Best practices in collecting online data with Asian, Black, Latino, and White respondents: evidence from the 2016 Collaborative Multiracial Post-election Survey

Matt A. Barreto, Lorrie Frasure-Yokley, Edward D. Vargas & Janelle Wong

Published online: 04 Jan 2018.

Submit your article to this journal ⬚

View related articles ⬚

View Crossmark data ⬚

Check for updates

Best practices in collecting online data with Asian, Black, Latino, and White respondents: evidence from the 2016 Collaborative Multiracial Post-election Survey

Matt A. Barreto[a], Lorrie Frasure-Yokley[a], Edward D. Vargas[b] and Janelle Wong[c]

[a]Department of Political Science, University of California, Los Angeles, CA, USA; [b]School of Transborder Studies, Arizona State University, Tempe, AZ, USA; [c]American Studies & Asian American Studies, University of Maryland, College Park, MD, USA

ABSTRACT

As the U.S. becomes increasingly diverse, new challenges to capturing this diversity emerge for survey researchers studying political attitudes and behavior. Sampling methods are no longer straightforward as simple random-digit-dial. Given the confluence of changing demographics and changing survey technology, we argue that researchers should carefully consider a stratified listed/density quota-sampling approach to multilingual surveys with large racial/ethnic minority samples. We examine the 2016 Collaborative Multiracial Post-election Survey, which implemented this approach with great success. Our approach resulted in collecting 10,145 completed surveys, in five languages, with large samples of Asian Americans, African-Americans, Latinos, and Whites. We conclude with a set of best practices or principles for online research of racial/ethnic minority populations that we hope will guide future social science research in this domain.

Introduction

As the U.S. becomes increasingly diverse, new challenges to capturing this diversity emerge for survey researchers studying political attitudes and behavior (Berry, Chouhoud, and Junn 2016). Sampling methods are no longer straightforward as simple random-digit-dial. This is because random selection methods often fail to include robust samples of geographically concentrated or relatively small populations, such as Asian Americans, Muslim Americans, or Caribbean Americans (cf., Heeringa et al. 2004; Jackson et al. 2004; Lien, Margaret Conway, and Wong 2004). The U.S. has also become more linguistically diverse, whereby English-only sampling leads to the exclusion of respondents who reside in largely immigrant communities (Lien, Margaret Conway, and Wong 2004; Ngo-Metzger, Legedza, and Phillips 2004; Wong et al. 2011; Barreto and Segura 2017). Moreover, in-person and telephone surveys of low-incidence populations yield few respondents and are cost prohibitive (Ramakrishnan et al. 2012). These issues are compounded by the ever-decreasing numbers of people who are willing to participate in surveys administered via landline telephone or cell phone (Schoeni et al. 2013). Given

---

the confluence of changing demographics and changing survey technology, we argue that researchers should carefully consider a stratified listed/density quota-sampling approach to multilingual surveys with large racial/ethnic minority samples. We examine the 2016 Collaborative Multiracial Post-election Survey (CMPS), which implemented this approach with great success. Our approach resulted in collecting 10,145 completed surveys, in five languages,[1] with large samples of Asian Americans, African-Americans, Latinos, and Whites.

Polling methods and minority populations

Berry, Chouhoud, and Junn (2016) provide an excellent overview of the advantages and disadvantages associated with various methods used to survey low-incidence populations. Such populations are defined as those "who share a common characteristic and make up a relatively small proportion of the broader population" including non-Whites. The methods they compare are (1) stratified designs, (2) list-based selection, and (3) density strategies. The primary disadvantages of these methods are the potential biases related to coverage. For example, some members of racial/ethnic minority groups may not have an "ethnic" name and will be excluded. Others may not live in high-density enclaves and may also be excluded. These exclusions may lead to systematic bias. However, when both listed and density-sampling are considered simultaneously, very few potential subjects are excluded. For example, name-matching will return any potential ethnic respondent, regardless of whether they live in a high-density immigrant enclave or in a majority-White suburban tract. Likewise, density-sampling will return any potential ethnic respondent, regardless if their name is José Sánchez or Bill Richardson. So minority sampling can be most effective when these approaches are combined. Adding stratification can further improve minority samples. To further reduce the possibility of bias, it is important that sampling relies on random selection, within key strata, to ensure that the resulting minority sample is not too heavily acculturated, or too heavily unacculturated. The advantages of these combined approaches are compelling. Listed and density approaches reduce cost, increase response rates, and allow researchers to better anticipate the language of interview (Berry, Chouhoud, and Junn 2016). Because many of the key characteristics of the targeted groups are known through government data sources, such as the U.S. Census and Current Population Surveys, the data from a stratified listed/density-sampling approach can incorporate post-stratification weights to better approximate known population parameters.

