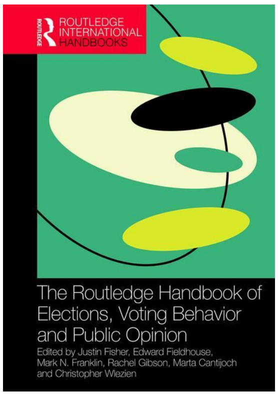

The Routledge Handbook of Elections, Voting Behaviorand Public Opinion

Justin Fisher, Edward Fieldhouse, Mark N. Franklin, Rachel Gibson, Marta Cantijoch, Christopher Wlezien

The quest for representative survey samples

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315712390.ch36
Laura Stoker, Andrew McCall
Published online on: 26 Sep 2017

36

THE QUEST FOR REPRESENTATIVE SURVEY SAMPLES

Laura Stoker and Andrew McCall

Survey research is in the midst of an era of extraordinary new developments and challenges. People have increasingly moved away from having landline telephones to having mobile phones or no phone at all, which poses new challenges for all aspects of telephone surveying. Response rates continue to deteriorate for face-to-face (FTF) and – especially – telephone surveys. Web-based surveying is booming, with platforms proliferating and the use of nonprobability samples on the rise.

An enormous literature on survey methods has arisen in response to these and other developments. The searchable, online bibliography created by Websm.org (www.websm.org/), which compiles materials on survey methods, contains nearly 5,000 entries from the past ten years alone. Efforts to generalize about developments in the field are hindered by the vastness of this literature and by the fact that technological, social, and political conditions bearing on survey research are rapidly changing and variable across nations. It is often not clear whether findings from studies carried out a decade ago still apply to today, nor whether findings from studies of one country are germane to another.

This chapter considers developments related to the quest for representative survey samples. We begin by discussing the decline in survey response rates, which has prompted new thinking about how response rates relate to sample bias, how to design a survey so as to minimize sample bias, what to use in lieu of the response rate as an index of the representativeness of a survey, and how to construct an optimal weighting scheme. We then briefly review the evolution of thinking about the value of nonprobability survey samples and conclude by drawing attention to two issues worthy of further research.

Declining response rates and their consequences

Survey response rates have continued their now decades-long pattern of decline. Most studies documenting response rate trends have focused on US-based surveys, but similar trends have been found worldwide (Groves 2011). A recent overview estimates the rate of decline to be three times larger for telephone than for FTF surveys in the US (Tourangeau and Plewes 2013). Whereas the response rates for the FTF American National Election Studies were at or above 70 percent for studies from 1980–1992, they dropped steadily to a low of 49 percent by 2012. The telephone-based US Survey of Consumer Attitudes showed even more dramatic response rate declines, from 72 percent in 1979 (Curtin, Presser, and Singer 2005) to 16 percent in 2013 (Dutwin and Lavrakas 2016). Over a more recent period (1997 to 2012), response rates for a "typical" Pew Research Center (Pew) telephone survey fell from 36 percent to 9 percent, while those for a "high effort" survey dropped from 61 percent to 22 percent (Kohut et al. 2012). The substitution of mobile phones for landlines is partially fueling these trends, as contact and non-response rates are lower for respondents contacted by mobile phones (Kohut et al. 2012). Yet, response rates are dropping for both landline and mobile phone samples (Dutwin and Lavrakas 2015, see also Brick and Williams 2013). Government-sponsored surveys in the US have managed to maintain high response rates, though they too have evidenced modest declines. Panel attrition rates appear to be holding steady (Schoeni et al. 2013).

Alarm over declining response rates has been tempered by research showing that the extent of bias in survey estimates is at best weakly related to the survey response rate. Some studies have followed one or more survey projects over time, examining whether the extent of bias grew as the response rate declined (see, for example, Kohut et al. 2012), while others have compared contemporaneous surveys to see if the extent of bias in a given survey can be predicted by its rate of response (see, for example, Groves 2006). In both cases, the answer is essentially no. Still other studies have experimentally varied respondent incentives (see, for example, Martin, Helmschrott, and Rammstedt 2014) or fieldwork strategies (see, for example, Groves and Peytcheva 2008) and showed that such variations have significant effects on response rates but insignificant effects on sample bias.

Furthermore, evidence is accumulating that the bias is often minimal even when response rates are low, especially if survey estimates are adjusted for non-response through effective weighting techniques (see, for example, Holbrook, Krosnick, and Pfent 2008; Keeter et al. 2000; Kohut et al. 2012). That said, it is also typical to find a large degree of variability across survey measures in the extent of bias that is evident. For example, weighted results from a Pew survey with a 9 percent response rate matched the benchmark Current Population Survey (CPS) data for voter registration, but exceeded the CPS results by 21 percent and 28 percent for the percentage contacting public officials and reporting volunteer work, respectively (Kohut et al. 2012).

To make sense of these findings, researchers have turned to formal models of survey non-response that were initially developed in the mid-1970s (Groves 2006; Bethlehem 2010; Bethlehem and Biffignandi 2011; Bethlehem, Cobben, and Schouten 2011). Most influential here is the stochastic non-response model, which depicts each potential respondent as having a latent probability, P, of responding to the survey. This model shows that bias in the sample mean of a given variable, Y, is influenced by three factors related to P. First of all, bias increases with the degree of correlation between P and the survey response itself (Y) – that is, with the extent to which people's likelihood of responding to the survey is correlated with the responses they would provide. Bias disappears if this correlation is zero. Otherwise, bias in the sample mean of Y grows as the mean of P declines – that is, as the overall response rate declines – and as the variance of P increases – that is, as the probability of responding becomes more variable within the population one is sampling from. The worst case scenario arises when potential respondents vary dramatically in their propensity to respond, when that propensity to respond is strongly correlated with the variable(s) of interest, and when the overall rate of response is low. Figure 36.1 illustrates how the magnitude of the bias depends upon all three factors considered together.[1]

