Statistical Science
2017, Vol. 32, No. 2, 279–292
DOI: 10.1214/16-STS597

Options for Conducting Web Surveys

Matthias Schonlau and Mick P. Couper

Abstract. Web surveys can be conducted relatively fast and at relatively low cost. However, Web surveys are often conducted with nonprobability samples and, therefore, a major concern is generalizability. There are two main approaches to address this concern: One, find a way to conduct Web surveys on probability samples without losing most of the cost and speed advantages (e.g., by using mixed-mode approaches or probability-based panel surveys). Two, make adjustments (e.g., propensity scoring, post-stratification, GREG) to nonprobability samples using auxiliary variables. We review both of these approaches as well as lesser-known ones such as respondent-driven sampling. There are many different ways Web surveys can solve the challenge of generalizability. Rather than adopting a one-size-fits-all approach, we conclude that the choice of approach should be commensurate with the purpose of the study.

Key words and phrases: Convenience sample, Internet survey.

1. INTRODUCTION

Web or Internet surveys[1] have come to dominate the survey world in a very short time (see Couper, 2000; Couper and Miller, 2008). The attraction of Web surveys lies in the speed with which large numbers of people can be surveyed at relatively low cost, using complex instruments that extend measurement beyond what can be done in other modes (especially paper). Nonetheless, there remain a number of concerns about the value of Web surveys, especially for those who need relatively precise estimates of general populations. We believe that part of the ongoing debate about the inferential value of Web surveys is rooted in broad generalizations about the mode of data collection, often ignoring the fact that there are many different types of Web surveys, and they serve many different purposes.

For instance, for general population surveys, no sampling frame or method exists that permits direct selection and invitation of sample persons to a Web survey. No complete list of e-mail addresses of the general population exists from which one can select a sample and send e-mailed invitations to a Web survey. However, for many other important populations of interest (e.g., college students, members of professional associations, registered users of Web services, etc.), such lists do exist and offer near-complete coverage of the population of interest. Similarly, no method exists for generating a random sample of e-mail addresses analogous to the random digit dial (RDD) sampling methods that made telephone surveys so popular in their heyday. This means that some other method must be used to sample and invite members of the general population to complete Web surveys (e.g., address based sampling [ABS] and mailed invitations). Others are using mixed-mode approaches (combining Web with mail or other modes) to address both the sampling and coverage problems associated with Web surveys. Because of these sampling challenges, a large number of alternative approaches have been developed to identify samples of Internet users (or broader populations) in order to make broader inferences. These methods are discussed in more detail below.

But there are also many other uses of Web surveys where coverage is not the primary concern. For groups with near-universal Internet access and an available sampling frame, nonresponse error may be a bigger

Matthias Schonlau is Professor, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave. West, Bldg. M3, Waterloo, Ontario, N2L 3G1, Canada (e-mail: schonlau@uwaterloo.ca). Mick P. Couper is Professor, Institute for Social Research, University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106, USA (e-mail: mcouper@umich.edu).

[1]We use the terms "Web survey," "Internet survey" and "online survey" interchangeably. Our specific focus is on surveys completed in a Web browser.

concern. Similarly, experiments have long relied on volunteers from selective segments of the population. The Internet broadly expands the range of people available for such experiments, and may reduce inferential concerns about experiments conducted on highly-selective groups. Similarly, randomized controlled trials (RCTs)—whether in health, economics or other fields—have benefited from the expanded access to many more—and more varied—subjects than was possible in the past.

For these reasons, Web surveys have not replaced all other modes of data collection [as some early proponents of the method proclaimed (see Couper, 2000)], but have added a powerful new set of tools to the survey researcher's toolkit. The question then becomes one of for which types of populations and research questions are Web surveys (either alone or in combination with other methods) the optimal approach? When should they be used, and when should they be avoided? This is fundamentally a question of fitness for use, or fitness for purpose. The inferential standards or requirement for all surveys are not the same, and understanding these different purposes or goals is important in choosing the most appropriate method for the research question at hand. Similarly, when evaluating whether a Web survey is "good" or "bad," we should always be asking "relative to what?" Understanding the variety of ways Web surveys can be designed and deployed can help researchers in choosing the best method (or methods) for the task.

The inferential issues thus vary across different research questions. For instance, coverage concerns may be much greater for those studying disadvantaged and/or marginalized populations. One would want to be cautious (for example) when using Web surveys to estimate quantities related to Internet access or use—the unbanked or underbanked, those in poverty, the uninsured, etc. Similarly, Web surveys may overestimate political interest and participation, but may be perfectly fine for studying the behavior or intentions of likely voters. Some point estimates, such as of population means or intercepts in regression models, may be biased, but the estimation of associations (or slope coefficients) may be less affected by the type of sample one uses. Sometimes appropriate auxiliary variables are available to permit adjustments to reduce bias; other times, no amount of adjustment may correct the inherent biases (Bethlehem, 2010). Given that one size does not fit all, sweeping generalizations about the value (or lack thereof) of Web surveys for all types of populations and research questions should be avoided. The field is now turning to a deeper exploration of when and where different types of Web surveys are most appropriate.

An overview of the paper is as follows: Section 2 reviews probability-based approaches; Section 3 discusses estimators that use both probability samples and nonprobability samples. Sections 4 covers nonprobability approaches. Section 5 reviews Web surveys with a nontraditional sampling method, respondent driven sampling. Section 6 reviews adjustments to nonprobability samples and Section 7 discusses their relative success in practice. Section 8 concludes with a discussion.

2. PROBABILITY-BASED APPROACHES

There are three main approaches for obtaining a probability sample with Web surveys.

2.1 Recruit Off-Line, Conduct Online

One option is to obtain a probability sample by contacting sample persons through a traditional survey mode (i.e., face-to-face, random digit dialling, or by mail) and asking them to complete a Web survey.

This method has two disadvantages. One, a traditional recruitment method comes with traditional costs. Opportunities for cost savings are reduced. Two, when respondents are contacted in one survey mode it may be hard to convert them to another survey mode. Web as the only response mode is therefore unattractive. For example, one such study recruited US high school students by mail, but was ultimately forced to allow a secondary response mode (Schonlau, Asch and Du, 2003). Nonetheless, many survey organizations are using mailed invitations because of the availability of address-based samples (including population registers), whether for stand alone Web surveys or for mixed-mode (mail and Web) surveys. This leads to the next option: a mixed-mode approach.

2.2 Mixed Mode

A survey mode refers to how respondents are contacted and/or how they respond (e.g., face-to-face, mail survey, Web survey, telephone). Mixed mode refers to allowing multiple response or contact modes. A traditional probability sample is usually drawn. Sample persons are then given the choice of responding in one of two or more survey modes. This approach is often used in populations where the only contact information is a mailing address and/or where the Internet status of the sample is unknown. Mixed-mode surveys can address coverage issues because some respondents may

be reached through one mode but not another. Mixed-mode surveys can also improve selective nonresponse if a respondent would respond in one mode but not another. There are two broad approaches to mixed-mode designs involving Web data collection: concurrent mixed modes or sequential mixed modes.

