Accounting for Complex Survey Designs: Strategies for Post-stratification and Weighting of Internet Surveys

Erin Hartman and Ines Levin*

October 4, 2019

Abstract

This chapter focuses on methods for analyzing data from Internet surveys with complex surveys designs in order to draw inferences that can be generalized to a target population of interest. We first review the central design issues and approaches for dealing with representativeness challenges that researchers commonly face when using online polling for persuasion research. Then, using data from a survey experiment on support for immigration reform, we demonstrate the importance of the careful choice of auxiliary information used when constructing weights for ensuring the generalizability of findings from non-representative Internet surveys.

1 Introduction

Survey research is the most popular method used by social scientists to measure the impact of persuasion attempts. Because of declining response rates of traditional survey modes (e.g. landline phone and mail surveys), high cost of alternative offline survey-administration techniques (e.g. cell phone surveys), and increased number of Internet users worldwide, practitioners are increasingly turning to web-based technologies for collecting public opinion data. Due to coverage and sampling limitations associated with Internet surveys, this technological change has not come without its challenges (Baker et al., 2016). In this chapter, we focus on methods for analyzing data from Internet samples in order to draw inferences (e.g. evaluate the success of persuasion efforts) that can be generalized to a target population of interest.

While the set of methods for recruiting and interviewing survey respondents is diverse and ever-expanding, the survey research process invariably involves relying on a sample of respondents to learn about the

*Hartman is Assistant Professor of Statistics and Political Science, University of California, Los Angeles; Levin is Associate Professor of Political Science, University of California, Irvine. This is a pre-publication version of a chapter published in the *Oxford Handbook of Electoral Persusasion*, edited by Elizabeth Suhay, Bernard Grofman, and Alexander Trechsel.

characteristics of a target population (e.g. whether a typical member of the American public is open to persuasion on a policy issue). Inferences about the population made from a sample may be more or less accurate depending on a number of factors, including whether the population the sample is drawn from covers the target population of interest; and whether the probability that a member of the population is selected for inclusion in the sample is known (or can be accurately estimated) by the researcher. These conditions are not easily met in the case of standard non-probability Internet surveys.

Internet surveys can be conducted at a relatively low cost and offer a number of advantages relative to offline survey modes. Since online questionnaires are self-administered, they are perceived as less intrusive by respondents, and this helps alleviate social-desirability and privacy concerns (Holbrook and Krosnick, 2010; Kreuter, Presser, and Tourangeau, 2008). Another advantage of Internet surveys is the availability of survey software (e.g. Qualtrics and SurveyMonkey) that can be used to easily create flexible survey instruments with complex features such as random variation of aspects of the questionnaire design. But since practitioners typically rely on opt-in Internet samples that are not representative of the target population of interest, non-probability Internet surveys often exhibit more bias than low-response rate probability phone samples (Dutwin and Buskirk, 2017).

Online surveys vary greatly in terms of methods used to recruit survey participants. Short surveys conducted over non-probability convenience samples (e.g. respondent recruitment via Amazon's Mechanical Turk) may contain scant information about participants, making it difficult to evaluate whether the sample resembles the population of interest in all relevant ways. Collecting more auxiliary information to use in adjustment methods can lessen the cost advantage of online surveys. While research on the generalizability of findings derived from convenience samples is far from conclusive (Berinsky, Huber, and Lenz, 2012; Coppock, 2018; Huff and Tingley, 2015), it has been noted that convenience samples could be particularly problematic in the context of survey experiments with heterogeneous treatment effects that vary over segments of the population (Coppock, Leeper, and Mullinix, 2018; Mullinix et al., 2015). If the probability of opting into the sample correlates with the source of treatment effect heterogeneity, a non-probability convenience sample may carry only minimal information about treatment effects among underrepresented groups and may result in inaccurate estimates of average treatment effects (Levin and Sinclair, 2018).

Polling companies have developed a variety of techniques for reducing bias in Internet samples. Knowledge Networks (now GfK) uses probability sampling implemented via list-assisted random digit dialing (RDD) to select adult members of the American public into a representative online panel. If a selected individual lacks Internet access, GfK provides free hardware and Internet service in exchange for frequent participation in Internet surveys. Other companies use different strategies for addressing the representativeness problem. YouGov, for example, has put together large online panels in a number of countries, with panel members recruited via a combination of web-based advertising and methods for recruiting hard-

to-reach individuals (including RDD). To circumvent the bias from non-probability sampling, YouGov developed a method denominated "sample matching" for identifying subsets of panelists resembling random samples drawn from the target population. Individuals that agree to join GfK's and YouGov's panels regularly receive email invitations to participate in online surveys and are offered incentives to encourage participation.

While techniques such as those developed by GfK and YouGov help address coverage and representativeness issues in online panels, they also greatly increase the cost of conducting Internet surveys. A more cost-effective alternative used by many researchers is to rely on non-probability, convenience samples drawn from online panels maintained by polling companies such as Qualtrics or SurveyMonkey, where panel recruitment is done primarily via web-based advertising and email invitations. Just like YouGov, Qualtrics and SurveyMonkey distribute surveys via email and reward participation by providing small incentives. The survey software provided by these companies allows the introduction of quotas to ensure a pre-specified level of participation of selected segments of the population. Because of the relatively low cost of these Internet surveys, researchers are often able to procure large online samples with broad coverage of the target population.

Regardless of the method used to recruit survey respondents, researchers should carefully consider aspects of the survey and experimental designs that could potentially affect the reliability and validity of survey responses. When using data from a sample to make inferences about a population, and when dealing with issues such as systematic non-response, researchers should also perform adequate statistical adjustments to obtain estimates that pertain to the target population of their studies. In the rest of this chapter we review approaches for dealing with design and representativeness issues that researchers commonly face when using online polling for persuasion research.

2 Design Issues

Researchers interested in conducting persuasion research often rely on experimental evidence, much of which is conducted on Internet panels. In this section we briefly discuss survey and experimental design issues. Each of these broad topics deserves a chapter of their own, and therefore we direct readers to reviews for additional information.