When these considerations are not factored in, online data for Blacks, Latinos, and Asian Americans are often unreliable and unusable (Couper 2011; Kennedy et al. 2016). Convenience samples collected from M-Turk, Google consumer, or low-budget panel providers such as Survey Sampling International are very likely to be skewed towards younger, U.S.-born, educated, and highly acculturated minority populations (Chang and Krosnick 2009; Taylor et al. 2009). We very strongly urge caution in drawing any inferences from observational analysis using these sampling approaches. We cannot be certain that the "results" are uniquely present in a convenience and unrepresentative sample or if they can be generalized to the larger population (Hoogendorn and Daalmans 2009; Szolnokin and Hoffmann 2013).

To this very point, the Pew Research Center undertook an analysis of multiple online panel providers to assess the accuracy of their African-American and Latino data. Their findings suggested that most online samples of minorities are unreliable:

> Across the nine nonprobability samples, the average estimated bias on benchmarked items was more than 10 percentage points for both Hispanics (15.1) and blacks (11.3). In addition, the online samples rarely yielded accurate estimates of the marginal effects of being Hispanic or black on substantive outcomes, when controlling for other demographics. These results suggest that researchers using online nonprobability samples are at risk of drawing erroneous conclusions about the effects associated with race and ethnicity. (Kennedy et al. 2016)

The goal of the 2016 CMPS was to include large and generalizable samples of Blacks, Latinos, and Asian Americans to allow for within-group comparison and analysis of an individual racial group or comparative analysis across groups (Jackson et al. 2004). The methodologies employed in the CMPS include both listed and density-sampling, all the while stratifying on key demographic characteristics to ensure a representative sample. We contracted with multiple sampling vendors who have proven expertise in collecting minority data-sets, and rely on lists that identified potential respondents by first, middle, or last name, or provided potential respondents by matching them to areas of high racial or ethnic density, or a combination of the two. In addition to working with sample vendors for our non-registered voter population, we pioneered a new approach to online completes for our registered voter sample that more closely approximates random telephone interviewing. The approach is called random-recruit-to-web (RRW) and has been implemented in mixed-mode phone–web research by political scientists Matt Barreto and Gary Segura in public opinion research among Latino voters (2014). The RRW method uses the official voter file of registered voters, which now contains about 40% of respondents with an email address either volunteered or matched through external databases. From these millions of records of email addresses, we sort them by race/ethnicity, stratified by age, and socioeconomic status (based on census tract). Then we randomly select email addresses to send an invite to take a survey. This is almost the exact approach taken to telephone surveys of registered voters.

The 2016 CMPS: how we built a cooperative of research scholars

In spring 2016, scholars were invited to collaborate on the 2016 CMPS. The goal of the project was to create the first cooperative, 100% user content-driven, multiracial, multieth-nic, multilingual, postelection online survey on race, ethnicity, and politics in the U.S. In full, 86 social scientists across 55 different colleges/universities joined the collaborative to self-fund the survey through the purchase of question content by contributors. A multidisciplin-ary group of researchers at varying stages of their academic careers contributed survey ques-tion content including video/audio stimuli and split-sample experimentation in question wording. Participating contributors included a diverse group of junior and senior faculty from large research institutions, Historically Black Colleges and Universities, Hispanic Ser-vicing Institutions, as well as graduate students and postdoctoral fellows.

