Public Opinion Quarterly, Vol. 70, No. 4, Winter 2006, pp. 499–529

AMERICAN PUBLIC OPINION IN THE 1930S AND 1940S
THE ANALYSIS OF QUOTA-CONTROLLED SAMPLE SURVEY DATA

ADAM J. BERINSKY

Abstract The 1930s saw the birth of mass survey research in America. Large public polling companies, such as Gallup and Roper, began surveying the public about a variety of important issues on a monthly basis. These polls contain information on public opinion questions of central importance to political scientists, historians, and policymakers, yet these data have been largely overlooked by modern researchers due to problems arising from the data collection methods. In this article I provide a strategy to properly analyze the public opinion data of the 1930s and 1940s. I first describe the quota-control methods of survey research prevalent during this time. I then detail the problems introduced through the use of quota-control techniques. Next, I describe specific strategies that researchers can employ to ameliorate these problems in data analysis at both the aggregate and individual levels. Finally, I use examples from several pubic opinion studies in the early 1940s to show how the methods of analysis laid out in this article enable us to utilize historical public opinion data.

The study of political behavior has flourished in recent decades with the development of long-term, time series data collected by both academic survey researchers and commercial pollsters. Our understanding of the dynamics of mass opinion and behavior prior to the 1950s has, however, been more limited. Expanding political behavior research to this earlier era is valuable

ADAM J. BERINSKY is an associate professor of political science at the Massachusetts Institute of Technology. For valuable discussion and advice regarding this article, I would like to thank several anonymous reviewers, Steve Ansolabehere, Jake Bowers, Bear Braumoeller, John Brehm, Andrew Gelman, Ben Hansen, Steve Heeringa, Sunshine Hillygus, Doug Kriner, Jim Lepkowski, Tali Mendelberg, Sarah Sled, Jim Snyder, Marco Steenbergen, seminar participants at Harvard University, MIT, the University of Chicago, the University of Michigan, and Yale University, and especially Eric Schickler. I am indebted to a team of research assistants who processed and recoded the data, including Gabriel Lenz, Colin Moore, Alice Savage, and Jonathan West. I would also like to thank Ellie Powell and Ian Yohai for superb research assistance and Matthew Gusella for assistance in preparing the manuscript. For financial support, I thank Princeton University and MIT. Address correspondence to the author; e-mail: berinsky@mit.edu.

doi:10.1093/poq/nfl021

for a number of reasons. In addition to extending the study of political trends back to the 1930s, we can answer questions of great historical importance. The decade from 1935 to 1945, after all, was like none other in American history. From the deep economic depression of the 1930s to the war years of the early 1940s, American politics underwent a radical transformation. Why, for example, did democracy prevail in the United States during the Depression years of the 1930s, given its failure in Germany and Italy? How did attitudes toward racial policy evolve in the 1930s and 1940s? Why did the public continue to support the war effort even as casualties mounted in the Atlantic and Pacific theaters?

Fortunately, there is a great deal of data concerning the public's views during this time. Starting in the mid-1930s, polling companies—such as Roper and Gallup—surveyed the public about important issues on a monthly basis. The polls from this time contain much information on questions of central importance to political scientists, historians, and policymakers. Furthermore, the surveys from this era are readily available; individual-level data for over 450 polls conducted before 1950 are available from the Roper Center for Public Opinion Research (http://www.ropercenter.uconn.edu).[1] It is somewhat surprising, then, that modern researchers have largely overlooked the opinion polls from this time. Some researchers—most notably Page and Shapiro (1992)—have used the aggregate poll data to study patterns of stability and change in public opinion. But this work is the exception. For example, Erikson, MacKuen, and Stimson's (2002) pathbreaking study of macropolitical trends begins in the early 1950s. Furthermore, contemporary studies of individual-level behavior using poll data collected before 1952 are rare (though see Baum and Kernell 2001; Caldeira 1987; Schlozman and Verba 1979; Verba and Schlozman 1977; and Weatherford and Sergeyev 2000).