2.1 Survey Design Considerations

There are many aspects of the survey design process that researchers should consider when designing their survey. There are four main components of survey error that can help guide a researcher's design: coverage errors, sampling error, response error, and non-response error (Thompson, 1997). These sources of error can all lead to bias, and contribute to "total survey error" (Groves and Lyberg, 2010).

Coverage errors relate to the population that the survey is sampled from as it compares to the population the researcher cares about, and

it can exhibit over- or under-coverage depending on if too many or too few individuals are included, respectively, relative to the target population of interest. Where possible, researchers should aim to conduct their survey on populations as similar to their primary population of interest as possible. If researchers are interested in the impact of persuasion messages on election day voters, limiting the survey to likely voters may be more appropriate than conducting the survey among all registered voters, a population which is overly inclusive and thus exhibits overcoverage. Many of the methods used to address non-response, described below, are also applicable for mitigating coverage errors.

Response error deals with the problems of construct validity and measurement error. Researchers should take care to ensure that their instrument measures their quantity of interest by carefully considering the wording of survey questions and using randomization to mitigate potential question- and response-order effects. Whenever possible, researchers should also collect information about respondent attention to the questionnaire via inclusion of attention and manipulation checks, recording of metadata such as time spent completing the survey and specific questions, and keeping track of the frequency of non-attitudes (i.e. selection of middle categories or "don't know" responses) and straightlining (i.e. selection of the same response for multiple contiguous questions, particularly in the context of questions presented in the same page or in a grid). This information may later be incorporated into statistical analyses in order to account for response error arising from inattentiveness (Alvarez et al., 2019; Berinsky, Margolis, and Sances, 2014).

Non-response error, probably the leading concern for modern researchers, deals with the non-random nature of those individuals who respond to surveys. Convenience samples, which have helped lower the cost of conducting surveys and survey experiments, have similar statistical concerns as surveys with severe non-response.

Sampling errors are the easiest error for researchers to control in the design stage when they are using probability sampling. Researchers can employ complex sampling designs, such as stratified or cluster sampling, which will increase the precision of their estimators. If stratifying, researchers should stratify on variables that are predictive of the outcome of interest, or are predictive of non-response. If conducting a clustered design, researchers should construct clusters which are homogeneous in their outcome and response propensity. Beyond stratification and clustering, there are many complex survey sampling methods that may be of use to researchers, including respondent driven sampling, capture-recapture methods, and adaptive methods (see Thompson (2012) for more details). Most online surveys employ convenience samples which means an individual's probability of being included in the sample, or inclusion probability, is unknown. However, where researchers are responsible for their own data collection, they should use methods that will increase the representativeness and precision of their design. When the sampling process is managed by a third party vendor, it is still important to request known inclusion probabilities where they exist, as this information is important for estimation.

2.2 Experimental Design Considerations

From a design standpoint, researchers interested in survey experiments have many of the same considerations as they do for the survey design itself. If researchers can control randomization in the survey, among respondents, then randomization coupled with a stable unit treatment value assumption (SUTVA) is sufficient to identify the sample average treatment effect (Fisher, 1937; Cox, 1958).

When conducting randomization, researchers should consider blocking on covariates that explain variance in the outcome or the treatment effect, which will increase the precision of their estimators. If researchers are concerned about spillover effects, they may want to consider a clustered design, where possible, to mitigate bias from a SUTVA violation.

Finally, researchers should conduct power analyses testing their design in order to ensure they have sufficient power to detect their effect. Given the low cost nature of online convenience samples such as Amazon's Mechanical Turk, researchers can often easily conduct pilot studies that can help inform their experimental design and power analyses.

There are a myriad of other decision that researchers must make when designing any experiment, whether conducted using a survey or not, and we direct researchers to some starting points for good experimental design. In particular, see Mutz (2011), Gerber and Green (2012), Nock and Guterbock (2010), Barabas and Jerit (2010), Hainmueller, Hopkins, and Yamamoto (2013), and Broockman, Kalla, and Sekhon (2017) for information on experimental design, in particular as it pertains to the design or generalizability of survey experiments or the use of surveys as an outcome.

3 Methods for Dealing with Non-Sampling Errors

Survey response rates have steadily declined over the last half century, leaving survey data subject to non-sampling errors such as non-response bias (Kohut et al., 2012). Convenience samples, such as those employed by many Internet survey firms, are similarly subject to non-sampling errors, primarily coverage errors. Of course, the main concern about these unrepresentative samples is that they lead to bias in our estimators and misleading conclusions in our research.

The primary way that researchers deal with these non-representative surveys is through post-hoc weighting of the respondent data. The goal of weighting is to adjust the observed sample in such a way that it looks representative of the target population (e.g. the population of all registered voters) on a set of auxiliary variables (e.g. age, gender, race, voting history). There are many different weighting methods available to researchers, discussed below, each of which has advantages and disadvantages. Given the trade-offs of each method, and the decisions of what auxiliary information to include in the method, there are many researcher degrees of freedom that can impact the quality of the survey weights.

Researchers studying electoral persuasion may wish to consider weighting their data for many reasons. If they are using surveys to collect out-

come data, non-response or non-random sampling may lead to a need to adjust the data so that it is representative of a target population, such as all registered voters. Researchers studying persuasion using experimental evidence may wish to generalize their experimental findings, which requires accounting for the sample selection mechanism and adjusting, or re-weighting, the experimental findings. Whether considering observational or experimental survey methods in the study of electoral persuasion, researchers may find themselves in need of a way of constructing a set of weights that make their sample representative on observable characteristics. Our goal is to help researchers be informed consumers of survey weights and to aid researchers in constructing their own survey weights best tailored to their own research needs. We primarily consider the problem of unit nonresponse, in which we construct one set of weights for all responding units, irrespective of item response patters. We discuss item nonresponse separately.