The survey's main focus was on attitudes about the 2016 Election and candidates, debates over immigration, policing, and racial equality, and experiences with racial discrimination across many facets of American life. The result of the project was a unique survey that included both a broad range of survey questions emanating from a diversity of disciplines,

including political science, sociology, psychology, public health, American studies, Latino studies, African-American studies, and Asian American studies. Users could submit questions for just one single racial group or common questions across all four racial groups, depending on their interest. In cases where two different users submitted very similar questions, the principal investigators worked to create a single common question.

We worked with each scholar or team of scholars to help them craft the best possible set of questions to ensure both comparability with longstanding studies such as the American National Election Studies (ANES), as well as new and innovative survey content regarding a host of political and policy related questions. About 60% of the question content in the 2016 CMPS is the same as or similar to longstanding survey questions used in the study of political science or the social sciences more broadly. After receiving all content, we formatted the instrument to determine if we needed to add content or further synthesize common content across the four samples. In our attempt to keep the overall survey at reasonable time/minutes, to minimize cost, and to allow for the unique specificity of some of the content purchased, some questions asked to the Latino or Asian sample, were not surveyed of the Black sample. On the other hand, some of the questions asked of the Black sample were not surveyed among the Latino or Asian samples.

## The 2016 CMPS: our innovative methodological approach

A total of 10,145 completed interviews were collected online in a respondent self-administered format from December 3, 2016, to February 15, 2017. The survey (and invitation) was available to respondents in English, Spanish, Chinese (simplified), Chinese (traditional), Korean, and Vietnamese. Because of the primary interest in the 2016 election, the project started with a large sample of registered voters to provide a large sample size for analyses. The data also included an adult sample of non-registered voters as well, including non-citizens. By examining both registered voters and those not on the rolls, studies of self-selection into participation become possible (Table 1).

The full data are weighted within each racial group to match the adult population in the 2015 Census ACS 1-year data file for age, gender, education, nativity, ancestry, and voter registration status. A post-stratification raking algorithm was used to balance each category within ±1% of the ACS estimates. It is very important that each racial group sample is weighted independently so that it matches the census estimates for the specific group, not the sample as a whole. Notably, our sampling methods yielded a fairly representative sample geographically, decreasing the need for costly oversamples, allowing researchers to analyze a large national multiracial sample of both registered and non-registered voters.

Overall, the survey contained 394 questions and median completion time of 43.2 minutes. In total, 298,159 email addresses were selected and sent invitations to participate in the survey and 29,489 people accepted the invitation and started the survey, for an effective response rate of 9.9%. Among the 29,489 people who started the survey, 11,868

Table 1. 2016 CMPS overall sample sizes, by race/ethnicity.

	Total	Latino	Black	Asian	White
Registered to vote	6024	1816	2002	1503	703
Not registered	4121	1187	1100	1503	331
Total	10,145	3003	3102	3006	1034
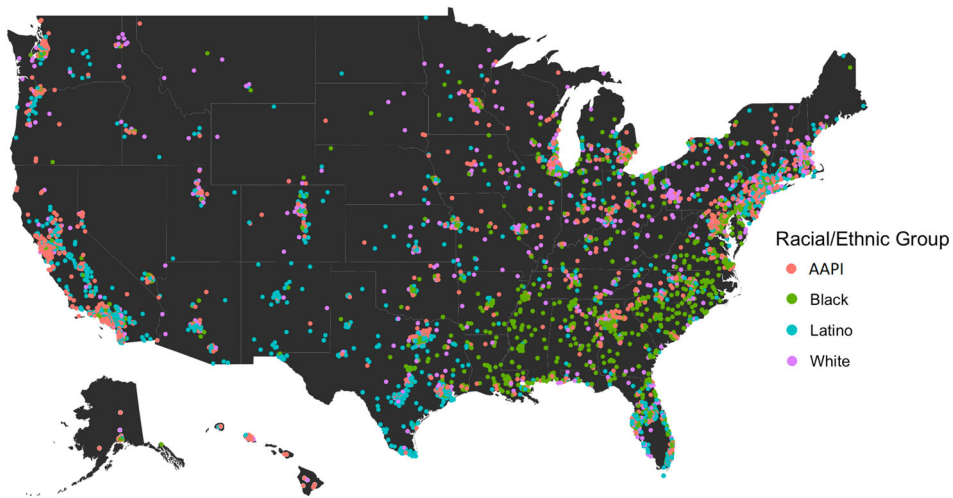
Figure 1. Map of 2016 CMPS respondents by race and geography.