In the concurrent mixed-mode approach, sample persons are typically sent a questionnaire by mail, with the option of completing the survey on the Web. This is also referred to as a mail survey with a Web option. In a meta-analysis of 19 experimental comparisons, Medway and Fulton (2012) found that providing a Web option in a mail survey had significantly lower response rates (OR = 0.87) than mail only. The authors give three possible reasons for this: First, being asked to make a choice increases respondent burden. Respondents may simply not get around to deciding, or a conscious thought process may lead them to dismiss both alternative choices. Second, the transition from mail to Web is too disruptive. Respondents may decide to respond on the Web, discard the paper questionnaire, but later they never follow through with the Web option. Finally, respondents may experience Internet connection problems which may discourage them sufficiently from participating. Despite the lack of empirical support, the mail-with-Web-option design is widely used, especially for large-scale data collection such as censuses (see Zewoldi, 2011). The reason for this is primarily to reduce costs rather than increase response rates.

The second broad class of approaches involves sequential mixed-mode designs. Rather than offering two response options simultaneously, sample persons are first sent an invitation (by mail) to complete a Web survey. Later, a mail questionnaire is offered to those who have not yet responded. Nonrespondents at that stage may be followed up using an interviewer-administered approach. While evidence on the effect of this approach on overall response rates is still quite mixed, starting with Web increases the proportion of Web responses, potentially reducing costs (see, e.g., Couper, 2012; Holmberg, Lorenc and Werner, 2010; Hughes and Tancreto, 2015). For this reason, sequential mixed-mode designs are increasingly being adopted for large-scale mail surveys such as the American Community Survey. Other statistical agencies are exploring mixed-mode surveys starting with the Web. For example, Statistics Netherlands draws a random sample from the population register, sends invitations by mail, solicits responses via the Web and then reapproaches nonrespondents by phone (if a number can be found). Recent research has focused on evaluating nonresponse bias and measurement error in sequential mixed-mode designs (e.g., Klausch, Schouten and Hox, 2015; Schouten et al., 2013).

Address-based sampling (ABS) is widespread. Usually this is used in mixed-mode designs (concurrent or sequential). For example, the American Community Survey (ACS) sends mailed invitations to households to complete the form online, following up with paper questionnaires and then CATI and CAPI for nonresponding households. Censuses in the UK, US, Canada, New Zealand, Australia, Japan and other countries use this approach.

2.3 Probability-Based Web Panels

Recognizing the desire of researchers to conduct Web surveys on a probability sample of respondents, several general population panels have sprung up in recent years.

These include in the Netherlands the LISS panel and the Dutch Immigrant Panel (www.lissdata.nl) (Scherpenzeel, 2011); in the USA the GfK Knowledge Panel (formerly Knowledge Networks), the Gallup panel (Callegaro et al., 2014), the USC's Understanding America Study (UAS, https://uasdata.usc.edu/), the Pew American Trends panel (Keeter and Weisel, 2015) and NORC's AmeriSpeak Panel (Dennis, 2015); in Germany the German Internet Panel (GIP) (Blom, Gathmann and Krieger, 2015) and the GESIS panel (www.gesis-panel.org); in France the ELIPSS panel (www.elipss.fr); in Norway the Norwegian Citizen panel (Høgestøl and Skjervheim, 2013); and in Sweden the Citizen panel (Citizen Panel, 2015). Some of these panels are restricted for internal use (e.g., GIP, Pew), others are open to broader academic use (e.g., LISS, GESIS Panel), and still others have a blend of uses, including commercial, academic and government clients (e.g., GfK Knowledge Panel, NORC's AmeriSpeak).

Most of these panels take a traditional offline approach to recruiting a probability sample. For example, GIP used a 3-stage probability sample with 250 primary sampling units. In each PSU, 22 addresses were sampled using a random route procedure. The GESIS panel drew a sample from German municipal registers. The UAS panel is recruited by inviting respondents of a (probability-based) mail survey to join the panel at the end of the survey. The Norwegian panel was able to obtain a probability sample from a national registry. The Pew panel recruited respondents from a large RDD phone survey. Gallup recruited via both

RDD and address-based sampling. The probability-based GfK Knowledge panel now also accepts volunteer members (http://join.knpanel.com/). The Swedish panel similarly consists of two separate probability and volunteer samples.

While Internet penetration rates these days are high, not everyone in the target population has Internet access. To represent the offline population, most panels (LISS, Dutch Immigrant Panel, GIP, GfK Knowledge Panel) provide such respondents with a free computer and Internet access. The ELIPSS panel offers every respondent (regardless of whether he or she has Internet access) the same tablet computer. This eliminates device or mode effects. The GESIS, Pew and Gallup Panels have chosen a mixed-mode approach: offline respondents are sent mail surveys. Usually, there is only a modest percentage of non-Internet respondents. However, in the GESIS panel 38% of survey respondents respond offline (GESIS Panel, 2015). The Pew panel recruited a little more than 10% mail-only panel members (Keeter and Weisel, 2015) which receive up to three mailings for each survey. The majority of Gallup Panel members respond via the Internet. About a fifth (22%) of the recruited AmeriSpeak Panel households are non-Internet households (Dennis, 2015) which are interviewed by telephone (as reported in Keeter and Weisel, 2015).

The advantage of such probability-based Web panels is that the samples can be reused for many surveys and, therefore, the recruiting costs per survey are much lower. Challenges include multiple layers of nonresponse, attrition, the concern that professional respondents may respond differently over time (panel fatigue) and the need to refresh the sample if it becomes too small or less representative due to attrition. In practice, the challenge of maintaining representativity over time in such panels is addressed by computing weights that adjust for nonresponse and by replenishing the panel every few years. This problem is common to all panels, whether online or not. As for responding differently over time, experiments conducted with trained respondents and fresh respondents lead to the same conclusions (Toepoel, Das and Van Soest, 2008), though trained respondents tend to satisfice more, including an increased tendency to straightline (Schonlau and Toepoel, 2015). Evidence of panel conditioning (respondents' answers are affected by participating in previous waves) was only found for knowledge questions (as it should) but not for "questions on attitudes, actual behavior, or expectations concerning the future" (Das, Toepoel and van Soest, 2011).

NORC's AmeriSpeak panel reports a cumulative AAPOR RR3 response rate of 13% to 20% for client surveys (Dennis, 2015). Such low cumulative response rates are a common problem for probability-based Web panels. However, data can be gathered at each stage of the recruitment process that may be helpful for nonresponse adjustment. Probability-based Internet panels are currently a big growth area.

3. COMBINED ESTIMATORS USING BOTH A NONPROBABILITY SAMPLE AND A PROBABILITY SAMPLE

Probability samples are expensive. Nonprobability samples are much cheaper but estimates based on nonprobability samples are often biased. An estimate based on combined probability and nonprobability samples may result in a lower mean squared error (MSE) than an estimate based on a probability sample alone (Elliott and Haviland, 2007, Ghosh-Dastidar et al., 2009, Schonlau, Fricker and Elliott, 2002, Appendix). Depending on whether the MSE of the combined estimator is smaller than the estimator based on the probability sample alone, either the combined or the probability-sample based estimate is used.