3.1 Estimation under Ideal Conditions

The basis of most survey design and estimation relies on the idea that we have a random sample of the population. For example, many times we take a simple random sample of the population. This means that if we wanted to study the average approval of the president among all students on a 10,000 student college campus, we might take a random sample of 1,000 students and ask them about approval. In this case, each student has an equal probability of being included in the sample, referred to as the sample inclusion probability, π_k. In this case, each student has a sample inclusion probability of $\pi_k = 1000/10000 = 0.1$. In this case, when every unit has an equal probability of being in the sample, the mean approval in the sample is a good estimator for the mean approval in the population.

Under ideal conditions, in which the sampling design is known and there is no non-response, researchers typically rely on estimators that weight inversely proportional to the known probability of being included in the sample. However, not every unit is required to have an equal probability of being included in the sample. If we had over-sampled freshmen students at twice the rate of other students, then we would want to give those units a weight of one half in a weighted mean. This method of weighting inversely proportional to the probability of being sampled is called the Horvitz-Thompson (HT) estimator, and it is an unbiased estimator of population totals and means.[1] In this case, each unit has weight $w_k = 1/\pi_k$. Given the desireable properties of these estimators, and their ease of implementation, researchers should aim to rely as much as possible on the known sample inclusion probabilities when deriving their survey weights. In most applications, researchers will find that non-response is non-random, and therefore the design-justified estimators described above are insufficient for recovering unbiased estimates in the target population. The most common way to adjust unrepresentative samples is to conduct weighting. These methods aim to reduce bias that occurs when samples

[1] The Hàjek estimator, which accounts for the probability of an observed sample, has lower mean squared error.

are unrepresentative of the population we seek to study.

3.1.1 Inverse Response Propensity Weighting

The HT estimator forms the conceptual basis for a common form of post-hoc weighting method—response propensity weighting (Little, 1988). When non-response exists, or a survey was conducted on a convenience sample, the research does not know the true sample inclusion probability. When this occurs, researchers must model the non-response pattern to estimate the probability an individual is included in the survey, sometimes referred to as a propensity score. This is typically done by comparing units in the survey to units in a population and estimating the probability of being included in the sample, referred to as $\hat{\pi}$, using some form of generalized linear model. This estimate of $\hat{\pi}_k$ is used to assign a weight of $w_k = 1/\hat{\pi}_k$ for each individual, much as was done with the known sample inclusion probability in the HT estimator. Assuming that $\hat{\pi}_k$ is an unbiased estimator of π_k, then the inverse response propensity weighting estimator returns unbiased estimates of population totals and means.

3.2 Calibration

In the presence of systematic nonresponse, or when sample inclusion probabilities are not known, information about the characteristics of the population may be used to develop weights for adjusting the sample to resemble the population. All common weighting methods, except for inverse response propensity weighting, are special cases of a class of weighting estimators called "calibration" estimators. The basic idea of calibration is to construct weights that make the sample exactly match the population on a set of user-specified characteristics. For example, we make our sample of college students match our student body exactly on the breakdown of gender, age, and where they went to high school. Deville and Särndal (1992) provide the general framework for calibration estimators, unifying weighting techniques such as generalized regression, post-stratification, and raking. Following Deville and Särndal (1992), the general structure for constructing calibration weights is described below.

1. We define a population U from which we have a sample S where all units in the sample must be included in the population ($S \subset U$), and where each individual k has some positive sample inclusion probability ($\pi_k > 0$).
2. The population has known characteristics which we define using auxiliary information vector $\mathbf{x}$. In particular, we know something about the population moments defined in an auxiliary vector $\mathbf{t_x} = \sum_U \mathbf{x_k}$. Population moments are simply characteristics about our population, such as total, mean or variances of covariates measured in the population. For example, we may have information on the average age in our population, the percentage which are women, and the distribution of educational attainment. This information is assumed to be accurately known.

3. Researchers start with an initial set of weights, d_k, which are typically defined as either one for all units, if nothing is known about the sampling design, or the inverse of the sample inclusion probability $1/\pi_k$, where π_k is the sample inclusion probability for unit k.
4. New weights, w_k, are then chosen to minimize some total distance, defined by a distance function, between the starting and ending weights, $\sum_k D(w_k, d_k)$, while simultanously ensuring that the weights make the sample look like the population on the desired population characteristics ($\mathbf{t_x} = \sum_k w_k \mathbf{x_k}$). Intuitively this means that the weights must ensure that our sample looks representative on our characteristics such as average age, percentage which are women, and distribution of educational attainment.

The researcher must decide what population constraints to include and what distance function to use. Using these new weights w_k, the researcher can estimate population totals as weighted sums ($y_{total} = \sum_k y_k w_k$) and population means as weighted means ($\bar{y} = \frac{1}{N} \sum_k y_k w_k$ where N is the population size). Different weighting methods require different assumptions about either the response or outcome model for unbiased estimates of population quantities. With an outline of the general calibration procedure, we turn now to the different decisions a researcher must make and the implications of those decisions.

3.2.1 Auxiliary Vector

The most consequential decision that a researcher must make is what information to include in the auxiliary vector that defines the population moments. Auxiliary information is information that researchers have about all units in the both the target population and the sample. For example, from a voter file we may know the age and gender distribution in the registered voter population. If we either match our online panel back to the voter file, or we ask individuals for a self-reported age and gender, [2] then we can construct weights to deal with observed imbalances in this auxiliary information. Särndal and Lundström (2005) discuss the properties of a good auxiliary vector, and they indicate a good auxiliary vector includes variables that explain: (1) what types of units are likely to respond or not, (2) variation in the outcome of interest, or (3) interesting subgroups of the data. Later in this chapter, we discuss under what conditions weighting can reduce bias; bias reduction is directly related to how well the auxiliary vector explains either the non-response pattern or the outcome of interest. In the case of experimental persuasion effects, we care not only about explaining the variation in the outcome of interest, which can increase precision, but also explaining variation in the treatment effect. Generalizability of experimental findings depends, to a great extent, on the correlation of factors that relate to who is included in our

[2] Typically we would consider auxiliary information to be information measured the same way across the full population. In the case of self-reported information, such as age, where people might misreport, there can be bias introduced if we adjust self-reported information to a different measurement in the auxiliary vector.

survey, such as our convenience Internet sample, and factors that moderate the treatment effect. Therefore, researchers should use substantive knowledge to consider factors that explain their primary research outcome as well as those that explain how their sample came to be.