One reason why scholars have ignored these data is that, from a modern standpoint, the data collection methods seem substandard. Principles of random selection took a backseat to interviewer discretion and concerns over survey cost. But acknowledging that the data were collected in ways that now appear questionable does not mean scholars should ignore these early public opinion polls. Whatever their flaws, these polls still provide insight into the beliefs of the mass public during a critical era of American history. The alternative, after all, is to consign a whole era of unique survey data to the dustbin of history.

In this article I lay out a strategy to analyze the public opinion data of the 1930s and 1940s. I first describe the quota-control methods prevalent during this time. I then detail the problems introduced through the use of quota-control

1. These data are extremely difficult to work with. Most of these surveys have not been touched for almost 60 years, and as a result, the data sets are not easily usable. Eric Schickler and I are undertaking a project designed to make the data from the 1930s and 1940s more easily accessible to the larger social science research community (funded under National Science Foundation, Political Science Program Grants SES-055043 and SES-0550199). We are recoding the individual files, preparing documentation, and will make available a series of weights to mitigate the biases resulting from quota-controlled sampling (described below).

techniques. Next, I describe specific strategies that researchers can employ to ameliorate these problems and study both aggregate trends in public opinion and individual-level relationships among variables of interest. At the aggregate level, I discuss methods of weighting; at the individual level, I advance the use of regression analysis with control variables. Finally, I use examples from several public opinion studies in the early 1940s to show how the methods of analysis laid out in this article enable us to utilize the data available to enrich our understanding of the shape and structure of public opinion during the years before the 1950s. Though the use of the methods advanced in this article does not always change our estimates of political behavior, those methods put the study of public opinion from the 1930s and 1940s on firm inferential ground.

Quota-Controlled Sampling: An Introduction

Modern opinion polls in the United States are conducted using a probability sampling method, such as random digit dialing telephone interviewing or a face-to-face multistage area probability sampling design.[2] While these methods are not equivalent to simple random sampling (SRS), statistical methods have been developed to account for the design components of modern survey sampling, such as clustering and stratification (see Kish 1965; Lohr 1999).

Such methods of probability sampling may be the norm in the United States today, but these sampling schemes have not always reigned supreme in survey research.[3] While probability sampling methods were well established in government agencies, such as the Census Bureau and the Works Progress Administration, by the late 1930s (Converse 1987), nongovernmental opinion polls in the United States before the 1950s were conducted using an entirely different methodology, namely quota-controlled sampling methods.[4]

2. For example, the National Election Study (NES) is conducted using a multistage area probability design. Sampling proceeds through a series of nested stages, with probability methods employed in all stages of sample selection. It should be noted that some contemporary organizations do use a form of quota sampling to select particular respondents from a given household. However, unlike the quota samples discussed here, the households in which the interview is conducted are selected randomly (see Voss, Gelman, and King 1995). This method of sampling, termed "probability sampling with quotas" by Sudman (1966), imposes stringent controls on the interviewer because the interviewer has no discretion in sampling households. Thus, these samples avoid many of the errors introduced by the wide interviewer discretion inherent in the quota sampling procedures used in the 1930s and 1940s.

3. In Europe today, many polls use modified quota sampling methods to collect data. These quota sampling methods avoid the most egregious errors of the quota-controlled sample polls discussed herein. However, many of the arguments that occurred in the United States's surveying community during the 1930s and 1940s continue abroad. For instance, advocates of probability sampling attacked quota sampling methods used in England after the 1992 preelection polls incorrectly predicted a Labour win (see Lynn and Jowell 1996; Worcester 1996).