When is this approach useful? The combined estimator only has a smaller MSE than the probability-sample estimator if the bias is very small, the probability sample is large (1000–10,000 observations) and the nonprobability sample is much larger still. This makes this approach cost efficient only if the nonprobability sample is much cheaper than the probability sample (Elliott and Haviland, 2007).

Ghosh-Dastidar et al. (2009) applied this approach to a study where the target population consisted of families with 3-5-year-old children. This inclusion criterion made an RDD phone survey expensive. At the same time a nonprobability sample could be bought from a marketing company with tens of thousands of families with young children. For 38 of 41 outcomes reported on, the combined estimator had a lower mean squared error relative to spending all resources on an RDD sample.

This approach has several drawbacks: One, if the bias is too large for any one parameter of interest, the nonprobability sample is wasted and the estimate must be based on the probability sample alone. Two, the complexity of the analysis increases. There has also been no work on regression with this approach. Three, the approach may be vulnerable to being "gamed" by

practitioners. For example, practitioners might strategically exclude observations from the nonprobability sample to reduce the estimated bias.

Despite these drawbacks, we think that this approach is under used in the practice of conducting Web surveys. In particular, applied studies are needed to establish to what extent the approach is practical, and under which circumstances the bias is reduced.

4. NONPROBABILITY APPROACHES

Given the relatively low cost and speed of conducting surveys on the Web, many approaches have been developed that focus primarily on getting large numbers and diverse groups of respondents to complete surveys. Some of these explicitly attempt population representation, others do so more implicitly, and still others make no claims about inference to a broader population. Since the advent of Web surveys, this is the area of most rapid growth.

The development of opt-in or access panels, in which panelists are recruited through a variety of nonprobability methods and invited to participate in surveys, saw rapid growth in the early years of Web surveys, especially in the domain of market research (Couper, 2000). In recent years, rising concerns about oversaturation and inattentive or fraudulent respondents may have slowed the growth of such panels, but there remain scores of panels across most countries with sufficiently large Internet populations (see AAPOR, 2010; Callegaro et al., 2014). Surveys administered to these panels are increasingly used by academics and other researchers for their convenience and affordability. Sometimes these are used to run experiments; other times they are used to generate descriptive statistics about populations.

Given rising concerns about the quality of access panels (see Faasse, 2005), increasing effort has been made to identify alternative approaches for recruiting large numbers of respondents for surveys. One such approach is called river sampling (see, e.g., Baker-Prewitt, 2010; DiSogra, 2008), involving the recruitment of people browsing the Web and directing them to a particular survey (i.e., a catch-and-release approach, as contrasted with the catch-and-retain approach to building opt-in panels). While the research is scarce, we know of no evidence that this method is superior in terms of data quality or inferential error than the access panel approach. Similarly, others have advocated using blended panels (in which the same survey is administered to members of several different panels, and the results aggregated; see, e.g. Lorch, Cavallaro and van Ossenbruggen, 2010) to protect against unusual results. Again, there is no evidence to suggest that this approach yields reliably better estimates than a single panel.

Intercept sampling usually refers to stopping passersby on the street for an interview on the spot. However, it can also refer to intercepting people online as they are browsing to trying to reach a particular website. A number of variations on the intercept approach[2] have been developed. For instance, Google Consumer Surveys intercepts users accessing restricted material, and requires the completion of two survey questions in order to access such material (see McDonald, Mohebbi and Slatkin, 2012). But independent evaluations of this approach are rare (see Keeter and Christian, 2012). A more recent approach, random domain intercept technology (Seeman, 2015; Seeman et al., 2016) exploits that fact that people make mistakes when browsing the Web, and redirects mistyped URLs and broken Web links to an invite to complete a short survey. While the methods and sources vary, these approaches all rely on "capturing" Internet users and inviting them to participate. The selection biases inherent in such approaches are largely unknown (and often unknowable).

With the recent rise in social media, there are a number of new approaches to obtaining volunteer samples for surveys or experiments. These include recruitment using Facebook or similar social media sites, and online exchanges such as Amazon's Mechanical Turk (see, e.g., Antoun et al., 2015; Berinsky, Huber and Lenz, 2012; Brickman Bhutta, 2012; Buhrmester, Kwang and Gosling, 2011; Nelson et al., 2014). In another example, Wang et al. (2015) recruited Xbox users to complete surveys leading up to the 2012 presidential election. While the raw data were biased toward younger persons and males, Wang et al. used post-stratification to bring estimates in line with other forecasts. Many of those using these self-selected approaches make no claims about representation, while others do make such claims, whether implicitly or explicitly (e.g., O'Donovan and Shave, 2007; Couper, 2007).

In their summary of nonprobability panels, the AAPOR (2010), p. 52 task force concluded that "Researchers should avoid nonprobability online panels

[2]Intercept sampling usually refers to stopping a (systematic) sample of those passing by for an interview on the spot (e.g., mall intercept survey; exit polls). However, it can also refer to intercepting people online as they are browsing or trying to reach a particular website (see Couper, 2000, p. 485).

when a key research objective is to accurately estimate population values." They further note that "... claims of 'representativeness' should be avoided when using these sample sources." On the other hand, the task force acknowledged that "There are times when a nonprobability online panel is an appropriate choice." We believe this would apply in equal measure to other nonprobability recruitment and selection methods. Given the wide variability in results from opt-in panels (see, e.g., Craig et al., 2013; Erens et al., 2014; Vonk, van Ossenbruggen and Willems, 2006; Yeager et al., 2011) and, by extension, other methods of subject recruitment (e.g., Antoun et al., 2015), this suggests that researchers should be cautious making broad inferential claims on the basis of a single study using nonprobability methods.

5. WEB SURVEYS WITH RESPONDENT-DRIVEN SAMPLING

Respondent-driven sampling (RDS) is a chain referral sampling technique. It was originally developed to recruit rare or hidden populations (e.g., HIV populations, drug users, homeless), using offline social networks (Heckathorn, 1997). RDS has now become a key methodology in AIDS surveillance at the US Centers for Disease Control and Prevention (CDC) (CDC, 2016).

In RDS, several seed respondents are purposively sampled and asked to recruit four (or some other number) of their friends, or people in their social network. The respondent does the recruitment; the interviewer never obtains names or contact information; hence the term respondent-driven sampling. The respondent receives a (financial) incentive both for completing the survey themselves and for each successful recruit. The survey contains a question about the number of a respondent's (eligible) friends to estimate network size. The use of coupons passed on from the respondent to his/her recruits allows recording of who recruited whom. The number of friends to be recruited has to be calibrated carefully: it has to be small enough to make paying incentives for long referral chains feasible and large enough to avoid the recruiting chain dying out. A methodological overview is given by Gile and Handcock (2010). RDS is a model-based technique and assumptions include the requirement that respondents recruit *at random* from their social network and that the recruiting chain is sufficiently long to render negligible the bias induced by the original sample.