The Y axis in Figure 36.1 is the extent of the bias in the sample mean of Y assuming that the standard deviation (SD) of Y is 20. Y could be thought of as a 0–100 feeling thermometer rating scale, where SDs in the range of 20 points are common. The X axis shows response rates ($\bar{P}$) varying from a low of 10 percent to a high of 95 percent. The four sets of results in the figure

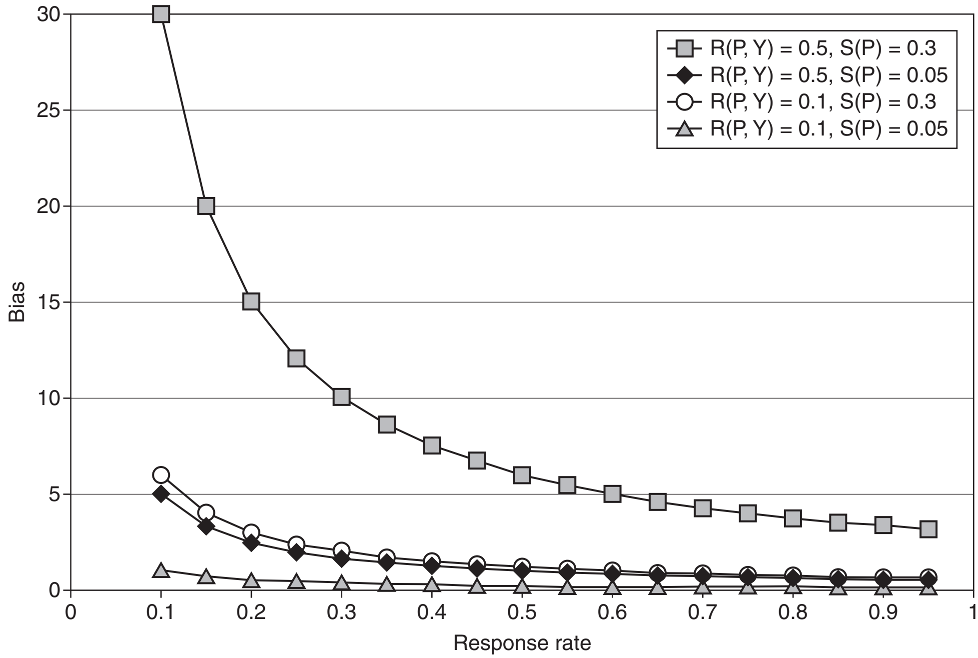

Figure 36.1 Bias in the sample mean as a function of non-response patterns

vary the correlation between Y and the probability of a response, P (either low at 0.10 or high at 0.50), and the SD of P (either low at 0.05 or high at 0.30).[2]

Two main conclusions can be drawn from Figure 36.1. First, bias is much worse when one faces the combination of a high correlation between P and Y and high variation in P. Even if that correlation is high ($R(P, Y) = 0.50$), bias will be minor so long as people do not vary dramatically in their propensity to respond ($S(P) = 0.05$) – even when the response rate is poor. Likewise, even if people vary dramatically in their propensity to respond ($S(P) = 0.30$), bias will be minor so long as the correlation between Y and P is low ($R(P, Y) = 0.10$) – again, even when the response rate is poor. In either of these scenarios, bias only begins to exceed 2 points on the 0–100 scale when the response rate dips below 30 percent. Second, the effect of declining response rates for bias in the estimation of population means is decidedly nonlinear, with the most serious problems arising when response rates become very low. For example, in each of the scenarios the effect on bias as the response rate drops from 30 percent to 10 percent is 3.5 times greater than the effect as the response rate drops from 70 percent to 30 percent.

Importantly, the stochastic non-response model applies just as well to self-selected survey samples as it does to probability survey samples. Although in the former there is no design-based probability of selection, each individual in the population of interest can still be thought of as having a given propensity to participate. Bias in opt-in surveys will be accentuated when those propensities are highly variable, highly correlated with target variables, and the overall rate of response is low. The model can also be elaborated to distinguish between bias that comes from population units being systematically underrepresented within the sample frame ("undercoverage") and bias due to non-response. As discussed by Bethlehem (2010), Callegaro, Manfreda and Vehovar (2015), Groves (2006), and Singer and Ye (2010), among others, these can work in complementary as well as contradictory directions to affect the overall probability of a response and its correlation with target variables. For example, older people are less likely than younger people to have internet access, but response rates among those who do have internet access are higher among the old than the young. This diminishes the extent of bias in web surveys for age and its correlates. On the other hand, bias in socioeconomic status (SES) and its correlates is accentuated by the fact that people with low SES are both less likely to have internet access and to respond to a survey if they do.

It is also worth emphasizing that a high correlation between *P* and *Y* will do more than bias sample estimates of population means. It can also affect inferences about associations between variables. When *Y* refers to a dependent variable, that correlation indexes the degree of what Heckman (1979) and a vast subsequent literature refers to as sample selection bias. Bivariate associations (e.g., correlations, regression coefficients) relating *Y* to any *X* will be biased toward zero, and multivariate associations will be biased in effectively unpredictable ways. The survey research literature does make this connection though infrequently and usually then in discussions about weighting (see, for example, Brick 2013; Winship and Radbill 1994). Designers as well as users of survey data need to worry about sample selection bias if they are estimating relationships among variables even if they are unconcerned about the accuracy of the sample means.

At least three developments in the field of survey research have been fueled by the emergence of a clearer understanding of how response rates relate to sample bias – new thinking about how to design the survey so as to minimize sample bias, about what to use in lieu of the response rate as an index of the representativeness of a survey, and about the development of optimal weighting schemes.