4. Until 1950, survey research firms almost exclusively used quota-controlled sampling. There were, however, vigorous debates between proponents of probability sampling (or area sampling) and supporters of quota-controlled sampling. The intellectual roots of this debate can be traced to the work of Neyman (1934), who developed the theory of probability sampling and argued that it was superior to methods of purposive sampling, such as quota-controlled sampling. Academic

Pollsters employed this sampling method both for commercial polls—such as those conducted by Roper and Gallup's American Institute of Public Opinion (AIPO)—and public-interest polls—those conducted by the Office of Public Opinion Research (OPOR) and the National Opinion Research Center (NORC). The details of quota-controlled sampling varied across survey organizations, but all polling firms used the same basic strategy. Under a quota-controlled sample, the survey researcher sought to interview predetermined proportions of people from particular segments of the population. The researcher controlled this process by dividing—or "stratifying"—the population into a number of mutually exclusive subpopulations (or strata) thought to capture politically relevant divisions (such as gender and region). Researchers then allocated the sample among these strata in proportion to their desired size.[5]

This desire to achieve representative samples through strict demographic controls represents the largest point of departure between the pollsters of the 1930s and 1940s—such as Gallup and Roper—and modern survey researchers in the United States. Gallup and Roper did not trust that chance alone would ensure that their sample would accurately represent the sentiment of the nation. Through the selection of particular interviewing locales and the construction of detailed quotas for their employees conducting interviews in those locales, these researchers presumed that they could construct a representative sample.[6]

researchers and members of the Census Bureau supported probability sampling, arguing that quota-controlled sampling had no basis in statistical theory. In particular, they noted that because the probabilities of inclusion in the sample of the population elements were unknown, the estimates of sampling error of quota samples that were based on sampling theory would be incorrect (Bershad and Tepping 1969; Hansen and Hauser 1945). In addition, some researchers faulted pollsters for setting fixed quotas, thereby "tampering with random sampling" (Warner 1939, p. 381; for more general critiques of quota-controlled sampling, see Hansen and Hauser 1945; Hauser and Hansen 1946; Johnson 1959). At the time, this debate became quite heated (see, for example, the exchange regarding Meier and Burke [1947] between Banks on the one hand and Meier and Burke on the other in Banks, Meier, and Burke [1948]). The debate was not simply one of "academics" versus "industry." As Converse (1987, p. 126) notes, "Social scientists were not, on the whole, very critical of the quota sample itself until the mid- and late-1940s. Those who were involved in survey work themselves accepted the practicality of the quota sample." Some academics even came to the defense of quota-controlled sampling (Wilks 1940). The debate over quota-controlled samples was never completely settled. But the fallout from the 1948 election—in which all the major polls that used quota sampling predicted a Thomas Dewey victory—hastened the demise of quota-controlled samples. Though Gallup never admitted that his sampling methods failed in 1948, he gradually moved to probability sampling methods through the early 1950s (Hogan 1997). On this point, it should be noted that, contrary to conventional wisdom, the Social Science Research Council (SSRC) study of the election did not place the blame for the failure of the preelection polls primarily on quota sample methods (Mosteller et al. 1949).

5. Specifically, quota-controlled sampling proceeded in three steps. First, the surveyor apportioned the sample to geographic regions. Next, he selected specific locales, such as cities and towns, in which to conduct interviews. These first two stages of the sampling process are similar in practice to cluster sampling (though the selection of clusters seems to have been done in a haphazard manner). Finally, he sent interviewers to these locales to collect interviews, constraining their selection of respondents through strict quotas on particular demographic characteristics.

6. For Gallup, a "representative" sample was one that represented the population of U.S. voters; for Roper, a "representative" sample was one that represented the full population.