RDS is typically conducted with a physical recruiting station at which interviewers can be contacted. This works well for individual cities but is expensive and does not scale to a broader region. Therefore, a small number of researchers have explored using RDS in conjunction with Web surveys. An early attempt on a college campus was so successful that recruiting had to be stopped after a single weekend (Wejnert and Heckathorn, 2008). Later attempts in different settings did not match this success. Bauermeister et al. (2012) conducted a Web-based RDS study on drug and alcohol consumption. Their study at a single university concluded after 2.5 months. Bengtsson et al. (2012) conducted a Web-based RDS study of men who have sex with men in Vietnam for about two months. Schonlau, Weidmer and Kapteyn (2014) attempted to build a Web panel of the US population using RDS, finding the recruiting process and respondents' reluctance to recruit their friends challenging.

Stein et al. (2014) used a Web implementation of RDS to study contact patterns. They encouraged respondents to contact their friends by Facebook. This required the use of an app that created private messages with a unique link. Contact by email was also allowed. The maximal number of recruitment waves in this pilot program was six. The study is remarkable in that it did not provide a payment incentive: the only incentive provided was respondent's ability to follow the progression of the recruitment tree online.

In practice, RDS assumptions are always violated to some degree. RDS estimators are asymptotically unbiased assuming a sufficiently large number of sampling waves and sufficiently low homophily in the network (and other conditions). Homophily is the tendency to associate with similar people; in the extreme case forming sub-networks of similar people that are not linked to one another. In practice, long recruiting chains are difficult to obtain and for moderately long chains asymptotic unbiasedness may not be reached. Respondents are also supposed to recruit "at random" from their social network. In practice, this is rarely the case and the impact on nonrandom recruiting on estimators is not well understood (see Gile and Handcock, 2010). The sensitivity of assumptions is discussed in Lu et al. (2012).

In summary, approaches to Web-based implementations of respondent driven sampling are still evolving and assumptions tend to be violated in practice. Implementation details are critical for convincing respondents to participate.

6. ADJUSTMENTS FOR WEB SURVEYS WITH NONPROBABILITY SAMPLES

Both opt-in Web panels and open-access Web surveys are nonprobability samples which are subject to selection bias and coverage error. Therefore, adjustments for correcting such errors are required. Adjustments rely on auxiliary variables—which should be related to both outcome variables and the propensity to respond (and thus be part of the sample)—that are measured in the nonprobability sample. Either the population totals of the auxiliary variables must be known from elsewhere (Section 6.1), or a probability-based reference sample must be available that also measures the auxiliary variables (Sections 6.2 and 6.3). Section 6.4 covers GREG estimation that can be used in both situations. All of these techniques rely on auxiliary adjustment variables which are discussed in Section 6.5.

Also see the overlapping exposition of this topic by Michael Elliott and Richard Valliant in their paper on nonprobability samples in this issue (Elliott and Valliant, 2017).

6.1 Post-Stratification to Population Totals

If the population distribution of the auxiliary variables is known, post-stratification can be used to reweight observations in the nonprobability sample to match the known population distribution on those variables. Cells or weighting classes are formed by crossing all categories of auxiliary variables. Any continuous auxiliary variables are first turned into categorical variables.

When only marginal totals of the auxiliary variables are known rather than their full distribution, or when there are so many strata that some strata are (nearly) empty, raking can be employed. Post-stratification and raking are discussed in this context, for example, in Valliant, Dever and Kreuter (2013), Chapter 14.2 and Bethlehem and Biffignandi (2011), Chapters 10.2.2 and 10.2.4. When the population distribution is not known, estimates from a high quality probability survey can be used as the basis for post-stratification or raking.

Such adjustments may lead to large weights. In practice, most analysts trim large weights. This represents a bias-variance trade-off where trimmed weights lead to biased estimates with lower variance.

Post-stratification is probably the most popular technique for adjustments for Web surveys in part because only cell totals—not individual-level data—are required. Socio-demographic variables are used most often as auxiliary variables because their population totals are more likely to be available from censuses or reference samples. The assumption is that adjusting on the available auxiliary variables will reduce the bias of estimates of all other variables in the survey.

6.2 Propensity Scoring with a Reference Sample

Critical for the success of any adjustment are good auxiliary variables. However, often only demographic variables are available from external sources. To be able to calibrate the Web survey to auxiliary variables it is sometimes useful to conduct a second smaller reference survey with a probability sample (e.g., RDD). RDD is typically used as the reference survey, because few (if any) can afford a CAPI survey simply to adjust a nonprobability Web survey. This second survey only asks a subset of 10–15 questions that are used as auxiliary variables and acts as a reference sample to correct the selection bias from the nonprobability sample. This reference survey can be reused for multiple Web surveys of the same target population, increasing cost efficiency. When a reference survey is available, typically propensity scoring is used to adjust for possible bias. This approach was pioneered for Web surveys by Harris Interactive, a commercial survey company (Taylor et al., 2001). [Of course, propensity scoring has a long history predating the Harris Interactive approach (see, e.g., Little and Rubin, 2002).]

The propensity score is defined as the conditional probability $\rho(X_k)$ that respondent k with auxiliary values X_k responds (or self-selects) into the nonprobability sample (e.g., Bethlehem and Biffignandi, 2011, Chapter 11):

$$\rho(X_k) = P(R_k = 1 | X = X_k),$$

where R_k is an indicator of membership in the nonprobability sample. The propensity score is usually estimated with a logistic regression of R_k on the combined survey and reference samples. Also see the corresponding discussion in the paper by Elliot and Valliant in this issue, and Lee (2006), Lee and Valliant (2009) and Schonlau et al. (2009) for more details and examples.

The purpose of the propensity score is to balance the samples with respect to the auxiliary variables (Valliant, Dever and Kreuter, 2013). That is, groups of respondents in the nonprobability sample and the reference survey with the same propensity score have roughly the same distribution of auxiliary variables.

The extent to which balance is achieved should be tested.

The propensity score adjustment can proceed in one of several ways: One, the propensity score can be partitioned into five strata. Five is by far the most common choice because Cochran (1968) found that five strata suffice to remove most of the removable bias. Observations in the same stratum are assumed to be balanced. The propensity strata estimator of the population mean of y, $\bar{y}_{\text{ps}}$ is

$$\bar{y}_{\text{ps}} = \frac{1}{n}\sum_{h=1}^{5} n_h \bar{y}^{(h)},$$

where h indexes the 5 strata, n_h is the sample size in stratum h of the reference sample, $n = \sum_h n_h$ is the sample size of the reference sample, and $\bar{y}^{(h)}$ is the mean of outcome y in stratum h in the nonprobability sample. Note y is only measured in the nonprobability sample, not in the reference sample. This is a post-stratification estimator where the reference sample defines the strata based on the propensity score.