Survey researchers have long sought to design surveys so as to optimize the response rate conditional on a budget constraint, thinking that as the response rate increased, so too would the representativeness of the sample. Researchers now realize that efforts to improve response rates can be ineffective or even backfire if they increase the variance of *P* or increase the correlation of *P* and *Y*. Groves and Heeringa (2006: 448–451) describe a study – probably not atypical – in which the procedures used in the last phase of fieldwork led interviewers to focus their efforts on completing interviews with people who were judged to have a high probability of responding. These procedures were a cost-effective way to maximize the response rate that the study could achieve. However, it is likely that such procedures would have increased the variance of *P* and the correlation of *P* with any number of *Ys*, at least when compared to procedures that would have attempted to improve response rates among those with a lower propensity to respond.

Survey researchers are now increasingly aiming to develop fieldwork procedures that follow "adaptive" (Luiten and Schouten 2013; Schouten, Calinescu, and Luiten 2013) and "responsive" (Groves and Heeringa 2006; Särndal 2011) design principles. Adaptive design makes use of data available ahead of time to set forth fieldwork strategies that will enhance representativeness. Responsive design breaks the fieldwork period into phases, using data from earlier phases to select fieldwork procedures for later phases, with the same objective in mind. For example, data from initial fieldwork can yield estimates of *P* and proxies for *Ys*, which can be used to determine bias-minimizing procedures for the phases to come. Researchers are also looking anew at procedures developed to increase response rates, such as the use of mixed mode designs (see, for example, Couper 2011; Stern, Bilgen, and Dillman 2014), interviewer incentives (see, for example, Peytchev et al. 2010) and respondent incentives (see, for example, Pforr et al. 2015; Singer and Ye 2013), among others (Kreuter 2013). Research has consistently shown, for example, that respondent incentives increase response rates, with the effects of money greater than those of gifts or lotteries and increasing with the size of the payment, and with the effects of prepaid incentives greater than those of contingent incentives. Now the urgent questions concern whether and which procedures reduce non-response bias: Do incentives have larger effects on those already predisposed to respond? Is non-response bias reduced by providing larger incentives for those predisposed against responding?[3]

Efforts are also underway to develop new measures for indexing the quality or representativeness of a survey sample, which thus far includes R-indicators, H-indicators, and the FMI Index. The R-indicator is the standard deviation of $\hat{P}$, as estimated using data on both non-respondents and respondents for a full set of covariates, while the Partial R-indicator is the same but estimated using only a subset of the covariates (Schlomo, Skinner, and Schouten 2012; Schouten, Cobben, and Bethlehem 2009). The R-indicator is proposed as a measure of overall survey quality, while the Partial R-indicator can be used to determine which covariates are especially relevant to the probability of a response. One problem is that neither R-indicator incorporates data on target variables, *Ys*, and thus neither considers the extent to which *P* and *Y* are correlated. Särndal and Lundström's (2010) H3 is comparable to the R-indicator in that it aims to index the representativeness of a survey, overall, using data on auxiliary variables for respondents and non-respondents but not data on *Y*. However, their H1 measure is designed to indicate the sample quality for a given target variable, *Y*, as is the FMI measure developed by Wagner (2012). This is a new and quickly evolving area of research, and before long we will likely see new indices of survey quality that consider *P* as well as many target variables (*Ys*).

An additional stream of burgeoning research concerns the optimal weighting scheme to use in order to adjust for sample bias. Using weights will eliminate (reduce) bias if there is no (less) variation in the propensity to respond within each cell of the weighting scheme, which eliminates (reduces) any correlation between *P* and *Y* within each cell and in the weighted analysis. Since weighting can and usually will increase standard errors (SEs), it is not helpful to use auxiliary variables for weighting that are uncorrelated with *Y*. Indeed, the optimal auxiliary variables for weighting will be strongly correlated with both *P* and *Y* (Little and Vartivarian 2005). It has become increasingly clear that post-stratification weights based on demographics are inferior to weights using a broader set of auxiliary variables and formed through a two-step procedure: (1) weighting (or adjusting design weights) for non-response, and (2) then adjusting through calibration/raking (Kolenikov 2016; Krueger and West 2014; Rota and Laitila 2015).

The traditional method for developing non-response weights estimates the probability of a response (*P*) using variables that are available for both respondents and non-respondents. One thrust of the recent survey research literature is the importance of gathering as much information as possible on non-respondents so as to improve the non-response adjustments (Kreuter et al. 2010; Krueger and West 2014; Olson 2013).

A more recently developed and perhaps more promising alternative makes use of gold-standard reference samples – high-quality samples thought to yield accurate population benchmark estimates, like the CPS or the American Community Survey in the US – for non-response weighting. This method estimates *P* using variables that are common to both samples. Since this "propensity score weighting" method does not require data on non-respondents, it can be used for weighting nonprobability samples. A second virtue is that a potentially much wider set of variables can be utilized when building the non-response model, including attitudinal and behavioral measures in addition to demographics and paradata, though of course the relevant questions must have been asked on each survey. However, there is nothing approaching a consensus as to what questions are essential to include. Studies evaluating the method have differed widely in the variables they incorporate (Berrens et al. 2003; Duffy et al. 2005; Lee 2006; Lee and Valliant 2009; Loosveldt and Sonck 2008; Schonlau et al. 2009). For example, Schonlau et al. (2009) used race, gender, age, income, self-assessed general health, and home ownership, while Berrens et al. (2003) used questions on trust in government, personal efficacy, whether the respondent owned a retirement account, and whether the respondent had read a book, traveled, or participated in a sport recently. As Mick Couper (2013) has argued, building a richer understanding of how survey respondents differ from non-respondents remains one of the most important challenges to be confronted in the decades ahead.

Finally, the data from most major surveys are released along with a single set of survey weights. However, it is clear that a single set of weights will be of limited value in reducing sample bias compared to weights constructed to be optimal for a given *Y* or set of *Ys*. Optimal weights would be constructed using non-response covariates strongly related to the *Y*(*s*) in question and take into account item non-response as well as unit non-response. Efforts to develop such optimal weights and use them in estimation are underway (Caughey and Wang 2014; Andridge and Little 2011; Särndal and Lundström 2010).