A quota-controlled sample—and only a quota-controlled sample—was the microcosm of America desired by the pollsters.[7]

The tight control exerted over respondent selection slipped once the interviewers were assigned their quotas. Polling organizations wanted to obtain completed interviews quickly to speed the data collection process and keep costs to a minimum. Once in the field, interviewers were given wide discretion in selecting the particular people they chose to interview.[8] Interviewers could go wherever they chose to secure interviews, as long as they stayed within their large geographic assignment. Interviews were conducted in people's homes and on the street. Potential respondents who did not wish to be interviewed were simply replaced with more willing citizens. As long as the interviewers met their specific quotas on age, gender, and economic class, the pollsters directing the major surveys of the period were satisfied.[9]

Table 1 summarizes the sampling schemes of the three main survey organizations active in the late 1930s and early 1940s: Roper (for *Fortune* magazine), Gallup, and NORC.[10] Clearly, there was a great deal of variation across the organizations in their specific polling practices, but there was also a significant amount of common ground. Each survey organization set quotas on the geographic regions where interviewing occurred. In addition, every organization sought to control the distribution of gender and economic status through the imposition of quotas.[11]

7. The importance of obtaining descriptive representation by control, rather than descriptive representation by pure random sampling, is clear from Gallup's writings. Gallup believed that exerting tight control over the sampling procedures improved the performance of the polls. In his 1944 book, *A Guide to Public Opinion Polls*, Gallup argued that his polling organization "goes one step further" than probability samples by stratifying the population into its major groups and then "randomly sampling" within those groups. Such a sampling procedure, Gallup argued, was used by polling organizations because it "served the added purpose of revealing what each group in the population thinks" (1944, p. 98). Neyman (1934), however, showed 10 years earlier that purposive sampling was inferior to random sampling. However, he proposed the use of limited controls through the method of random stratified sampling.

8. The arbitrary nature of respondent selection is made clear by the *Journal of Educational Sociology*'s description of Roper's sampling procedures: "interviewers go out on foot or by car and use their own judgment—for which the requirements are high—regarding which doorbells to ring, what shanties to visit, who to approach in the country store to get the specified proportions in each class" ("*Fortune* Survey" 1940, p. 252).

9. The reduction of interviewer discretion is the largest change in quota procedures from the 1930s to the present. Firms that use quota methods—such as MORI in Britain—today give strict instructions regarding quasi-random procedures for respondent selection (Worcester 1996).

10. This chart represents the compilation of sometimes contradictory information from a variety of sources because there is no single source for this information. While Converse (1987) provides an excellent history of the survey research enterprise, she does not describe the specific sampling procedures used by the different firms. In the chart I make a distinction between "hard quotas"—those quantities controlled through a strict distribution—and "soft quotas"—those variables where interviewers were instructed to get a "good distribution."

11. The economic class variables were especially tricky. "Class" was defined in relation to the particular geographic context in which the interview took place. For instance, Roper used an economic-level designation that took account of variations across geographic regions and the size of place in average income levels (Roper 1940b). Similarly, Gallup classified his respondents into six categories, ranging from "wealthy" to "on relief." The polling firms set their quotas in relation to these classifications, in a somewhat arbitrary manner. These measures were dropped by the 1950s (Smith 1987).

Table 1. Quota-Controlled Sampling Schemes

	Strata Selection (Purposive Selection by Central Office)		Quota Controls (Guide Respondent Selection by Interviewers)	
	Geographic Region	Interviewing Area Selection: Size of Place	Hard Quota (Distribution set)	Soft Quota (Distribution encouraged)
AIPO[a]	Census Region: Cases assigned to South and non-South on the basis of previous presidential election turnout States: State quotas determined directly from total sample size in proportion to the state's contribution to the total vote in the previous presidential election[b]	Possible interviewing locales stratified by region and regional city/size strata. Sample selected to provide a broad geographic distribution of cases Sample assigned to size classes in proportion to their presidential vote in the previous presidential election	Gender	Age: Interviewers instructed to get "a good spread" Economic Class: (wealthy/average+/average/poor+/poor/on relief): Interviewers seek a distribution Occupation[c]
Roper[d]	Census Region: Cases assigned to regions in proportion to census numbers States: Sample apportioned to states within geographic regions on the basis of votes in the previous presidential election	Cities: considered by themselves Non-Cities: Stratify counties by size, select counties By 1940, 80 sampling places were selected[e]	Gender Age 1938–1943: 21–39/40+ 1943–1961: 21–34/35–49/50+ Economic Class (A–D, Negro) Occupation	