While different estimation strategies for constructing weights have different assumptions that can have important consequences for bias and variance reduction, often times the most significant factor is what variables applied researchers include in these statistical techniques. The Pew Research Center conducted a naturalistic simulation study of over 30,000 opt-in internet panel respondents in which they compared different weighting techniques that use a basic or more extensive set of covariates for the auxiliary vector in the adjustment procedure (Pew Research Center, 2018). They then use simulated samples from the 30,000 panelists in sample sizes ranging from 2,000 to 8,000 and apply adjustment techniques with the different auxiliary variable sets to see if they can recover known high quality benchmark values. Ultimately they find that even with extensive auxiliary variable sets, all of the statistical methods have a fair amount of remaining bias, with none reducing bias below 6 percentage points across 24 benchmark values. However, they find that most of that bias reduction, roughly one third, comes from the addition of the richer covariate sets. The more complex statistical techniques, holding the the covariate set constant, do not provide significantly more bias reduction.

Constructing a strong auxiliary vector is difficult, and a lot of current research considers data-driven methods for selecting covariate measures, but no automated method will serve as a panacea for overcoming problems with sample-selection and non-response. Researchers must use their substantive knowledge about the research question at hand and the response mechanism in their survey when constructing a good auxiliary vector.

3.2.2 Initial Weight

Survey researchers often rely on design-based inference to derive the properties of their estimators. As discussed earlier, when researchers know the sample inclusion probability for each unit, there are estimators with known, desirable properties such as unbiasedness (such as the HT) or low mean squared error. Under non-ideal conditions, such as with severe non-response, this means that d_k, the initial weight, should incorporate known sample inclusion probabilities π_k whenever possible. The properties of the calibration estimators, as originally derived by Deville and Särndal (1992) assume that d_k is the inverse of the known inclusion probability, and the goal is to deviate from these weights as little as possible, thus staying as true to the design-justified estimator as possible.

Of course, if researchers are relying on convenience samples, or they are not provided sample inclusion probabilities for a probability sample, this is not possible. In such cases researchers can either begin with unity, setting all the initial weights to one, i.e. $d_k = 1 \; \forall \; k$, or they can begin with the survey weights provided by the survey firm. If survey weights are provided by the firm, these may incorporate valuable information unavailable to the researcher such as initial inclusion probabilities or sensitive auxiliary information available to the firm but not the researcher. See Deville and

Method	$D(w_k, d_k)$	Bias Reduction	
		Response Model	Outcome Model
Generalized Regression	$\propto (w_k - d_k)^2$	linear in $\mathbf{x}_k$	linear in $\mathbf{x}_k$
Post-stratification	$\propto (w_k - d_k)^2$	homogenous within strata	linear in strata
Raking	$\propto w_k log(w_k/d_k)$	additive w/o interaction	additive w/o interaction

Table 1: Common calibration methods correspond to different distance functions, and lead to bias reduction under different assumptions.

Särndal (1992) or Kott (2016) for extensions when initial weights which deviate from the inverse of the sample inclusion probability are used.

3.2.3 Method of Weighting

There are many common weighting approaches that survey methodologists use when constructing post-hoc weights—primarily post-stratification and raking—each of which has advantages and disadvantages. Different commonly use weighting methods correspond to different distance metrics in the calibration framework, which we briefly discuss here. See Berinsky (2006) for a thorough treatment of these commonly used weighting techniques. Table 1 shows how the common weighting methods relate to different distance metrics, and the conditions under which they lead to near zero bias (Särndal and Lundström, 2005) in the estimation of population totals and means.

There are three common weighting methods seen in political surveys: generalized regression, post-stratification, and raking. Generalized regression uses a variant of linear regression to construct the weights.[3] The advantage of the generalized regression estimator is that it is easy to compute and can take in categorical as well as continuous auxiliary information. A main draw back of using linear weights is that it can result in negative weights. Likely for this reason, it is rarely used in Political Science.

Post-stratification is a special case of linear weighting in which the population moment constraints are made up of a set of population proportions of subgroups of the data.[4] The auxiliary vector, $\mathbf{x}$ is then composed of indicators for which subgroup an individual is in.

Post-stratification weights are the easiest weights to compute. They are determined by dividing the population proportion for that subgroup by the sample proportion.[5] For example, if we were to weight just on gender,

[3]Generalized regression corresponds to $D(w_k, d_k) \propto (w_k - d_k)^2$. Generalized regression, also called linear, weights are determined using:

$$w_k = d_k + d_k \times \mathbf{x}_k^\top (\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^\top)^{-1} (\mathbf{t}_\mathbf{x} - \sum_{k \in S} d_k \mathbf{x}_k)$$

where, once again, the $\mathbf{t}_\mathbf{x}$ define the population moment constraints for the corresponding auxiliary information $\mathbf{x}$.

[4]These subgroups must be defined in such a way that every unit is in one, and only one, subgroup.

[5]To construct an individual weight, first define a strata variable H which indicates strata membership for individual k. For an individual in strata h, the weight is then:

and we found that our sample was 40% women, but our population was 50% women, each woman would get a weight of $0.5/0.4 = 1.25$.

An advantage of post-stratification is that, if respondents are missing completely at random within strata, then we will eliminate bias in our estimator. The main drawback of post-stratification is that the data requirements are high. For precise estimates, we need reasonably large sample sizes within each stratum. Perhaps more problematically, if any strata are empty in the sample, but not in the population, the estimator is not defined because we cannot divide by a sample proportion of zero. This requires, then, that researchers change their target population to only those strata which are represented in the sample or otherwise relax their assumptions on the the nonresponse pattern.