Nonprobability survey samples

Writing in 1999, the prominent American pollster Walter Mitofsky condemned internet polling based on nonprobability samples, stating that "the willingness to discard the use of sampling frames as a means of selecting a sample and then the feeble attempts at manipulating the resulting bias … undermine the credibility of the survey process" (Mitofsky 1999: 26). The current consensus remains skeptical of the value of nonprobability samples for population-based inferences, though puts it less stridently. The oft-quoted 2010 Report on Online Panels from the American Association for Public Opinion Research (AAPOR) concluded:

> Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values. There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population.
>
> *(AAPOR Standards Committee 2010: 758)*

Similar conclusions are presented in a recent report to the US National Science Foundation (Krosnick et al. 2015) and in an important new volume on internet panel research (Callegaro et al. 2014). Major news organizations in the US, such as the Associated Press and the *New York Times*, generally limit their reporting of surveys to those based on probability samples.[4]

Yet, that consensus is, in certain respects, starting to give way. A first development is signaled by a change in AAPOR's guidelines about efforts to make population-based inferences from self-selected samples. The initial position taken by AAPOR was that any reporting of SEs when working with surveys using self-selected samples was "misleading" and that researchers working with opt-in panels should use the following wording when describing their study: "Because the sample is based on those who initially self-selected for participation [in the panel] rather than a probability sample, no estimates of sampling error can be calculated."[5] However, in 2015 AAPOR revised its Code of Professional Ethics and Practices to allow for the reporting of SEs, "provided that the measures are accompanied by a detailed description of how the underlying model was specified, its assumptions validated and the measure(s) calculated," suggesting a variety of possible methods for estimating SEs, including resampling techniques, Bayesian Credibility Intervals, and Taylor Series Linearization.[6] Although research into the use of these techniques for self-selected samples is just getting started, recent studies have advocated jackknife or bootstrap procedures as optimal for SE estimation with nonprobability samples (Isaksson, Lee, and Sweden 2005; Lee and Valliant 2009; Enderle and Münnich 2014). AAPOR's 2013 Report on the Task Force on Nonprobability Sampling provides much more information on the research and thinking that led to this revision (Baker et al. 2013).[7]

Second, scholars are gaining a better appreciation of the diverse methods that can be used to select and weight respondents from opt-in internet panels and how these bear on sample quality. Many firms allow researchers to design a sample to meet demographic quotas on selected variables, sometimes also constructing weights to remedy remaining imbalances. Other firms use sample matching strategies, which require the use of data from gold-standard reference samples. One variant combines data from online panelists and the reference sample using variables common to both to estimate the propensity to fall into one or the other group, then builds the online survey sample so that its distribution of response propensities matches that of the reference sample (Terhanian and Bremer 2012). A second variant, used by YouGov, seeks to find respondents within the online panels that are best matches to respondents within the reference sample, which is more successful if the pool of online panelists is very large (Rivers 2007; Rivers and Bailey 2009). Either way, weights are constructed using the reference sample and other data on the target population in order to mitigate remaining imbalances. As described earlier, these methods frequently use a rich set of matching/weighting variables, including attitudinal and behavioral measures and paradata in addition to demographic data.

Research is accumulating on how these variations matter to sample representativeness. Studies distributing the same questionnaire to respondents from different online panels demonstrate substantial variation across the panels in how well the results match population benchmarks (Yeager et al. 2011; Gittelman et al. 2015; Kennedy et al. 2016). Quota sampling will yield representativeness on the quota cells and calibration will do so for the calibration variables, but these techniques do not eliminate – or, often, even mitigate – bias on other measures (Malhotra and Krosnick 2007; Yeager et al. 2011). Expanding the set of quota variables does not appear to help (Gittelman et al. 2015). Those within any given quota cell who respond to the survey remain markedly different from those who do not. Sample matching strategies, however, have much more promise, especially when combined with propensity score and calibration weighting (Baker et al. 2013; Rivers 2007; Gittelman et al. 2015; Kennedy et al. 2016; Ansolabehere and Schaffner 2014; Ansolabehere and Rivers 2013; Sanders et al. 2007; Vavreck and Iyengar 2011; Simmons and Bobo 2015). Just how effective, of course, depends on the variables used in the matching/weighting – ideally highly correlated with target variables and the propensity to respond (Little and Vartivarian 2005), while also being exogenous to the phenomena being studied (Kennedy et al. 2016; Rivers 2016).

That is not to say that well-designed and weighted samples drawn from opt-in internet panels are equal in quality to well-designed and weighted probability samples. Studies have typically found more sample bias in opt-in surveys than in surveys conducted FTF, by telephone or over the web with respondents selected using probability methods (see reviews in AAPOR Standards Committee 2010; Callegaro et al. 2014; Fieldhouse and Prosser in this volume). However, few of these studies use opt-in survey samples that were developed and weighted using what we now think to be best practices (Baker et al. 2013). Those that do use best practices show more promising results, but still typically find one or more variables for which the opt-in sample is less accurate (see, for example, Ansolabehere and Schaffer 2014; Simmons and Bobo 2015).[8] Similarly, even the best opt-in surveys have on occasion fared worse than probability-based surveys in predicting electoral outcomes. For example, YouGov underestimated Tory support in the 2015 British elections more so than did the British Election Study, conducted FTF. The main reason, according to YouGov, was that their weighted sample of young people (who tend to vote Labour) were more politically engaged than young people in the electorate overall.[9] Voter turnout in the US also tends to be more inflated in Cooperative Congressional Election Study

(CCES) surveys carried out with YouGov samples than it is in data from ANES surveys based on probability sampling and conducted FTF. Both samples have a greater percentage of validated voters than they should, but the percentage of those falsely claiming to have voted is twice as high in the CCES as it is in the ANES (Ansolabehere and Hersh 2012).[10]