potential respondents were terminated due to quotas being full, which resulted in 17,621 who were eligible to take the survey and of which 10,145 completed the full questionnaire for a cooperation rate of 57.6%. Respondents were given a $10 or $20 gift card as compensation for their participation. Data for registered voters come from the national voter registration database email sample, and respondents were randomly selected to participate in the study and confirmed they were registered to vote before starting the survey. A map of the distribution of our respondents nationally by race can be found in Figures 1 and 2.

For the non-registered sample, email addresses were randomly selected from various online panel vendors. This group was selected from one of six online panels of respondents and confirmed that they were not registered to vote before starting the survey. Programming and data collection for the full project were overseen by Pacific Market Research in Renton, WA.

It is important to note that our registered voter and non-registered respondents come from somewhat different sources. Because it is possible to start with an official list of registered voters, and to obtain validated vote history from the voter file, we used that as a starting point. Research on random digit dialing and self-reported registration and voting status suggests considerable overreporting and false-reporting does occur (Shaw, de la Garza, and Lee 2000; Karp and Brockington 2005), and we avoided that by starting with an actual list of people registered to vote. For our non-registered sample, there was no such "list" to start with, and thus we had to rely on a variety of panel vendors noted above. In this case, it was important that we asked the panel vendors to de-dupe the lists and remove anyone who was a registered voter from their lists before sending out invites to participate in the survey. In the end, for both types of respondents, we apply weights so that they both match the Census universe estimates for age, socioeconomic status, nativity, and more.

Assessing the characteristics and accuracy of the 2016 CMPS

With any data collection project, the final product is only as good as the accuracy of the data. In the next section, we assess the CMPS on two parameters: first, the basic

Latino respondents

Black respondents

AAPI respondents
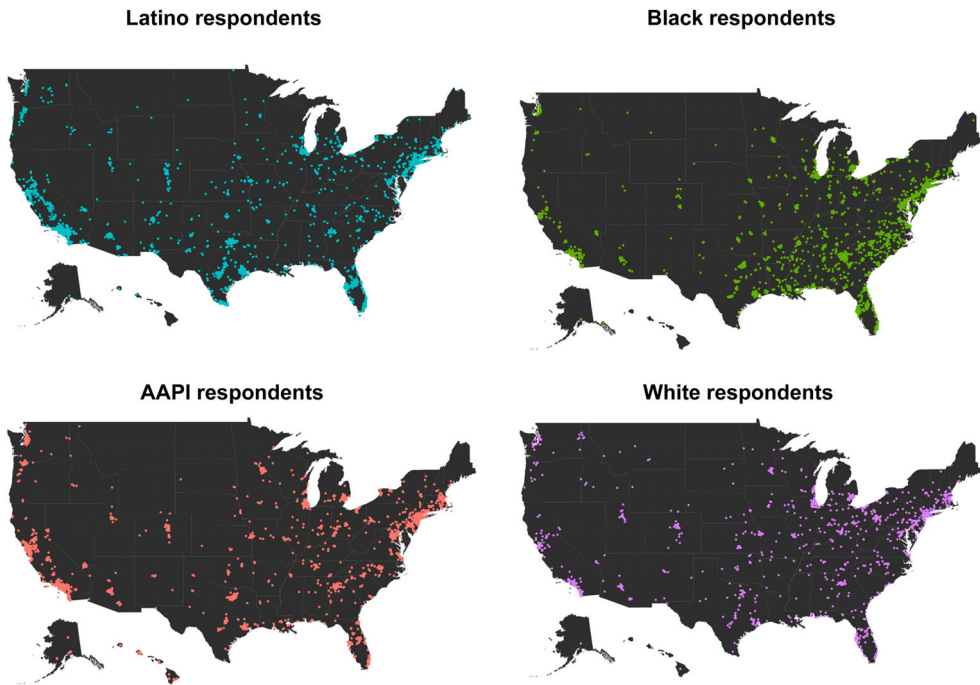
White respondents

Figure 2. Map of 2016 CMPS Respondents by race and geography.