One of the most common weighting methods is raking, an iterative method that often has weaker data requirements than post-stratification. In raking, researchers typically define population moments using only means of the population characteristics. For example, they define only the average age and the proportion of women, rather than defining the average age separately by gender. While raking can take in any population moments, including interacted margins, most survey weights are constructed using only the marginal mean.[6]

Raking, or iterative proportional fitting, iteratively post-stratifies to the marginal population constraints, using initial weights set to the weights calculated in the previous iteration, until the weights reach a user-specified level of convergence between iterations. The advantage of raking, particularly over post-stratification, is that it requires only knowledge of marginal population moments such as means rather than information about each characteristic in each subgroup.[7] While it has the same data requirements as commonly implemented linear weights, it has the advantage that it will enforce the weights to be positive. Its major drawback is that it requires strong modelling assumptions about how the auxiliary variables relate to either the outcome or the non-response model, namely that at least one of them is linearly additive in the included auxiliary variables. (Särndal and Lundström, 2005).

4 Additional Considerations

4.1 Trimming Weights

In the face of extremely non-representative samples or large auxiliary vectors weighting methods can yield very large, or extreme, weights. Some

$$w_{k:H=h} = \frac{\sum_U I(H=h)}{N} * \left(\frac{\sum_{k\in S} I(H=h)}{n}\right)^{-1}$$

where $I(H=h)$ is an indicator if that individual is in strata s.

[6]Raking corresponds to a distance metric $D(w_k, d_k) \propto w_k log w_k/d_k$ (Deville, Särndal, and Sautory, 1993). In many cases, the solution is the same as for a related weighting technique called "maximum entropy weighting".

[7]While the method does not require users to only match population margins, this is the most common formulation of population weights.

survey researchers suggest the use of methods that trim the extreme weights. The primary concern with extreme weights is increased sampling variance. Potter and Zheng (2015) discuss four common approaches to principled methods for trimming weights. Trimming extreme weights can induce bias, especially if extreme weights are also correlated with extreme values in the population. They indicate that weight trimming is not necessary when the weights and the outcome are negatively correlated, but can be useful for reducing variance when the extreme weights are associated with extreme values on the outcome. They suggest using past data, where possible, to determine if there is a strong correlation between the survey data and the weights to evaluate the impact on the sampling variance and possible bias. They find mixed evidence on the value of trimming weights, indicating that in some cases point and interval estimates with trimmed weights can be misleading.

4.2 Item Nonresponse

Up until this point, we have primarily considered unit non-response, in which researchers construct a set of weights for all responding units. However, in many surveys even respondents who choose to participate in the survey often do not respond to all questions. This process of selectively choosing which questions to respond to is a problem of item nonresponse.

In some cases, particularly for questions that are not used as an outcome, researchers can use multiple imputation (Rubin, 2004). Multiple imputation methods impute missing items using models and the observed joint distribution among other respondents. Common estimation techniques include MICE (Buuren and Groothuis-Oudshoorn, 2010) and AMELIA II (Honaker, King, and Blackwell, 2011). Researchers can impute missing values on auxiliary variables before calibration.

For missing items on outcomes that researchers wish to analyze, multiple imputation might not be appropriate. In this case, researchers may wish to construct a set of calibration weights, calibrating only individuals who responded to the item of interest. Recall that a good auxiliary vector is one that explains response propensity and the outcome. If researchers account for item nonresponse with a unique set of weights, they may also wish to consider a unique auxiliary vector for each question given the nonresponse pattern for the question and prognostic variables for that outcome. Accounting for item non-response in this way is likely only practical for researchers considering only a few outcomes of interest.

4.3 Panel Attrition

In panel designs where units are interviewed at two or more points in time, unit nonresponse may become a problem in later waves of the study. This may happen if researchers are unable to reestablish contact with individuals that participated in the initial waves, or if some respondents decline to be reinterviewed. In the event of wave nonresponse, inclusion probabilities or survey weights applicable to the first wave of the study may be unsuitable for adjusting data from later waves, as respondents retained in the panel may differ systematically from those lost due to panel attrition.

Common fixes used to address this problem include: model-based imputation of missing values based on responses given by all respondents in earlier waves and other respondents during the re-interview process; accounting for wave nonresponse by constructing wave-specific survey weights; and recruiting new respondents for participation in the later waves to compensate for respondents lost to attrition. For a detailed discussion of the advantages and drawbacks of some of these solutions, see Kalton 1986 and Chen et al. 2015.

4.4 Generalization of Experimental Findings

The study of electoral persuasion often uses experimental evidence (e.g. Kalla and Broockman (2018)). The use of randomization allows for causally identified effects of persuasion messages and methods in the sample, however in most scenarios experiments are not conducted on a random sample of the population. This limits the generalizability of the experimental findings. Recent literature has investigated the additional assumptions that are needed in order to recover the population average treatment effects from randomized trials (Cole and Stuart, 2010; Buchanan et al., 2018; Hartman et al., 2015; Kern et al., 2016; Pearl and Bareinboim, 2014; Stuart et al., 2011; Tipton, 2013; Xie, 2013). The most common approach that this literature addresses generalization of experimental findings is by accounting for the sample selection mechanism, and adjusting, or reweighting, the sample accordingly. If the selection process is modeled correctly, then the re-weighted experimental findings can be used to estimate the treatment effect in the target population. The degree of bias in estimation of the target population effect depends on how correlated the sampling process and the treatment effect heterogeneity variables are (Levin and Sinclair, 2018). Miratrix et al. (2017) find that the variance increase from applying survey weights, under the assumption they correctly account for the sampling mechanism, can swamp the bias reduction, and thus lead to increased mean squared error when the sampling mechanism is not highly correlated with treatment effect heterogeneity. As the bias in the sample is more extreme, however, weighting can be important.

If researchers do choose to reweight their experimental findings, one decision they need to make is whether to weight the treatment and control groups separately or not. Randomization will, in expectation, give balanced treated and control groups. However, any given randomization may yield slight imbalances between the treated and control groups. If researchers are weighting because they are interested in recovering a population effect, calibrating the treatment and control group separately will provide a treated and control group that are balanced in their representation of the population. Researchers should check for extreme weights if they use this approach, particularly in small samples.