Still, the advantages of using probability-based sampling (plus weighting) instead of sample matching (plus weighting) may disappear when response rates for the former dip into the low double or single digits, as Doug Rivers has long argued (see, for example, Rivers 2009, 2016; see also Baker et al. 2013). An important new study conducted by the Pew Research Center suggests as much (Kennedy et al. 2016). Kennedy et al. solicited nine surveys from eight opt-in panel vendors to compare to surveys of Pew's probability-based panel (ATP), using an identical questionnaire. The ATP response rate was in the 3.4–3.7 percent range. The study compared sample means and multivariate coefficients obtained from the various samples to benchmarks from CPS, examined how well coefficients from the online surveys predicted outcomes in the CPS data, and performed a series of analyses to determine the value added by sample selection vs. sample weighting procedures. In almost every analysis, Sample "I" outperformed ATP and the rest of the opt-in samples, and as Rivers (2016) revealed, Sample I was YouGov. The study judged the YouGov survey superior in its sample selection design as well as in its weighting procedures. At the same time, the study showed that all 10 web surveys reported much more political and, especially, civic engagement than did the CPS after weighting, and tended to be especially inaccurate in depicting the characteristics of sample subgroups including Blacks, Hispanics, and the young. The report concluded that Sample I (YouGov) had developed a "better methodology" that produces "a more representative, more accurate national survey than the competition within the online nonprobability space," while also noting that the ATP performance was "mixed" (p. 5). Although just one study, the Kennedy et al. (2016) report will undoubtedly spur a further conversation on the relative merits of probability-based vs. opt-in web survey samples.

Looking forward

Most of the research on sample quality has focused on how an unrepresentative sample will yield a misleading portrait of the target population writ large, without considering how the inferences regarding subpopulations are affected. Yet, the erosion of response rates and proliferation of nonprobability samples may be affecting our inferences concerning some subgroups more than others. If it is hard to encourage certain groups to participate in surveys – e.g., Blacks or Hispanics, young people, the less educated – then it would not be surprising to find that, among such groups, the people who do participate are especially unlike their non-responding counterparts. Results from the 2012 ANES comparison of web and face-to-face surveys are illustrative. The discrepancy in self-reported 2008 turnout between the web sample (from KnowledgePanel, response rate 2 percent) and the face-to-face sample (response rate 49 percent) was almost five times greater for those with a high school degree or less (14.4 percent) than for those with a college degree or more (2.9 percent), after weighting. The comparable figures for self-reported turnout in 2012 were 9.0 percent vs. –1.5 percent. Without denying the importance of overall sample representativeness, future research should delve more deeply into sample biases affecting our understanding of subgroup differences and how best to alter survey design and weighting procedures to mitigate them.

Little attention has also been given to how bias in survey samples affects interpretation of treatment effects from survey experiments. One prong of the existing research has considered how well results from survey experiments carried out on nonprobability samples compare to benchmarks obtained from population-based survey experiments, which has tended to be reassuring (Berinsky, Huber, and Lenz 2012; Mullinix et al. 2015, also see Barabas and Jerit 2010; Hainmueller, Hangartner, and Yamamoto 2015). Other work examines the question of whether survey-experimental data should be weighted, keeping in mind the fact that the size of the weight is inversely related to the magnitude of the estimated probability of a survey response, $\hat{P}$ (Solon, Haider, and Woodlridge 2015; Levin and Sinclair 2016; see also Cole and Stuart 2010; Hartman et al. 2015; and Stuart et al. 2011, who consider clinical trials not surveys). If a treatment effect is homogeneous or has heterogeneity that is unrelated to P, then the sample average treatment effect (SATE) gives an unbiased estimate of the population average treatment effect (PATE) and a weighted analysis is inadvisable as it will only inflate the SEs. If the treatment effect varies with P and the missing data are ignorable after weights are applied (a heroic assumption), then the weighted analysis will also yield an unbiased estimate of PATE, though not if the treatment effect is estimated while controlling for covariates (Solon, Haider, and Wooldridge 2015). Otherwise, SATE will be a biased estimator of PATE.

The practical advice offered by Solon, Haider, and Woodlridge (2015) is to examine whether the treatment effect is heterogeneous with respect to the weights by analyzing the data with and without weighting, and then to report and discuss both sets of results. It may, however, be more transparent and informative for researchers to simply display and discuss how estimated treatment effects vary with $\hat{P}$. Since P can be estimated for virtually any sample using the propensity score weighting techniques described earlier, even those working with convenience samples should be able to bring this evidence to bear on the generalizability of their results. Levin and Sinclair (2016) and Hartman et al. (2015) go further, proposing techniques for directly estimating the population average treatment effect on the treated (PATT). Levin and Sinclair show how three different techniques for matching treatment and control subjects can be modified to incorporate survey weights. Hartman et al. recommend estimating treatment effects within matched groups of treatment and control subjects, and then estimating PATT by aggregating these effect sizes weighted by the population proportion for each matched set. As this literature develops and starts to affect standards of practice, it should yield a much richer sense of how the quest for representative samples bears on the goal of generalizing sample-based causal inferences.

Notes

1 Bethlehem (2010: 172–173) shows that the bias is approximately $[R(P, Y)\star S(P)\star S(Y)] \div \bar{P}$, the formula used to create Figure 36.1.

2 A SD of 0.05 for P approximates the situation where the distribution is highly skewed (e.g., Beta distribution parameters of 0.2 and 8), while the SD of 0.30 for P approximates the situation where the distribution is uniform or even modestly U-shaped (Beta parameters of 1 and 1, or 0.2 and 0.7).