Two, the inverse propensity score can be used as weights: $w_i = (1 - \rho(x_i))/\rho(x_i)$ where $\rho(x_i)$ is the estimated response propensity. The weights adjust to the population corresponding to the probability sample rather than to the population corresponding to the two combined samples. This is analogous to the use of weights in observational studies when targeting the population of the treated rather than the combined population of treated and untreated. Then

$$\bar{y}_{\text{pw}} = \frac{\sum_{i=1}^{n_n} w_i y_i}{\sum_{i=1}^{n_n} w_i},$$

where $\bar{y}_{\text{pw}}$ refers to the propensity weight estimate, i indexes respondents in the nonprobability sample, n_n is the sample size of the nonprobability sample, y is an outcome of interest. Both methods will increase the variance of estimates. Very large weights arising from inverse propensity scores can be trimmed representing the usual bias-variance trade off.

Whereas the weighting class adjustments only require aggregate data, the propensity score modelling requires individual-level data. The success of the propensity scoring adjustments depends strongly on whether the available auxiliary variables adequately capture the difference between the nonprobability sample and the reference sample with respect to all outcomes of interest. This is the so-called strong ignorability assumption (Little and Rubin, 2002). If the adjustment only works partially for a given sample, most likely the strong ignorability assumption is not (fully) met.

6.3 Sample Matching with a Reference Sample

Rivers (Rivers, 2007; Vavreck and Rivers, 2008; Rivers and Bailey, 2009) proposed "sample matching." For each respondent of a probability sample a matching respondent from a large nonprobability sample is found using auxiliary variables. Collectively, the matched respondents in the nonprobability sample are referred to as the "matched sample." Only respondents from the (cheaper) nonprobability sample complete the survey; respondents of the probability sample only act as a reference sample. Estimates are produced based on the matched sample using sampling weights from the corresponding matches in the probability sample. Elliott and Valliant briefly mention this technique as a variation of "sample matching" in their paper in this issue (Elliott and Valliant, 2017).

Using a reference sample and the idea that only respondents in the nonprobability sample complete the survey is very similar to the propensity scoring approach described in Section 6.2. A key difference is that matching is performed on several auxiliary variables rather than on the one-dimensional propensity score.

Vavreck and Rivers (2008) found that estimates based on this approach have a smaller mean squared error than those based on random digit dialing in predicting election results for the 2006 Cooperative Congressional Election Study. As with the propensity scoring adjustment, the choice of the auxiliary variables is crucial for this approach to succeed. Bethlehem (2016) concluded that "with respect to nonresponse bias reduction, sample matching has no substantial advantages over stratified sampling and post-stratification estimation."

6.4 Generalized Regression Estimation (GREG)

The GREG estimator uses auxiliary variables as x-variables in a linear regression setting. The GREG estimator of the population mean of y is defined as

$$\bar{y}_{\text{GR}} = \bar{y} + (\bar{X} - \bar{x})^t b,$$

where $\bar{y}$ is the sample mean of an outcome variable, $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)^t$ is the vector of sample means of p auxiliary variables, $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)^t$ is the vector of population means of the auxiliary variables and $b = (b_1, b_2, \ldots, b_p)$ is the vector of regression coefficients (e.g., Bethlehem and Biffignandi, 2011, Section 10.2.3). The coefficients b are estimated using the usual least squares regression estimator. For unequal probability sampling inclusion probabilities

should be taken into account. The equation above makes an adjustment only if the sample means $\bar{x}$ differ from the population means $\bar{X}$. It can be shown that post-stratification is a special case of GREG estimation where the auxiliary variables are categorical.

In practice, GREG estimation is rarely used for adjustments in nonprobability samples (but see Dever, Rafferty and Valliant, 2008) presumably because it is less well known and many nonprobability samples do not bother with adjustments to begin with.

6.5 Auxiliary Variables

Each of the approaches described above uses auxiliary variables, although the way they are used differs between methods. Importantly, auxiliary variables must be measured in the Web survey and in the reference survey (if available). For auxiliary variables to be actually useful (i.e., to reduce bias), they must be related to both the propensity to respond (or the probability of being selected) and related to the key outcome variables of interest. Many adjustment schemes focus on the first condition and ignore the second condition. It is much harder to focus on the second condition in part because there are often many variables of interest. Bias and variance are reduced if both conditions hold (Bethlehem, 2010).

What auxiliary variables might be useful? In addition to demographic variables, so-called webographic or attitudinal variables have been proposed that capture the difference between the online and the offline population. There is no universally recognized suite of questions. Typical questions include "Do you often feel alone?" and "In the last month have you read a book?" Some additional questions are listed in Appendix A in Schonlau et al. (2004). While such questions are designed to adjust for undercoverage, by adjusting to a reference survey they also adjust for selection bias from nonresponse or other sources. For example, if the reference survey finds 20% of the population often feel alone, by adjusting to that number nonresponse in the Web survey is accounted for also.

Harris Interactive, a commercial Web survey company, pioneered the use of webographic questions but their suite of questions is not disclosed. (However, questions could be guessed by observing unusual questions near the end of their questionnaires). To our knowledge, few vendors of nonprobability (opt-in) panels use propensity score adjustment and webographic variables. This may reflect less on the merit of this method and more on the unwillingness of clients to pay for the extra work.

7. ADJUSTMENTS IN PRACTICE

The purpose of adjustments is to improve the estimators, that is, to reduce their mean squared error. For probability-based surveys, such adjustments typically proceed in separate steps (e.g., Valliant, Dever and Kreuter, 2013; Haziza and Beaumont, 2017):

- Base weights correct for unequal probability of selection.
- Weights are adjusted for sample units with unknown eligibility (for surveys where not all sample units are eligible).
- Nonresponse weights correct for nonresponse error.
- Weights are adjusted (e.g., via post-stratification) to correct for coverage error and to reduce variance (e.g., trimming).

In nonprobability surveys, adjustments are usually performed in a single step. Base weights in nonprobability surveys do not exist. Separate adjustments for sample units with unknown eligibility and nonresponse adjustment are not needed. Therefore, the concern is calibration to control totals from census, register or (probability) survey data. Often this involves only demographic variables because no other variables are available. Less often other variables such as attitudinal, lifestyle or "webographic" variables are used in propensity score adjustments.

A number of studies have explored to what extent the adjustments remove bias in practice. Yeager et al. (2011) compared seven nonprobability panels to a probability panel and RDD telephone survey. They found that probability samples were generally more accurate than nonprobability panels. On a range of unadjusted outcome measures, the average absolute error were 3.6% and 4% for two probability samples as compared to a range of 4.8% to 8.9% for 7 nonprobability Web surveys. Post-stratification reduced the average absolute error in both probability samples to 2.9% and 3.4% but only in 4 out of the 7 nonprobability samples (ranging from 4.5% to 6.6%).

Using a large number of questions from the Health and Retirement Study (HRS), Schonlau et al. (2009) compared sample means in the Internet access subsample of the HRS compared to the full, large HRS probability sample. They found the differences of the means were reduced from 6.5% before the propensity adjustment to 3.7% after the propensity adjustment, on average. Comparing an RDD and a nonprobability Web survey, Schonlau et al. (2004) found the difference in 8 of 37 outcomes not to be statistically significant.