Given calibration on small samples can lead to large weights and increased variance, weighting the treated and control units together, and appealing to the unbiasedness of the randomization in expectation, may be preferable. Most weighted survey experiments use weights that calibrate the whole sample together. If researchers use the combined data when conducting calibration, they should check that the weighted popu-

lation targets for both the treated and control groups are similar.

Additionally, some authors have found that many survey experiments generalize from convenience samples to nationally representative surveys, seemingly because many survey experiments do not show a great deal of heterogeneity (Coppock, Leeper, and Mullinix, 2018; Mullinix et al., 2015). This indicates researchers should test for heterogeneity in the effect before pursuing a reweighting strategy. Researchers should be careful when considering the generalizability of their results, especially when heterogeneity is present.

5 Application

In this section we illustrate the application of weighting and post-stratification techniques using data from a survey experiment embedded in the post-election wave of the University of Georgia (UGA) module of the 2014 Cooperative Congressional Election Study (Schaffner and Ansolabehere, 2017). This nationwide online survey was carried out by YouGov/Polimetrix (now YouGov) during the two weeks following the November 6, 2014 General Election. Here we use data from an unmatched sample consisting of members of YouGov's opt-in online panel that had not been matched to a target sample of the population nor weighted to YouGov's sampling frame. The total size of the unnmatched sample that saw questions in UGA's module is 1,548 respondents. Here, however, we focus on 994 respondents that participated in the second wave of the study, dropping four respondents that did not answer the question used to measure the outcome variable in our analysis. Data from the 2014 UGA CCES module (matched and unmatched samples) are available for download through the Harvard Dataverse Network (Hare, 2017).

Before analyzing the results of the survey experiment, we weighted the sample to all 103,771 adults in the November 2014 Current Population Survey (CPS). The CPS is a monthly survey conducted by the United States Census Bureau, with participant households selected via multi-stage stratified sampling and interviews carried out in person and over the phone. We selected CPS data to construct our population benchmark because of the survey's large sample size, national coverage, and use of probability sampling techniques. The five auxiliary variables we used for weighting our sample to the CPS were: Census region, age group, gender, ethnicity, and educational attainment. In Table 2 we present the marginal distribution of each of these variables for the CPS sample (weighted using the composite final weight available in the CPS data set), YouGov's matched sample, and unmatched samples (unweighted and weighted using linear weighting, raking, and post-stratification).

Compared to the adult American population according to the Current Population Survey (see first column of Table 2), the unmatched CCES module (see "No weights" column of Table 2) over-represents middle-aged groups and people who have college degrees, and under-represents non-white ethnic groups. While some of these differences are attenuated in YouGov's matched CCES module (see "Matched sample" column of Table 2), clear discrepancies remain, as non-whites are still severely

	Census (weighted CPS)	Matched sample	Unmatched sample			
			No weights	Raking	Linear weighting	Post-stratification
Region						
Northeast	18.1	17.3	18.0	18.1	18.1	17.9
Midwest	21.3	23.8	23.4	21.3	21.3	21.9
South	37.2	35.6	34.3	37.2	37.2	38.2
West	23.4	23.3	24.2	23.4	23.4	21.9
Age						
18 - 29 years	21.3	19.5	14.5	21.3	21.3	19.0
30 to 44 years	25.2	23.5	16.8	25.2	25.2	24.6
45 to 64 years	34.5	38.5	46.9	34.5	34.5	37.6
65+ years	19.0	18.5	21.8	19.0	19.0	18.8
Gender						
Male	48.2	46.5	44.6	48.2	48.2	44.9
Female	51.8	53.5	55.4	51.8	51.8	55.1
Ethnicity						
White	65.2	76.1	77.4	65.2	65.2	78.4
Black	11.6	9.5	9.3	11.6	11.6	7.7
Hispanic	15.3	7.6	6.9	15.3	15.3	9.2
Other	7.8	6.7	6.5	7.8	7.8	4.7
Education						
No HS	11.9	11.1	2.4	11.9	11.9	3.5
HS graduate	29.5	26.9	26.6	29.5	29.5	27.8
Some college	28.7	34.1	34.0	28.7	28.7	33.4
College graduate	30.0	27.9	37.0	30.0	30.0	35.4

Table 2: Population and sample characteristics by weighting scheme.

under-represented in the matched sample. Part of the explanation for this discrepancy could be that YouGov constructs its sampling frame (used to select a subset of respondents from the opt-in sample that resemble a random sample from the target population) as to mirror adult citizens in the American Community Survey, and that we only consider 867 respondents in the matched sample (out of 1,000) that participated in the second wave of the study and provided responses for the outcome variable of interest.

The last three columns of Table 2 show the marginal distribution of the five auxiliary variables found after adjusting the unmatched sample using raking, linear weighting, and post-stratification. Raking and linear weighting yield distributions that match the population targets (i.e. marginal distributions shown in the first column of Table 2) for these two procedures. Post-stratification produces weighted marginal distributions that resemble those in the CPS sample less closely, which can be explained by the fact that this method uses different population targets (proportions of the population within strata constructed based on values of all auxiliary variables, rather than population marginal distributions).

