that changes attrition and especially attrition conditional on potential covariates or outcomes of interest. Viewing, in a real-time dashboard, the behavior of your respondents allows you to iterate survey design quickly and efficiently.

VI. Lessons From Previous Virtual Lab Studies

Table 3 compares costs and attrition numbers for the three example studies cited in the manual. It's worth noting some major drivers of cost:

Stratification increases representation but costs more. This tradeoff highlights the danger of not stratifying, low costs of digital recruiting can be attractive but it comes at a cost!

Recruitment time: Short recruitment periods force higher per-day ad budgets and higher per-impression ad costs. Cost-per-acquisition does not generally stay constant at different levels of daily budget. As the budget increases, the cost-per-acquisition increases as well. This is a natural consequence of the auction format of digital advertising, ads are competing with other advertisers for a limited number of daily viewers.

Attrition in each wave during panel studies can be very high and this naturally increases advertising cost exponentially. Techniques to reduce attrition have been discussed in the previous section, but we do not yet have rigorous evidence to show what exactly works. ITALY experienced extremely low attrition compared to the other two example studies. This might be driven by the difference in the countries, but also by the fact that ITALY took place during a country-wide Covid-19 confinement and the subject of the survey was Covid-19. Potentially, respondents had both more time on their hands to fill surveys and were also interested in the subject matter.

Table 3. Stratification, Recruitment Time and Attrition

	PFI	**ITALY**	**MNM**
Clicks on ad	33,000	3,500	302,000
Cost per click (US$)	0.66	0.10	0.20
Incentives	Lottery ticket	Amazon voucher	Mobile credit
Stratification of recruiting campaign	Yes	No	Yes
Type of study	Longitudinal (RCT)	Longitudinal	Longitudinal (RCT)

Individuals' participation time	60 minutes	60 minutes	40 minutes
Length of study (base-to-endline)	4 months	3 months	6 months
Initial sample size	5,200	1,220	18,800
Final sample size (percentage of initial sample size)	620 (12%)	600 (50%)	Study In Progress

VII. Future Research

Recruiting respondents with digital ads and asking them questions via chatbot is a relatively new approach. Virtual Lab was built as a research tool to take advantage of these new technologies and test out new techniques. As such, there is much work to be done!

Methods to measure bias of ads and ad optimization

Digital ad platforms generally have a very large population onto which they can target ads. The sample of individuals that are both shown ads and respond to them, however, is not a random sample from that population. It is worth pointing out several factors (among many) driving that non-randomness: A) the bids of other companies for users, which is based on those users value to them B) the prediction engine of the ad platform which determines who is most likely to click on your ad and C) the ad creative (image and copy) itself, which greatly determines who actually clicks on the ad.

If those factors lead to a strongly non-random sample from the potential population, then understanding how they relate to your outcome or interact with your treatment is crucial to knowing whether they introduce a bias in-and-of-themselves. Luckily, all of those factors can be partially controlled by the researcher: A) bids can be increased to outbid other advertisers, B) daily budgets can be increased, which forces ad platforms to show your ad to more people, including those who may not be those they predict most likely to click and C) ad creative (image and text) can be swapped out and tested against each other.[9] Coming up with a reasonable methodology for doing these sensitivity checks is absolutely vital to doing research with respondents recruited via digital advertising.

Mode effects and Attrition of Digital Advertising

Attrition is a major concern in any survey process, but in messenger surveys and/or digitally-recruited participants little is known about specific mode effects on attrition. Our

[9] Note that, as opposed to traditional A/B testing done with creative by ad companies, here we are not suggesting that ads be compared in order to choose the one that is most effective, but rather to be compared in downstream analysis to see if the ads recruit qualitatively different audiences. For example, if the intervention shows a differential treatment effect on the audiences recruited by different ads or the estimated population parameter differs in the audiences recruited by the different ads.

experience with the example studies in India and Italy suggest that potentially these modes might be “easy in, easy out”: individuals opt-in to start a survey very easily and, potentially, without a lot of commitment, which might make them opt back out again with equal ease. Understanding how these digital modes of recruitment and interaction affect attrition is therefore an important area for further research.

Mode effects and Attrition of Chatbot Surveying

Every survey format has a “mode effect” that influences the way respondents answer questions. Some initial research shows that chat does, indeed, have mode effects that differ from those of a traditional web survey (Toepoel et al. 2020). More work is needed to fully understand these effects. It is of special interest to see how attrition across multi-wave surveys differs between chatbot surveying and other modes (web survey + email, for example).