3 Efforts to reduce non-response bias could also end up increasing measurement error. Recent studies have found that people with a low propensity to respond tend to provide poorer quality data when they do respond, e.g., more item non-response, less consistent responses, more straightlining, and less information in response to open-ended questions (Fricker and Tourangeau 2010; Dahlhamer 2012; Roberts, Allum, and Sturgis 2014).

4 See http://graphics8.nytimes.com/packages/pdf/politics/20110511quick_checklist.pdf for the NYT and http://commonsensej.blogspot.com/2007/11/ap-style-other-recent-updates.html for the AP.

5 www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Opt-In-Surveys-and-Margin-of-Error.aspx.

6 www.aapor.org/getattachment/Education-Resources/For-Researchers/AAPOR_Guidance_Nonprob_Precision_042216.pdf.aspx. See, also, the 2014 joint ESOMAR/WAPOR Guideline on Opinion Polls and Published Surveys, available at http://wapor.org/esomarwapor-guide-to-opinion-polls/.

7 AAPOR still, however, uses the acronym SLOP for "self-selected opinion polls" in the section of their website on "Bad Samples" (www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Bad-Samples.aspx).

8 Mode differences can arise due to differences in sample quality or in response quality. It is often difficult to determine which cause is producing differences. The literature has, however, consistently demonstrated less social desirability bias in web surveys than in those conducted FTF or by telephone, with recent studies also demonstrating that web surveys elicit more negative attitudes (see, for example, Klausch, Hox, and Schouten 2013; Pew 2015; Ye, Fulton, and Tourangeau 2011; Tourangeau and Yan 2007). There is also fairly strong evidence that web surveys elicit more item non-response, speeding, and straightlining (see, for example, Heerwegh and Loosveldt 2008). See Fieldhouse and Prosser (in this volume) for further details.

9 https://yougov.co.uk/news/2015/05/08/general-election-opinion-polls-brief-post-mortem/ and https://yougov.co.uk/news/2015/12/07/analysis-what-went-wrong-our-ge15-polling-and-what/.

10 Since social desirability bias has consistently been found to be lower in self-administered surveys than in FTF surveys, this high rate of misreporting in the CCES is likely tied to sample composition.

References

AAPOR Standards Committee (2010) "AAPOR Report on Online Panels," *Public Opinion Quarterly*, vol. 74, no. 4, Winter, 711–781.

Andridge, R. R. and Little, R. J. (2011) "Proxy Pattern-Mixture Analysis for Survey Nonresponse," *Journal of Official Statistics*, vol. 27, no. 2, June, 153–180.

Ansolabehere, S. and Hersh, E. (2012) "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate," *Political Analysis*, vol. 20, no. 4, Autumn, 437–459.

Ansolabehere, S. and Rivers, D. (2013) "Cooperative Survey Research," *Annual Review of Political Science*, vol. 16, May, 307–329.

Ansolabehere, S. and Schaffner, B. F. (2014) "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison," *Political Analysis*, vol. 22, no. 3, Summer, 285–303.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013) "Summary Report of the AAPOR Task Force on Nonprobability Sampling," *Journal of Survey Statistics and Methodology*, vol. 1, no. 2, November, 90–143.

Barabas, J. and Jerit, J. (2010) "Are Survey Experiments Externally Valid?," *American Political Science Review*, vol. 104, no. 2, May, 226–242.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012) "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, vol. 20, no. 3, Summer, 351–368.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., and Weimer, D. L. (2003) "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples," *Political Analysis*, vol. 11, no. 1, Winter, 1–22.

Bethlehem, J. (2010) "Selection Bias in Web Surveys," *International Statistical Review*, vol. 78, no. 2, August, 161–188.

Bethlehem, J. and Biffignandi, S. (2011) *Handbook of Web Surveys*, Volume 567, London: John Wiley & Sons.

Bethlehem, J., Cobben, F., and Schouten, B. (2011) *Handbook of Nonresponse in Household Surveys*, Volume 568, London: John Wiley & Sons.

Brick, J. M. (2013) "Unit Nonresponse and Weighting Adjustments: A Critical Review," *Journal of Official Statistics*, vol. 29, no. 3, June, 329–353.

Brick, J. M. and Williams, D. (2013) "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys," *The ANNALS of the American Academy of Political and Social Science*, vol. 645, no. 1, January, 36–59.

Callegaro, M., Baker, R., Bethlehem, J., Goritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (eds.) (2014) *Online Panel Research: A Data Quality Perspective*, Chichester: John Wiley & Sons Ltd.

Callegaro, M., Manfreda, K. L., and Vehovar, V. (2015) *Web Survey Methodology*, Los Angeles: Sage.

Caughey, D. and Wang, M. (2014) "Bayesian Population Interpolation and Lasso-Based Target Selection in Survey Weighting," Paper presented at the 2014 Summer Meeting of the Society for Political Methodology, University of Georgia, July 24–26.

Cole, S. R. and Stuart, E. A. (2010) "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial," *American Journal of Epidemiology*, vol. 172, no. 1, July, 107–115.

Couper, M. P. (2011) "The Future of Modes of Data Collection," *Public Opinion Quarterly*, vol. 75, no. 5, 889–908.

Couper, M. P. (2013) "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys," *Survey Research Methods*, vol. 7, no. 3, 145–156.

Curtin, R., Presser, S., and Singer, E. (2005) "Changes in Telephone Survey Nonresponse Over the Past Quarter Century," *Public Opinion Quarterly*, vol. 69, no. 1, Spring, 87–98.

Dahlhamer, J. M. (2012) "The Intersection of Response Propensity and Data Quality in the National Health Interview Survey (NHIS)," Paper presented at the 2012 Joint Statistical Meeting of the American Statistical Association, San Diego, July 28–August 2.

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005) "Comparing Data from Online and Face-To-Face Surveys," *International Journal of Market Research*, vol. 47, no. 6, 615–639.

Dutwin, D. and Lavrakas, P. (2016) "Trends in Telephone Outcomes, 2008–2015," *Survey Practice*, vol. 9, no. 3, 1–9.