We next analyze the results of a survey experiment on attitudes toward immigration reform, comparing estimates of average treatment effects across weighting schemes. Respondents randomly assigned to different experimental conditions were exposed to different versions of a question that varied in terms of information on elite support for policy change. Respondents in the control group were asked: "In 2013, immigration leg-

	Matched sample	Unmatched sample			
		No weights	Raking	Linear weighting	Post-stratification
Groups	-6.4	-7.8	-4.1	-7.7	-9.0
	(5.6)	(3.8)	(5.3)	(4.8)	(3.5)
Party	-4.7	-8.4	-3.7	-4.9	-8.0
	(5.6)	(3.9)	(5.2)	(4.9)	(3.4)

Table 3: Treatment effects among matched and weighted unmatched samples.

islation was introduced in Congress that would provide a path to legal status for undocumented immigrants meeting certain requirements, such as passing background checks, and paying taxes and a penalty fee. How much do you support immediate action on immigration reform providing a path to legal status for undocumented immigrants meeting certain requirements?" Respondents in the 'groups' condition saw a similar question, except that before being asked about their level of support they were also told that "Political groups and organizations advocating for more restrictive immigration policies oppose immediate action on the bill since they believe immigration reform should wait until the U.S.-Mexican border is secure." Respondents in the 'party' condition were told that "House Republicans oppose immediate action on the bill since they believe immigration reform should wait until the U.S.-Mexican border is secure."

We next evaluate, for different weighting schemes, the average effect of exposure to 'groups' and 'party' experimental conditions (relative to the control group) on the likelihood of reporting support for reform (either weak or strong). Estimates of average treatment effects are shown in Table 3 (standard errors between parentheses). All estimates suggest that, on average, exposure to the experimental cues (that is, being told that groups or parties oppose legislation that would provide a path to legal status for undocumented immigrants until the U.S.-Mexican border is secure) reduces support for the legislation in question. While estimates are statistically significant at conventional levels in the unadjusted unmatched sample, as well as when the unmatched sample is weighted via post-stratification; adjusted effects are not statistically significant on either the matched sample or for the raking and linear weighting schemes—for these procedures, adjusted estimates are smaller in magnitude and have larger standard errors.

If a respondent attribute is important for explaining both survey participation and variation in treatment effects, excluding this variable from the weighting procedure may lead to inaccurate estimates of average treatment effects. To illustrate this point, we reestimated average treatment effects without adjusting for the under-representation of non-whites in the unmatched sample. We expected this change to be consequential, as non-whites react very differently to experimental cues compared to whites (among non-whites, being told that groups or parties oppose immediate action on immigration increases support for the legislation, rather than decrease it as among whites). Indeed, dropping ethnic group from the auxiliary information passed as input to the weighting schemes leads to

	Matched sample	Unmatched sample			
		No weights	Raking	Linear weighting	Post-stratification
Groups	-6.4	-7.8	-7.0	-8.9	-5.1
	(5.6)	(3.8)	(4.9)	(4.7)	(3.8)
Party	-4.7	-8.4	-5.7	-6.5	-4.0
	(5.6)	(3.9)	(4.8)	(4.7)	(3.8)

Table 4: Treatment effects among matched and weighted unmatched samples (without adjusting for ethnicity).

markedly different estimates of average treatment effects (see last three columns of Table 4), particularly for post-stratification where estimates are now considerably smaller and do not attain conventional levels of statistical significance. These results confirm the notion that what matters the most is not how weights are constructed, but what auxiliary variables go into the calibration procedure.

6 Conclusion

Online polling presents significant challenges for researchers relying on this form of data collection for learning about the effectiveness of persuasion tactics, as key features of the sampling design (e.g. inclusion probabilities) are often not known for Internet surveys. Opt-in Internet samples may differ systematically in important ways from the population of interest of research studies. This can be troublesome for persuasion research, particularly when the effect of exposure to persuasion attempts varies over segments of the population. Steps may be taken during the data collection process to ensure the sample covers most of the target population. For example, special procedures may be put in place to recruit hard-to-reach participants. In the absence of serious coverage error, researchers may rely on statistical adjustments to estimate treatment effects that apply to the target population. In this chapter we reviewed a general framework for drawing population-level inferences from non-representative data, known as calibration. This framework encompasses common techniques for post-hoc weighting of survey data, including post-stratification and raking.

Using data from a survey experiment on support for immigration reform, we illustrated the importance of the careful choice of auxiliary information for the calibration procedure for ensuring the generalizability of findings from non-representative Internet surveys. In selecting auxiliary variables to be used for developing calibration weights, researchers are generally advised to rely on their substantive knowledge. In the case of experimental research, in particular, researchers should include variables associated with inclusion in the opt-in sample, the outcome of interest, as well as variables accounting for variation in treatment effects across segments of the population. Researchers should also stay attuned to advances in survey research and computational methods that could facilitate the selection and measurement of auxiliary information—e.g. data-driven

approaches for selecting auxiliary vectors (Caughey and Hartman, 2017) and design-based approaches for collecting information on propensity to respond (Bailey, 2017).

References

Alvarez, R Michael, Lonna R Atkeson, Ines Levin, and Yimeng Li (2019). "Paying attention to inattentive survey respondents". In: *Political Analysis* 27.2, pp. 145–162.

Bailey, Michael A (2017). "Designing surveys to account for endogenous non-response". Annual meeting of the Society for Political Methodology, University of Wisconsin-Madison. URL: https://polmeth.polisci.wisc.edu/Papers/SelectionModels_March2017.pdf.

Baker, Reg, Scott Keeter, Courtney Kennedy, and Andrew Mercer (2016). *Evaluating survey quality in today's complex environment*. Tech. rep. American Association for Public Opinion Research. URL: https://doi.org/10.7910/DVN/WTPQIJ.

Barabas, Jason and Jennifer Jerit (2010). "Are survey experiments externally valid?" In: *American Political Science Review* 104.2, pp. 226–242.

Berinsky, Adam J (2006). "American public opinion in the 1930s and 1940s: The analysis of quota-controlled sample survey data". In: *International Journal of Public Opinion Quarterly* 70.4, pp. 499–529.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk". In: *Political Analysis* 20.3, pp. 351–368.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances (2014). "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys". In: *American Journal of Political Science* 58.3, pp. 739–753.

Broockman, David E, Joshua L Kalla, and Jasjeet S Sekhon (2017). "The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs". In: *Political Analysis* 25.4, pp. 435–464.

Buchanan, Ashley L, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero (2018). "Generalizing evidence from randomized trials using inverse probability of sampling weights". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 95, p. 1082.

Buuren, S van and Karin Groothuis-Oudshoorn (2010). "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* 4.3, pp. 1–68.