Propensity score adjustments generally reduced differences but did not eliminate them.

Schnorf et al. (2014) compared 6 different surveys on privacy attitudes. They found respondents of opt-in panels had much lower privacy discomfort on average than respondents of probability samples, Amazon's Mechanical Turk or respondents of Google Consumer Surveys.

Vonk, van Ossenbruggen and Willems (2006) compared 19 different Dutch panels to Statistics Netherlands data. They found that individual response history is an important adjustment variable and conclude "when samples are balanced on such factors as number of invitations received and number of surveys completed, panel membership no longer influences research outcomes." (Vonk, van Ossenbruggen and Willems, 2006, p. 76). Intriguingly, respondents that belong to more than one panel—62% of respondents—are also the respondents most likely to respond (i.e., complete a survey). Pasek and Krosnick (2010) found significant and substantial differences in opinions and behaviours measured between an opt-in Internet and RDD tracking survey on attitudes to the US Census. Further, the socio-demographic composition of the RDD sample more closely resembled that the US population as compared to the opt-in Internet sample.

In summary, adjustments for inference from nonprobability Web surveys typically (but not always) reduce biases. Adjustments typically increase the variance of the estimates because they lead to more heterogeneous weights, inflating the design effect.

7.1 The 2015 UK Elections

A high profile example of nonprobability sampling with adjustments gone wrong are the polls leading up to the 2015 UK elections. Multiple online and phone polls predicted a virtual dead heat between Conservatives and Labour. Surprisingly, almost all polls were within one percentage point of one another. On election day, however, the Conservatives won by approximately seven points, stunning the public and pollsters alike. All polls had used quota sampling: Online polls recruited respondents from nonprobability online panels and phone polls recruited respondents from customer databases and other methods. The commission charged with the analysis of this phenomenon found the primary cause to be unrepresentative samples (Sturgis et al., 2016). A late swing (time trend) towards Conservatives played only a minor role and there was no discernible mode effect (online polls vs. phone polls). The low variability among the polls was thought to be consistent to herding behaviour, that is, correcting a poll to be more consistent with earlier polls. Sturgis et al. (2016) concluded that switching to probability samples in future elections—while ideal—is unrealistic because it would be prohibitively expensive. Instead, the study recommended improving representation in weighting cells (using traditional techniques, i.e., more reminders, incentives) and finding better auxiliary variables that are correlated with both the propensity to cast a vote *and* the voting outcome.

8. DISCUSSION

In this paper, we have briefly reviewed a number of different ways in which Web survey samples can be obtained, and some ways in which statistical adjustment can be made to reduce the biases that may be inherent in some of the approaches. There are several key messages.

One is that there are a wide variety of ways in which samples can be drawn for Web surveys. Treating all Web surveys as the same, and evaluating them uniformly (e.g., as "good" or "bad") is a risky practice. The different approaches have different strengths and weaknesses and should be evaluated relative to their stated purpose and the inferential claims that they make. Achieving broad population representation, whether through a mixed-mode approach or a probability-based panel, may be most expensive and time-consuming. This may undermine the attraction of Web surveys for many, and may be unnecessary and wasteful in many cases.

A second, and related, point is that surveys are done for many different purposes. Some of the objectives or uses of Web surveys include the following:

- Pre-testing survey instruments.
- Exploratory research on low-incidence or hard-to-reach populations.
- Experiments.
- Trend analysis (in stable populations).
- Correlation/regression analysis.
- Prevalence estimates (whether full population or key subgroups).

The inferential demands of these uses may be quite different. The further one goes down the list, the more important a probability sample may be. Or, if not a probability sample, the more important careful adjustments may be needed to compensate for biases that may affect estimates. Not all research endeavors require a high-quality probability sample (see also AAPOR, 2010).

It is important to understand the variety of purposes and map the appropriate method to those purposes. There is growing acknowledgement (AAPOR, 2013) that nonprobability surveys have also a place, and even probability surveys—especially those with very low response rates—suffer from inferential errors.

Third, statistical adjustments are not a silver bullet that can "fix" all inferential problems with nonprobability samples. Sometimes they work, and sometimes they do not. Even with a single survey, they may reduce bias for some estimates or some subgroup or domain of interest, but not for others. While the "optimal" set of auxiliary variables that jointly explain both the variation in the propensity to respond and the variation in the outcomes can reduce bias and variance, in practice statistical adjustments often come with a penalty of increased variance. However, this is often not reflected in the standard errors and confidence intervals generated from such samples. There have been some efforts to generate alternative measures of statistical uncertainty (e.g., credibility intervals as an analog to margins of sampling error used in probability sampling), but these are not without controversy (see AAPOR, 2012). Increasingly, surveys of all types (whether probability-based or not) are relying on model-based or model-assisted estimation, and the success of any adjustments depends in large part on the quality of the models. This leads to our next point.

Fourth, for those attempting to make broad generalizations from nonprobability surveys, adjustment should not be viewed as an afterthought. Given the importance of the auxiliary variables used for adjustment, choosing such variables should be an explicit part the design process. These should be based on a careful consideration of the sources of biases that may exist given the selection method employed, and given the outcome variables of interest. There is no one set of auxiliary variables that will be equally effective across all topics and populations. Careful consideration of the variables needed will lead to more thoughtful application of adjustment methods that may more effectively reduce bias in estimates. Further, the focus should not solely be on bias reduction, but the methods should also be chosen or evaluated based on their effect on the precision of estimates. Good estimates are those that minimize mean squared error (both bias and variance) for a given cost, that is, constrain them to an acceptable level given the study's purpose. There is a trade-off between cost and precision. Increasing sample size will reduce variance but not bias. With almost-unlimited budgets, improving precision will require a probability sample with high response rates and low noncoverage to achieve relatively unbiased estimates. Such budgets are now out of reach of most researchers.

Fifth, in order to increase our understanding of when and how best to use Web surveys, openness in reporting is critical. The purpose of the survey and the intended use of the resulting estimates should be made explicit at the outset. Further, especially for surveys making broad claims of representation, a careful analysis of the potential bias and the likely effect this may have on estimates, is important. Similarly, claims of representation should be accompanied by indicators of the uncertainty of the estimates. If statistical adjustments are used, researchers should provide more details on the model specification and the variables used in the models. Ideally, sensitivity analyses would be conducted to evaluate the effect different models have on estimates. Increasingly, journals are requiring that replication materials be made available. Details of the recruitment and selection process, and the data collection protocol are also needed. This will allow others to evaluate the quality of the survey, and independently judge the value of the findings, rather than simply relying on the claims of the author or designer. We believe such increased openness will help move the field of Web survey design and analysis forward. Efforts like the Transparency Initiative of AAPOR and the Disclosure Requirements of the National Council on Public Polls are trying to encourage the survey community in a similar direction.

In conclusion, our view is that Web surveys are valuable tools in the survey researcher's toolkit. They do not replace other tools, but expand the range of tools available. There is not one tool that is optimal for all purposes and circumstances. Web surveys are a very valuable method, but should be used appropriately.