Enderle, T. and Münnich, R. (2014) "Accuracy of Estimates in Access Panel Based Surveys," in Engel, U., Jann, B., Lynn, P., Scherpenzeel, A. and Sturgis, P. (eds.) *Improving Survey Methods: Lessons from Recent Research*, Abingdon: Routledge: 250–265.

Fricker, S. and Tourangeau, R. (2010) "Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys," *Public Opinion Quarterly*, vol. 74, no. 5, December, 934–955.

Gittelman, S. H., Thomas, R. K., Lavrakas, P. J., and Lange, V. (2015) "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples," *Journal of Advertising Research*, vol. 55, no. 4, December, 368–379.

Groves, R. M. (2006) "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, vol. 70, no. 5, December, 646–675.

Groves, R. M. (2011) "Three Eras of Survey Research," *Public Opinion Quarterly*, vol. 75, no. 5, December, 861–871.

Groves, R. M. and Heeringa, S. G. (2006) "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 169, no. 3, July, 439–457.

Groves, R. M. and Peytcheva, E. (2008) "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly*, vol. 72, no. 2, Summer, 167–189.

Hainmueller, J., Hangartner, D., and Yamamoto, T. (2015) "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, February, 2395–2400.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015) "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 178, no. 3, June, 757–778.

Heckman, J. J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica*, vol. 47, no. 1, January, 153–161.

Heerwegh, D. and Loosveldt, G. (2008) "Face-To-Face Versus Web Surveying in a High-Internet-Coverage Population Differences in Response Quality," *Public Opinion Quarterly*, vol. 72, no. 5, December, 836–846.

Holbrook, A., Krosnick, J. A., and Pfent, A. (2008) "The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms," in Lepkowski, J. M., Tucker, C., Brick, J. M., deLeeuw, E. D., Japec, L., Lavrakas, P. J., Link, M. W., and Sangster, R. L. (eds.) *Advances in Telephone Survey Methodology*, Hoboken: John Wiley & Sons: 499–528.

Isaksson, A., Lee, S., and Sweden, S. (2005) "Simple Approaches to Estimating the Variance of the Propensity Score Weighted Estimator Applied on Volunteer Panel Web Survey Data – a Comparative Study," *SRMS Proceedings*, 3143–3149.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., and Presser, S. (2000) "Consequences of Reducing Nonresponse in a National Telephone Survey," *Public Opinion Quarterly*, vol. 64, no. 2, Summer, 125–148.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016) *Evaluating Online Nonprobability Surveys*, Washington, DC: Pew Research Center, available online: www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/ [accessed July 25, 2016].

Klausch, T., Hox, J. J., and Schouten, B. (2013) "Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions," *Sociological Methods and Research*, vol. 42, no. 3, August, 227–263.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., and Christian, L. (2012) "Assessing the Representativeness of Public Opinion Surveys," Paper presented at the 2012 Annual AAPOR Conference, Orlando, May 17–20.

Kolenikov, S. (2016) "Post-Stratification or Non-Response Adjustment?," *Survey Practice*, vol. 9, no. 3, 1–12.

Kreuter, F. (2013) "Facing the Nonresponse Challenge," *The ANNALS of the American Academy of Political and Social Science*, vol. 645, no. 1, January, 23–35.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010) "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 173, no. 2, April, 389–407.

Krosnick, J. A., Presser, S., Fealing, K. H., Ruggles, S., and Vannette, D. (2015) *The Future of Survey Research: Challenges and Opportunities*, Stanford University: The National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences Subcommittee on Advancing SBE Survey Research, available online: www.nsf.gov/sbe/AC_Materials/The_Future_of_Survey_Research.pdf [accessed March 11, 2016].

Krueger, B. S. and West, B. T. (2014) "Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments," *Public Opinion Quarterly*, vol. 78, no. 4, Winter, 795–831.

Lee, S. (2006) "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys," *Journal of Official Statistics*, vol. 22, no. 2, June, 329–349.

Lee, S. and Valliant, R. (2009) "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment," *Sociological Methods and Research*, vol. 37, no. 3, February, 319–343.

Levin, I. and Sinclair, B. (2016) "Causal Inference with Complex Survey Designs," in Atkeson, L. R. and Alvarez, M. (eds.) *The Oxford Handbook of Polling and Polling Methods*, Oxford: Oxford University Press, available online: www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190213299.001.0001/oxfordhb-9780190213299-e-4.

Little, R. J. and Vartivarian, S. (2005) "Does Weighting for Nonresponse Increase the Variance of Survey Means?," *Survey Methodology*, vol. 31, no. 2, 161–168.

Loosveldt, G. and Sonck, N. (2008) "An Evaluation of the Weighting Procedures for an Online Access Panel Survey," *Survey Research Methods*, vol. 2, no. 2, 93–105.

Luiten, A. and Schouten, B. (2013) "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 176, no. 1, January, 169–189.

Malhotra, N. and Krosnick, J. A. (2007) "The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples," *Political Analysis*, vol. 15, no. 3, Summer, 286–323.

Martin, S., Helmschrott, S., and Rammstedt, B. (2014) "The Use of Respondent Incentives in PIAAC: The Field Test Experiment in Germany," *Methods, Data, Analyses*, vol. 8, no. 2, 223–242.

Mitofsky, W. J. (1999) "Pollsters.com," *Public Perspective*, vol. 10, no. 4, July, 24–26.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., and Freese, J. (2015) "The Generalizability of Survey Experiments," *Journal of Experimental Political Science*, vol. 2, no. 2, January, 109–138.

Olson, K. (2013) "Paradata for Nonresponse Adjustment," *The ANNALS of the American Academy of Political and Social Science*, vol. 645, no. 1, January, 142–170.