Caughey, Devin and Erin Hartman (2017). "Target selection as variable selection: Using the Lasso to select auxiliary vectors for the construction of survey weights". Annual meeting of the Society for Political Methodology, University of Wisconsin-Madison. URL: `https://polmeth.polisci.wisc.edu/Papers/caughey_hartman_lasso_weighting_polmeth2017.pdf`.

Chen, Qixuan, Andrew Gelman, Melissa Tracy, Fran H Norris, and Sandro Galea (2015). "Incorporating the sampling design in weighting adjustments for panel attrition". In: *Statistics in medicine* 34.28, pp. 3637–3647.

Cole, Stephen R and Elizabeth A Stuart (2010). "Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial". In: *American journal of epidemiology* 172.1, pp. 107–115.

Coppock, Alexander (2018). "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach". In: *Political Science Research and Methods* 7.3, 1—16.

Coppock, Alexander, Thomas J Leeper, and Kevin J Mullinix (2018). "The generalizability of heterogeneous treatment effect estimates across samples". In: *Proceedings of the National Academy of Sciences* 115.49, 12441—12446.

Cox, D R (1958). *Planning of experiments*. Oxford, England: John Wiley & Sons.

Deville, Jean-Claude and Carl-Erik Särndal (1992). "Calibration estimators in survey sampling". In: *Journal of the American Statistical Association* 87.418, pp. 376–382.

Deville, Jean-Claude, Carl-Erik Särndal, and Olivier Sautory (1993). "Generalized raking procedures in survey sampling". In: *Journal of the American statistical Association* 88.423, pp. 1013–1020.

Dutwin, David and Trent D. Buskirk (2017). "Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability Internet samples to low response rate probability samples". In: *Public Opinion Quarterly* 81.S1, pp. 213–239.

Fisher, Ronald Aylmer (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Gerber, Alan S and Donald P Green (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Groves, Robert M and Lars Lyberg (2010). "Total survey error: Past, present, and future". In: *Public opinion quarterly* 74.5, pp. 849–879.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto (2013). "Causal inference in conjoint analysis: Understanding multidi-

mensional choices via stated preference experiments". In: *Political Analysis* 22.1, pp. 1–30.

Hare, Chris (2017). *CCES 2014, team module of University of Georgia (UGA)*. URL: https://doi.org/10.7910/DVN/WTPQIJ.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon (2015). "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 178.3, pp. 757–778.

Holbrook, Allyson L and Jon A Krosnick (2010). "Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity". In: *Public Opinion Quarterly* 74.2, pp. 328–343.

Honaker, James, Gary King, Matthew Blackwell, et al. (2011). "Amelia II: A program for missing data". In: *Journal of statistical software* 45.7, pp. 1–47.

Huff, Connor and Dustin Tingley (2015). ""Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents". In: *Research & Politics* 2.3.

Kalla, Joshua L and David E Broockman (2018). "The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments". In: *American Political Science Review* 112.1, pp. 148–166.

Kalton, Graham (1986). "Handling wave nonresponse in panel surveys". In: *Journal of Official Statistics* 2.3, pp. 303–314.

Kern, Holger L, Elizabeth A Stuart, Jennifer Hill, and Donald P Green (2016). "Assessing methods for generalizing experimental impact estimates to target populations". In: *Journal of Research on Educational Effectiveness* 9.1, pp. 103–127.

Kohut, Andrew, Scott Keeter, Carroll Doherty, Michael Dimock, and Leah Christian (2012). "Assessing the representativeness of public opinion surveys". In: *Washington, DC: Pew Research Center*.

Kott, Philip S (2016). "Calibration weighting in survey sampling". In: *WIREs Computational Statistics* 8, pp. 39–53.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau (2008). "Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity". In: *Public Opinion Quarterly* 72.5, pp. 847–865.

Levin, Ines and Betsy Sinclair (2018). "Causal inference with complex survey designs: Generating population estimates using survey weights". In: *Oxford handbook of polling and survey methods*. Ed. by L R Atkeson and R M Alvarez. New York, NY: Oxford University Press, pp. 299–315.

Little, Roderick JA (1988). "Missing-data adjustments in large surveys". In: *Journal of Business & Economic Statistics* 6.3, pp. 287–296.

Miratrix, Luke W, Jasjeet S Sekhon, Alexander G Theodoridis, and Luis F Campos (2017). "Worth weighting? How to think about and use sample weights in survey experiments". In: *arXiv preprint arXiv:1703.06808*.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese (2015). "The generalizability of survey experiments". In: *Journal of Experimental Political Science* 2.2, pp. 109–138.

Mutz, Diana C (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Nock, Steven L and Thomas M Guterbock (2010). "Survey experiments". In: *Handbook of survey research* 2, pp. 837–865.

Pearl, Judea and Elias Bareinboim (2014). "External validity: from do-calculus to transportability across populations". In: *Statistical Science* 29.4, pp. 579–595.

Pew Research Center (2018). *For weighting online opt-in Samples, what matters most?* Tech. rep.

Potter, Frank and Yuhong Zheng (2015). "Methods and issues in trimming extreme weights in sample surveys". In: *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 2707–2719.

Rubin, Donald B (2004). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Särndal, Carl-Erik and Sixten Lundström (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Schaffner, Brian and Stephen Ansolabehere (2017). *CCES common content, 2014*. URL: `https://doi.org/10.7910/DVN/XFXJVY`.

Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf (2011). "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 174.2, pp. 369–386.

Thompson, Mary (1997). *Theory of sample surveys*. Vol. 74. CRC Press.

Thompson, Steven K (2012). *Sampling, Third Edition*. Hoboken, NJ: John Wiley & Sons.

Tipton, Elizabeth (2013). "Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts". In: *Journal of Educational and Behavioral Statistics* 38.3, pp. 239–266.

Xie, Yu (2013). "Population heterogeneity and causal inference". In: *Proceedings of the National Academy of Sciences* 110.16, pp. 6262–6268.