REFERENCES

AAPOR (2010). *AAPOR report on online panels*. American Association for Public Opinion Research, Deerfield, IL.

AAPOR (2012). Understanding a 'credibility interval' and how it differs from the 'margin of sampling error' in a public opinion poll. Available at http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/DetailedAAPORstatementoncredibilityintervals.pdf.

AAPOR (2013). *Report of the AAPOR Task Force on Nonprobability Sampling*. American Association for Public Opinion Research, Deerfield, IL.

ANTOUN, C., ZHANG, C., CONRAD, F. G. and SCHOBER, M. F. (2015) Comparisons of online recruitment strategies for convenience samples Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods*. DOI:10.1177/1525822X15603149.

BAKER-PREWITT, J. (2010). Looking beyond quality differences: How do consumer buying patterns differ by sample source? Paper presented at the CASRO panel conference, New Orleans, LA.

BAUERMEISTER, J. A., ZIMMERMAN, M. A., JOHNS, M. M., GLOWACKI, P., STODDARD, S. and VOLZ, E. (2012). Innovative recruitment using online networks: Lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (WebRDS) strategy. *J. Stud. Alcohol Drugs* **73** 834–838.

BENGTSSON, L., LU, X., NGUYEN, Q. C., CAMITZ, M., HOANG, N. L., NGUYEN, T. A., LILJEROS, F. and THORSON, A. (2012). Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. *PLoS ONE* **7** e49417.

BERINSKY, A. J., HUBER, G. A. and LENZ, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* **20** 351–368.

BETHLEHEM, J. (2010). Selection bias in Web surveys. *Int. Stat. Rev.* **78** 161–188.

BETHLEHEM, J. (2016). Solving the nonresponse problem with sample matching? *Soc. Sci. Comput. Rev.* **34** 59–77.

BETHLEHEM, J. and BIFFIGNANDI, S. (2011). *Handbook of Web Surveys*. Wiley, New York.

BLOM, A. G., GATHMANN, C. and KRIEGER, U. (2015). Setting up an online panel representative of the general population the German Internet panel. *Field Methods* **27** 391–408.

BRICKMAN BHUTTA, C. (2012). Not by the book: Facebook as a sampling frame. *Sociol. Methods Res.* **41** 57–88.

BUHRMESTER, M., KWANG, T. and GOSLING, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* **6** 3–5.

CALLEGARO, M., BAKER, R. P., BETHLEHEM, J., GÖRITZ, A. S., KROSNICK, J. A. and LAVRAKAS, P. J. (2014). *Online Panel Research*: *A Data Quality Perspective*. Wiley, New York.

CDC (2016). National HIV behavioral surveillance (NHBS). Available at http://www.cdc.gov/hiv/statistics/systems/nhbs/index.html.

CITIZEN PANEL (2015). The Citizen Panel at the University of Gothenburg. Available at http://lore.gu.se/surveys/citizen.

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **25** 295–313.

COUPER, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opin. Q.* **64** 464–494.

COUPER, M. P. (2007). Issues of representation in eHealth research (with a focus on Web surveys). *Am. J. Prev. Med.* **32** S83–S89.

COUPER, M. P. (2012). Assessment of innovations in data collection technology for understanding society. Report to the Economic and Social Research Council, UK. Available at http://eprints.ncrm.ac.uk/2276/.

COUPER, M. P. and MILLER, P. V. (2008). Web survey methods: Introduction. *Public Opin. Q.* **72** 831–835.

CRAIG, B. M., HAYS, R. D., PICKARD, A. S., CELLA, D., REVICKI, D. A. and REEVE, B. B. (2013). Comparison of US panel vendors for online surveys. *J. Med. Internet Res.* **15** e260.

DAS, M., TOEPOEL, V. and VAN SOEST, A. (2011). Nonparametric tests of panel conditioning and attrition bias in panel surveys. *Sociol. Methods Res.* **40** 32–56. MR2758298

DENNIS, M. (2015). Technical overview of the AmeriSpeak panel—NORC's probability-based research panel. Technical report.

DEVER, J. A., RAFFERTY, A. and VALLIANT, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Surv. Res. Methods* **2** 47–62.

DISOGRA, C. (2008). River samples: A good catch for researchers? Knowledge Networks Newsletter. Available at http://www.knowledgenetworks.com/accuracy/fall-winter2008/disogra.html.

ELLIOTT, M. N. and HAVILAND, A. (2007). Use of a Web-based convenience sample to supplement a probability sample. *Surv. Methodol.* **33** 211–215.

ELLIOTT, M. R. and VALLIANT, R. (2017). Inference for nonprobability samples. *Statist. Sci.* **32** 249–264.

ERENS, B., BURKILL, S., COUPER, M. P., CONRAD, F., CLIFTON, S., TANTON, C., PHELPS, A., DATTA, J., MERCER, C. H., SONNENBERG, P. et al. (2014). Nonprobability Web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *J. Med. Internet Res.* **16** e276.

FAASSE, J. (2005). Panel proliferation and quality concerns. In *Proceedings of ESOMAR Conference on Worldwide Panel Research*: *Developments and Progress*, *Budapest*, *Hungary* 159–169, ESOMAR, Amsterdam. [CD].

GESIS PANEL (2015). GESIS panel study descriptions (version 11.0.0). Available at http://www.gesis.org/en/services/data-collection/gesis-panel/gesis-panel-data-usage/.

GHOSH-DASTIDAR, B., ELLIOTT, M. N., HAVILAND, A. M. and KAROLY, L. A. (2009). Composite estimates from incomplete and complete frames for minimum-MSE estimation in a rare population an application to families with young children. *Public Opin. Q.* **73** 761–784.

GILE, K. J. and HANDCOCK, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.

HAZIZA, D. and BEAUMONT, J.-F. (2017). Construction of weights in surveys: A review. *Statist. Sci.* **32** 206–226.

HECKATHORN, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.

HØGESTØL, A. and SKJERVHEIM, Ø. (2013). The Norwegian Citizen Panel: 2013 first wave. Available at http://www.uib.no/en/citizen.

HOLMBERG, A., LORENC, B. and WERNER, P. (2010). Contact strategies to improve participation via the Web in a mixed-mode mail and Web survey. *J. Off. Stat.* **26** 465–480.

HUGHES, T. and TANCRETO, J. (2015). Refining the Web response option in the multiple mode collection of the American Community Survey. Paper presented at the European Survey Research Association Conference, Reykjavik.

KEETER, S. and CHRISTIAN, L. (2012). *A Comparison of Results from Surveys by the Pew Research Center and Google Consumer Surveys*. Pew Research Center for the People & The Press, Washington, DC.

KEETER, S. and WEISEL, R. (2015). Building Pew Research Center's American Trends Panel. Pew Research Center report. Available at http://www.pewresearch.org/files/2015/04/2015-04-08_building-the-ATP_FINAL.pdf.