Pew Research Center (2015) *From Telephone to the Web: The Challenge of Mode of Interview Effects in Public Opinion Polls*, Washington, DC: Pew Research Center, available online: www.pewresearch.org/2015/05/13/from-telephone-to-the-web-the-challenge-of-mode-of-interview-effects-in-public-opinion-polls/ [accessed July 28, 2016].

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010) "Reduction of Nonresponse Bias in Surveys Through Case Prioritization," *Survey Research Methods*, vol. 4, no. 1, 21–29.

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Frässdorf, M., Hajek, K., Helmschrott, S., Kleinert, C., Koch, A., and Krieger, U. (2015) "Are Incentive Effects on Response Rates and Nonresponse Bias in Large-Scale, Face-To-Face Surveys Generalizable to Germany? Evidence from Ten Experiments," *Public Opinion Quarterly*, vol. 79, no. 3, Autumn, 740–768.

Rivers, D. (2007) "Sampling for Web Surveys," Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, July 29–August 2.

Rivers, D. (2009) "Second Thoughts About Internet Surveys," *The Huffington Post*, June 9, 2009, available online: www.huffingtonpost.com/guest-pollster/doug_rivers_b_724621.html.

Rivers, D. (2016) "Pew Research: YouGov Consistently Outperforms Competitors on Accuracy," *YouGov*, May 13, 2016, available online: https://today.yougov.com/news/2016/05/13/pew-research-yougov/.

Rivers, D. and Bailey, D. (2009) "Inference from Matched Samples in the 2008 US National Elections," Paper presented at the 2009 Joint Statistical Meetings, Washington, DC, August 1–6.

Roberts, C., Allum, N., and Sturgis, P. (2014) "Nonresponse and Measurement Error in an Online Panel," in Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. J., Krosnick, J. A., and Lavrakas, P. J. (eds.) *Online Panel Research: A Data Quality Perspective*, London: Wiley: 337–362.

Rota, B. J. and Laitila, T. (2015) "Comparisons of Some Weighting Methods for Nonresponse Adjustment," *Lithuanian Journal of Statistics*, vol. 54, no. 1, 69–83.

Sanders, D., Clarke, H. D., Stewart, M. C., and Whiteley, P. (2007) "Does Mode Matter for Modeling Political Choice? Evidence from the 2005 British Election Study," *Political Analysis*, vol. 15, no. 3, Summer, 257–285.

Särndal, C. E. (2011) "The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation," *Journal of Official Statistics*, vol. 27, no. 1, March, 1–21.

Särndal, C. E. and Lundström, S. (2010) "Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias," *Survey Methodology*, vol. 36, no. 2, 131–144.

Schoeni, R. F., Stafford, F., McGonagle, K. A., and Andreski, P. (2013) "Response Rates in National Panel Surveys," *The ANNALS of the American Academy of Political and Social Science*, vol. 645, no. 1, January, 60–87.

Schonlau, M., Van Soest, A., Kapteyn, A., and Couper, M. (2009) "Selection Bias in Web Surveys and the Use of Propensity Scores," *Sociological Methods and Research*, vol. 37, no. 3, February, 291–318.

Schouten, B., Calinescu, M., and Luiten, A. (2013) "Optimizing Quality of Response Through Adaptive Survey Designs," *Survey Methodology*, vol. 39, no. 1, 29–58.

Schouten, B., Cobben, F., and Bethlehem, J. (2009) "Indicators for the Representativeness of Survey Response," *Survey Methodology*, vol. 35, no. 1, 101–113.

Shlomo, N., Skinner, C. J., and Schouten, B. (2012) "Estimation of an Indicator of the Representativeness of Survey Response," *Journal of Statistical Planning and Inference*, vol. 142, no. 1, January, 201–211.

Simmons, A. D. and Bobo, L. D. (2015) "Can Non-Full-Probability Internet Surveys Yield Useful Data? A Comparison with Full-Probability Face-to-Face Surveys in the Domain of Race and Social Inequality Attitudes," *Sociological Methodology*, vol. 45, no. 1, August, 357–387.

Singer, E. and Ye, C. (2013) "The Use and Effects of Incentives in Surveys," *The ANNALS of the American Academy of Political and Social Science*, vol. 645, no. 1, January, 112–141.

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015) "What are We Weighting for?," *Journal of Human Resources*, vol. 50, no. 2, Spring, 301–316.

Stern, M. J., Bilgen, I., and Dillman, D. A. (2014) "The State of Survey Methodology Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design," *Field Methods*, vol. 26, no. 3, August, 284–301.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 174, no. 2, April, 369–386.

Terhanian, G. and Bremer, J. (2012) "A Smarter Way to Select Respondents for Surveys," *International Journal of Market Research*, vol. 54, no. 6, 751–780.

Tourangeau, R. and Plewes, T. J. (eds.) (2013) *Nonresponse in Social Science Surveys: A Research Agenda*, Washington, DC: National Academies Press.

Tourangeau, R. and Yan, T. (2007) "Sensitive Questions in Surveys," *Psychological Bulletin*, vol. 133, no. 5, September, 859–883.

Vavreck, L. and Iyengar, S. (2011) "The Future of Political Communication Research: Online Panels and Experimentation," in Shapiro, R. Y. and Jacobs, L. J. (eds.) *Oxford Handbook of Public Opinion and the Media*, Oxford: Oxford University Press: 156–168

Wagner, J. (2012) "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias," *Public Opinion Quarterly*, vol. 76, no. 3, Autumn, 555–575.

Winship, C. and Radbill, L. (1994) "Sampling Weights and Regression Analysis," *Sociological Methods and Research*, vol. 23, no. 2, November, 230–257.

Ye, C., Fulton, J., and Tourangeau, R. (2011) "More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice," *Public Opinion Quarterly*, vol. 75, no. 2, Summer, 349–365.

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 Opinion Quarterly*, vol. 75, no. 4, Winter, 709–747.