Table 2. Latino demographics and political attitudes.

Sample	2016 CMPS (online) unweighted All adults N = 3003	2016 CMPS (online) All adults N = 3003	2008 CMPS (phone) Registered voters N = 1577	2016 ANES (in-person and online) Eligible voters N = 450	2015 LNHIS (phone and online) All adults N = 1493
Foreign born	29%	46%	42%	33%	52%
U.S. born	71%	53%	58%	65%	47%
HS or less	34%	60%	44%	50%	46%
Some college	35%	25%	37%	19%	22%
College or more	32%	15%	19%	31%	29%
Less than $40k	43%	52%	44%	42%	40%
40k–80k	28%	26%	34%	29%	22%
$80k or greater	20%	19%	22%	23%	16%
Homeowner	38%	34%	50%	47%	n/a
Median age	35	35	52	40	45
Male	32%	48%	49%	52%	46%
Female	68%	52%	51%	46%	54%
Spanish media	43%	52%	53%	n/a	n/a
Linked fate	58%	57%	58%	61%	53%
Obama fav.	78%	81%	81%	60%	n/a
Democrat	48%	51%	51%	49%	44%
Republican	16%	14%	15%	15%	11%
Independent/other	40%	36%	34%	35%	26%

Note: All data are weighted unless specified.

demographic characteristics of the samples, and second on more substantive political attitudes. In Tables 2–4, we compare the 2016 CMPS sample to the 2008 CMPS, the 2012 and 2016 ANES, the 2016 National Asian American Survey (NAAS), 2012 Pew Asian

Table 3. Black demographics and political attitudes.

Sample	2016 CMPS (online) unweighted All adults N = 3102	2016 CMPS (online) All adults N = 3102	2008 CMPS (phone) Registered voters N = 944	2012 ANES (in-person and online) Eligible voters N = 409	2016 ANES (in-person and online) Eligible voters N = 398
Foreign born	6%	12%	9%	8%	9%
U.S. born	94%	88%	90%	92%	90%
HS or less	31%	47%	46%	49%	49%
Some college	37%	33%	32%	33%	68%
College or more	37%	20%	20%	18%	32%
Less than $40k	50%	53%	33%	58%	61%
40k–80k	12%	21%	22%	24%	24%
$80k or greater	27%	16%	14%	18%	16%
Homeowner	34%	33%	62%	44%	35%
Median age	40	40	53	45	45
Male	31%	47%	54%	47%	42%
Female	69%	53%	46%	53%	57%
Linked fate	67%	66%	60%	64%	77%
Obama fav.	92%	92%	96%	94%	82%
Democrat	68%	69%	68%	72%	70%
Republican	4%	5%	2%	3%	6%
Independent/ other	27%	26%	30%	25%	22%

Note: All data are weighted unless specified.

Table 4. Asian American demographics and political attitudes.

Sample	2016 CMPS (online) unweighted All adults (N = 3006)	2016 CMPS (online) All adults (N = 3006)	2016 ANES (in-person and phone) Eligible voters (N = 148)	2012 *Pew Asian American (phone) All adults (N = 3511)	2016 **NAAS Pre-election (phone) All adults (N = 2800)
Foreign born	46%	76%	59%	76%	68%
U.S. born	54%	24%	41%	24%	32%
HS or less	15%	26%	18%	26%	27%
Some college	19%	20%	15%	19%	15%
College or more	66%	54%	66%	54%	58%
Less than $40k	23%	29%	26%	32%	43%**
40k–80k	33%	33%	22%	21%*	15%**
$80k or greater	44%	38%	53%	47%*	48%
Homeowner	52%	49%	66%	56%	61%
Median age	36	40	39	43	55
Male	40%	49%	52%	50%	46%
Female	60%	51%	47%	50%	54%
Asian lang. media	25%	22%	N/A	N/A	23%***
Linked fate	61%	59%	70%	N/A	N/A
Obama fav.	76%	79%	N/A	65%	69%
Democrat	42%	42%	34%	33%	40%
Republican	20%	18%	24%	18%	13%
Independent/other	38%	40%	42%	49%	47%