Automatic Stratification

As mentioned in the motivation section on integrating recruitment and surveying, many opportunities arise in an integrated platform to automate the stratification process based on real-time outcome data: both for optimizing the number of respondents per stratum and for selecting the variables that make up the strata themselves.

Given the technology to create integrated platforms, like Virtual Lab, and the APIs of modern digital advertising tools, online optimal stratification methods can have important potential benefits in increasing the accuracy of any number of surveys or polls. We believe this is a very interesting area for statistical methodological research that addresses these problems and proposes solutions that work in this context.

Comparing accuracy of results to gold-standard surveys

Similar to previous results using post-stratification to estimate election or survey outcomes (Wang et al. 2015; Goel, Obeng, and Rothschild 2015), it will be important to show how well samples recruited via digital advertising can replicate gold-standard probability sampling results such as those by national census-based surveys (i.e the General Social Survey or the European Social Survey).

In the motivation section of this manual, we lay out a hypothesis that purposefully targeted subpopulations recruited via large digital advertising platforms can improve upon results from pure convenience samples such as XBox players or MTurk workers. If such an improvement exists and is substantial, targeted digital advertising could quickly gain prominence as a proven and affordable tool for accurate survey recruitment.

Bibliography

Battaglia, Michael P., David C. Hoaglin, and Martin R. Frankel. 2009. “Practical Considerations in Raking Survey Data.” Survey Practice 2 (5): 1–10. https://doi.org/10.29115/sp-2009-0019.

Biemer, Paul P. 2010. “Total survey error: Design, implementation, and evaluation.” Public Opinion Quarterly 74 (5): 817–48. https://doi.org/10.1093/poq/nfq058.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. “in Survey Sampling Calibration Estimators.” Journal of the American Statistical Association 87 (418): 376–82. http://www.jstor.org/stable/2290268.

Gelman, Andrew, and Thomas C. Little. 1997. “Poststratification Into Many Categories Using Hierarchical Logistic Regression.”

Goel, Sharad, Adam Obeng, and David Rothschild. 2015. “Non-Representative Surveys: Fast, Cheap, and Mostly Accurate,” 27.

Groves, Robert M., Eleanor Singer, James M. Lepkowski, Steven G. Heeringa, and Duane F. Alwin. 2010. Survey methodology. https://doi.org/10.4324/9780429314254-2.

Grow, André, Daniela Perrotta, Emanuele Del Fava, Jorge Cimentada, Francesco Rampazzo, Sofia Gil-Clavel, and Emilio Zagheni. 2020. “Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations.” https://doi.org/10.31235/osf.io/ez9pb.

Keeter, Scott, Nick Hatley, Courtney Kennedy, and Arnold Lau. 2017. “What Low Response Rates Mean for Telephone Surveys,” 1–39. http://www.pewresearch.org/wp-content/uploads/2017/05/RDD-Non-response-Full-Report.pdf.

Kolenikov, Stas J. 2016. “Post-stratification or a non-response adjustment?” Survey Practice 9 (3): 1–12. https://doi.org/10.29115/SP-2016-0014.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. 2018. “For Weighting Online Opt-In Samples, What Matters Most?” Pew Research Center.

Neyman, Jerzy. 1934. “On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection Author (s): Jerzy Neyman Source : Journal of the Royal Statistical Society , Vol . 97 , No . 4 (1934), pp . 558-625 Pub.” Journal of the Royal Statistical Society 97 (4): 558–625.

Perrotta, Daniela, André Grow, Francesco Rampazzo, Jorge Cimentada, Emanuele Del Fava, Sofia Gil-Clavel, and Emilio Zagheni. 2020. “Behaviors and attitudes in response to the COVID-19 pandemic: Insights from a cross-national Facebook survey,” 1–17. https://doi.org/10.1101/2020.05.09.20096388.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel, and Andrew Gelman. 2018. “Disentangling Bias and Variance in Election Polls.” Journal of the American Statistical Association 113 (522): 607–14. https://doi.org/10.1080/01621459.2018.1448823.

Toepoel, Vera, Peter Lugtig, Bella Struminskaya, Anne Elevelt, and Marieke Haan. 2020. “Adapting surveys to the modern world: Comparing a research messenger design to a regular responsive design for online surveys.” Survey Practice 13 (1): 1–10. https://doi.org/10.29115/sp-2020-0010.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. “Forecasting elections with non-representative polls.” International Journal of Forecasting 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants.” Population and Development Review 43 (4): 721–34. https://doi.org/10.1111/padr.12102.