KLAUSCH, T., SCHOUTEN, B. and HOX, J. J. (2015). Evaluating bias of sequential mixed-mode designs against benchmark surveys. *Sociol. Methods Res.*. DOI:10.1177/0049124115585362.

LEE, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *J. Off. Stat.* **22** 329–349.

LEE, S. and VALLIANT, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **37** 319–343.

LITTLE, R. J. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

LORCH, J., CAVALLARO, K. and VAN OSSENBRUGGEN, R. (2010). Sample blending: $1 + 1 > 2$. Paper presented at the CASRO Panel Conference, New Orleans, LA.

LU, X., BENGTSSON, L., BRITTON, T., CAMITZ, M., KIM, B. J., THORSON, A. and LILJEROS, F. (2012). The sensitivity of respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **175** 191–216. MR2873802

MCDONALD, P., MOHEBBI, M. and SLATKIN, B. (2012). Comparing Google Consumer Surveys to existing probability and non-probability based Internet surveys. White paper. Google, Mountain View, CA. Available at http://www.google.com/insights/consumersurveys/static/358002174745700394/consumer_surveys_whitepaper.pdf.

MEDWAY, R. L. and FULTON, J. (2012). When more gets you less: A meta-analysis of the effect of concurrent Web options on mail survey response rates. *Public Opin. Q.* **76** 733–746.

NELSON, E. J., HUGHES, J., OAKES, J. M., PANKOW, J. S. and KULASINGAM, S. L. (2014). Estimation of geographic variation in human papillomavirus vaccine uptake in men and women: An online survey using Facebook recruitment. *J. Med. Internet Res.* **16** e198.

O'DONOVAN, G. and SHAVE, R. (2007). British adults' views on the health benefits of moderate and vigorous activity. *Prev. Med.* **45** 432–435.

PASEK, J. and KROSNICK, J. A. (2010). Measuring intent to participate and participation in the 2010 census and their correlates and trends: Comparisons of RDD telephone and non-probability sample Internet survey data. Technical report 2010:15, Statistical Research Division of the US Census Bureau, Washington DC.

RIVERS, D. (2007). Sampling for Web surveys. Paper presented at the Joint Statistical Meetings in Salt Lake City, UT.

RIVERS, D. and BAILEY, D. (2009). Inference from matched samples in the 2008 US national elections. In *Proceedings of the Joint Statistical Meetings* 627–639.

SCHERPENZEEL, A. (2011). Data collection in a probability-based Internet panel: How the LISS panel was built and how it can be used. *BMS. Bull. Méthodol. Sociol.* **109** 56–61.

SCHNORF, S., SEDLEY, A., ORTLIEB, M. and WOODRUFF, A. (2014). A comparison of six sample providers regarding online privacy benchmarks. In *SOUPS Workshop on Privacy Personas and Segmentation*, Menlo Park, CA.

SCHONLAU, M., ASCH, B. J. and DU, C. (2003). Web surveys as part of a mixed-mode strategy for populations that cannot be contacted by e-mail. *Soc. Sci. Comput. Rev.* **21** 218–222.

SCHONLAU, M., FRICKER, R. and ELLIOTT, M. N. (2002). *Conducting Research Surveys via E-mail and the Web*. RAND Corporation, Santa Monica, CA.

SCHONLAU, M. and TOEPOEL, V. (2015). Straightlining in Web survey panels over time. *Surv. Res. Methods* **9** 125–137.

SCHONLAU, M., WEIDMER, B. and KAPTEYN, A. (2014). Recruiting an Internet panel using respondent-driven sampling. *J. Off. Stat.* **30** 291–310.

SCHONLAU, M., ZAPERT, K., PAYNE SIMON, L., HAYNES SANSTAD, K., MARCUS, S. M., ADAMS, J., SPRANCA, M., KAN, H., TURNER, R. and BERRY, S. H. (2004). A comparison between responses from a propensity-weighted Web survey and an identical RDD survey. *Soc. Sci. Comput. Rev.* **22** 128–138.

SCHONLAU, M., VAN SOEST, A., KAPTEYN, A. and COUPER, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociol. Methods Res.* **37** 291–318. MR2649462

SCHOUTEN, B., VAN DEN BRAKEL, J., BUELENS, B., VAN DER LAAN, J. and KLAUSCH, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Soc. Sci. Res.* **42** 1555–1570.

SEEMAN, N. (2015). Use data to challenge mental-health stigma. *Nature* **528** 309. Available at http://www.nature.com/news/use-data-to-challenge-mental-health-stigma-1.19033.

SEEMAN, N., TANG, S., BROWN, A. D. and ING, A. (2016). World survey of mental illness stigma. *J. Affective Disorders* **190** 115–121.

STEIN, M. L., VAN STEENBERGEN, J. E., CHANYASANHA, C., TIPAYAMONGKHOLGUL, M., BUSKENS, V., VAN DER HEIJDEN, P. G. M., SABAIWAN, W., BENGTSSON, L., LU, X., THORSON, A. E. et al. (2014). Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: A pilot study in Thailand. *PLoS ONE* **9** e85256.

STURGIS, P., BAKER, N., CALLEGARO, M., FISHER, S., GREEN, J., JENNINGS, W., KUHA, J., LAUDERDALE, B. and SMITH, P. (2016). Report of the inquiry into the 2015 British general election opinion polls. Available at http://eprints.ncrm.ac.uk/3789/1/Report_final_revised.pdf.

TAYLOR, H., BREMER, J., OVERMEYER, C., SIEGEL, J. W. and TERHANIAN, G. (2001). The record of Internet-based opinion polls in predicting the results of 72 races in the November 2000 US elections. *Int. J. Mark. Res.* **43** 127–136.

TOEPOEL, V., DAS, M. and VAN SOEST, A. (2008). Effects of design in Web surveys comparing trained and fresh respondents. *Public Opin. Q.* **72** 985–1007.

VALLIANT, R., DEVER, J. A. and KREUTER, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, Berlin. MR3088726

VAVRECK, L. and RIVERS, D. (2008). The 2006 cooperative congressional election study. *J. Elect. Publ. Opin. Part.* **18** 355–366.

VONK, T., VAN OSSENBRUGGEN, R. and WILLEMS, P. (2006). A comparison study across 19 online panels (NOPVO 2006). In *Access Panels and Online Research, Panacea or Pitfall?* (I. Stoop and M. Wittenberg, eds.). *DANS Symposium Publications* **4**. Aksant Academic Publishers, Amsterdam.

WANG, W., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* **31** 980–991.

WEJNERT, C. and HECKATHORN, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociol. Methods Res.* **37** 105–134.

YEAGER, D. S., KROSNICK, J. A., CHANG, L., JAVITZ, H. S., LEVENDUSKY, M. S., SIMPSER, A. and WANG, R. (2011).

Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opin. Q.* **75** 709–747.

ZEWOLDI, Y. (2011). Introduction, 2011. Seminar on New Technologies in Population and Housing Censuses: Country Experiences, 42nd session of the United Nations Statistical Commission, New York. Available at http://unstats.un.org/unsd/statcom/statcom_2011/Seminars/NewTechnologies/default.html.