Notes: All data is weighted unless specified.
Response categories for income differ across surveys: *Pew 2012 is 40–75k and 75k+; **2016 National Asian American Survey (NAAS) is 50k or less; 50–75k; 75k+; ***2016 NAAS: "Most" political information from ethnic media; Columns may not add up due to rounding.

American Survey, and the 2015 Latino National Health and Immigration Survey (LNHIS) for large representative telephone or mixed-mode surveys that contain sizable samples of Asian Americans, Latinos, African-Americans, Whites. Overall, we find very consistent

demographic characteristics of the 2016 CMPS online data-set to these other benchmark data-sets, suggesting that the demographics are accurate. Furthermore, we find considerable similarities in the political attitudes across the sample, providing additional evidence validating the 2016 CMPS underlying data.

Strengths of online survey design

While ideally we would have implemented a design that included both phone and internet modes, we do note here that internet-based surveys have distinct advantages over both phone and in-person methodologies. First, in addition to being much more cost-effective than traditional telephone or in-person survey methods, internet-based surveys do not have to be completed in a single-sitting. Respondents can pause and return to the survey, giving those who are interrupted or run out of time to complete the survey with another chance. Second, an internet-based survey eliminates interviewer effects. Third, internet surveys facilitate a greater range of questions that rely on visual content, such as a photo or video.

Remaining challenges with online survey design

While cost and efficiency are certainly big advantages, online data collection is not yet perfect, and challenges remain. First, there are two populations that we need to continue working towards increasing their population in online surveys: non-English language-dominant respondents and elderly respondents over age 70. While considerable advancements have been made in increasing the number of Spanish-dominant respondents in various Latino panels, the same cannot be said for Asian Americans whose first language is not English or immigrants from other parts of the world. In addition, the digital divide continues to persist among older Americans who either have less access to Internet technology or are simply less comfortable with self-administered computer, tablet, or smartphone surveys.

Best practices are critical

Given these challenges, we conclude with a set of best practices or principles for online research of racial/ethnic minority populations that we hope will guide future social science research in this domain:

Best practices in online surveys of racially diverse populations:

1) Cast a wide net, and let respondents self-identify their race or ethnicity. Do not give respondents any hint of which race or ethnicity you are looking for; always begin your survey and introduction with a generic statement that you are interested in a variety of opinions to ensure that everyone in [geography] has a chance to participate.
2) Use a multilingual invite and introduction so as not to exclude anyone at first contact. Allow respondents to select which language they are most comfortable in to complete the survey.

3) Just as phone and mail surveys require multiple follow-up invites, release your sample in replicates and do multiple recruitment follow-ups to selected sample pieces before you release any new, fresh sample.
4) When using listed samples, it is important to flag respondents by first, middle, or last name. Do not rely on surname-only approaches as it could miss married women or people with non-ethnic surnames, but who carry ethnic first or middle names.
5) Targeting potential racial/ethnic respondents by census block density helps increase your potential subject pool and move beyond a "name-only" approach. Practices #4 and #5 should be used in combination with one another.
6) Stratify and quota-sample[2] by key age, socioeconomic, and ethnic demographic points to ensure your data are balanced. Within each stratum, randomly select respondents.
7) Weight your final data to the best universe estimates within each racial group.

Notes

1. The initial invitation, as well as the first screen was available in multiple languages, and respondents could self-select which language they were most comfortable in completing the survey.
2. We recommend setting quota targets by age, education, gender, and nativity, such that the sample is divided into key stratum, and then respondents are randomly selected. However, each strata has a quota target, with some flexibility. For example, if the 2016 Census American Community Survey (ACS) estimates that 29% of Latino adults are between the ages of 18 and 29, and the researcher is targeting a final sample size of 1000 Latino adults, we recommend setting a quota target of 290 Latino respondents age 18–29, and so on, for each age group. When possible, use the latest Census ACS data as a guide for the quota targets.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Barreto, Matt, and Gary Segura. 2014. *Latino America: How America's Most Dynamic Population is Poised to Transform the Politics of the Nation*. New York, NY: Public Affairs Books.

Barreto, Matt, and Gary Segura. 2017. "Understanding Latino Voting Strength in 2016 and Beyond: Why Culturally Competent Research Matters." *Journal of Cultural Marketing Strategy* 2 (2): 190–201.

Berry, Justin, Youssef Chouhoud, and Jane Junn. 2016. "Reaching Beyond Low-Hanging Fruit," October. doi:10.1093/oxfordhb/9780190213299.013.1.

Chang, L., and J. A. Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet – Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73 (4): 641–678.

Couper, M. P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75 (5): 889–908.

Heeringa, Steven G., James Wagner, Myriam Torres, Naihua Duan, Terry Adams, and Patricia Berglund. 2004. "Sample Designs and Sampling Methods for the Collaborative Psychiatric Epidemiology Studies (CPES)." *International Journal of Methods in Psychiatric Research* 13 (4): 221–240. doi:10.1002/mpr.179.

Hoogendorn, A. W., and J. Daalmans. 2009. "Nonresponse in the Requirement of an Internet Panel Based on Probability Sampling." *Survey Research Methods* 3: 59–72.

Jackson, James S., Myriam Torres, Cleopatra H. Caldwell, Harold W. Neighbors, Randolph M. Nesse, Robert Joseph Taylor, Steven J. Trierweiler, and David R. Williams. 2004. "The National Survey of American Life: A Study of Racial, Ethnic and Cultural Influences on Mental Disorders and Mental Health." *International Journal of Methods in Psychiatric Research* 13 (4): 196–207. doi:10.1002/mpr.177.

Karp, Jeffrey A, and David Brockington. 2005. "Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries." *The Journal of Politics* 67 (3): 825–840.

Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. 2016. Evaluating Online Nonprobability Surveys: Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics. Pew Research Center. May 2.

Lien, Pei-te, M. Margaret Conway, and Janelle Wong. 2004. *The Politics of Asian Americans: Diversity and Community*. New York, NY: Taylor & Francis.

Ngo-Metzger, Q., A. T. R. Legedza, and R. S. Phillips. 2004. "Asian Americans' Reports of Their Health Care Experiences: Results of a National Survey." *Journal of General Internal Medicine* 19 (2): 111–119. doi:10.1111/j.1525-1497.2004.30143.x.

Ramakrishnan, Karthick, Jane Junn, Taeku Lee, and Janelle Wong. 2012. *National Asian American Survey, 2008. ICPSR31481-v2*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. doi:10.3886/ICPSR31481.v2.

Schoeni, Robert F., Frank Stafford, Katherine A. Mcgonagle, and Patricia Andreski. 2013. "Response Rates in National Panel Surveys." *The Annals of the American Academy of Political and Social Science* 645: 60–87. (Jan).

Shaw, Daron, Rodolfo O. de la Garza, and Jongho Lee. 2000. "Examining Latino Turnout in 1996: A Three-State, Validated Survey Approach." *The American Journal of Political Science* 44 (2): 338–346.

Szolnokin, Gergely, and Dieter Hoffmann. 2013. "Online, Face-to-Face and Telephone Surveys – Comparing Different Sampling Methods in Wine Consumer Research." *Wine Economics and Policy* 2: 57–66.

Taylor, P. A., N. M. Nelson, B. D. Grandjean, B. Anatchkova, and D. Aadland. 2009. Mode Effects and Other Potential Biases in Panel-based Internet Surveys Final Report, Wyoming Survey & Analysis Centre (WYSAC, University of Wyoming).

Wong, Janelle, S. Karthick Ramakrishnan, Taeku Lee, and Jane Junn. 2011. *Asian American Political Participation: Emerging Constituents and Their Political Identities*. 1st Edition. New York: Russell Sage Foundation.