Table 1. *(Continued)*

	Strata Selection (Purposive Selection by Central Office)		Quota Controls (Guide Respondent Selection by Interviewers)	
	Geographic Region	Interviewing Area Selection: Size of Place	Hard Quota (Distribution set)	Soft Quota (Distribution encouraged)
NORC	Census Region: Cases assigned to nine regions in proportion to census numbers	Interviewing locations chosen based on size of city	Gender[f] Age Economic Class (A–D, Colored) for non-farmers	Economic level for farmer; interviewers told to "see that the farmers you get represent a true cross-section, economically of your particular rural territory" (1943, 640).

[a] AIPO did the field work for the OPOR surveys from 1940 to 1943.

[b] The information concerning the allocation of cases to the states comes from the SSRC report on the 1948 preelection polls (Mosteller et al., 1949). It is possible that Gallup allocated his cases differently in the late 1930s and early 1940s. However, circumstantial evidence suggests that Gallup did not change his geographic allocation procedure. First, the distribution of cases among the states remained reasonably stable over the period. Second, in the 1936 election—when AIPO used a partial mail-balloting procedure—Gallup allocated the cases directly to the states (Katz and Cantril 1937). It should also be noted that Gallup's practice had the effect of malapportioning the distribution of cases between the South and the rest of the country. Because the sample was allocated to each state on the basis of previous presidential turnout, a disproportionately small number of cases was assigned to the South.

[c] Cantril et al. (1944, 148–49) reports that occupation was a "partially controlled variable"—akin to education—and notes that "the overrepresentation in the poll samples of the groups labeled professional, managers, and officials and the accompanying underrepresentation of the worker groups show that a definite occupation bias exists."

[d] This information comes in part from Roper (1940a).

[e] "*Fortune* Survey" 1940.

[f] Apparently, for some surveys NORC cross-classified the gender quotas with the other quota variables, such as age and economic class. For example, when crossing the gender and age quotas, NORC set specific quotas for men over 40, men under 40, women over 40, and women under 40 (Stephan and McCarthy 1957).

Quota-Controlled Sampling: Problems and Concerns

Whatever their potential appeal, in practice the quota-controlled sampling procedures used by the major survey research firms did not lead to samples representative of the U.S. population. The characteristics of the samples collected by the pollsters differed from the characteristics of the population, as measured by the census, in two ways. First, the exacting quotas set by particular pollsters sometimes guaranteed a sample unrepresentative of the population. Put another way, the sample consisted of *nonrepresentative strata*. Second, the lack of strict direction given to field workers regarding the standards for filling the defined quotas allowed—and perhaps even encouraged—interviewers to select unrepresentative respondents. Thus, the sample was plagued by *nonrandom selection within strata*. I discuss each of these two types of sample distortions in turn.

NONREPRESENTATIVE STRATA: REPRESENTING PERCEIVED VOTERS, NOT CITIZENS

Contrary to their populist rhetoric, not all pollsters were interested in obtaining descriptively representative samples of Americans. While Roper drew samples to conform to the census population figures, Gallup drew samples to represent each population segment in proportion to votes they usually cast in elections, rather than in proportion to number in the population (Mosteller et al. 1949; Robinson 1999). Thus Gallup's "representative" sample was intended to represent the voting public, not the full population of the United States. As a result, the proportion of the sample within each stratum does not accurately represent the full population proportions.

Women, southerners, and blacks voted at low rates in the 1930s and 1940s. These groups were therefore deliberately underrepresented in the Gallup samples. For instance, from the mid-1930s through the mid-1940s, Gallup designed his samples to be 65 to 70 percent male.[12] Similarly, the Gallup polls also contained a disproportionately small number of southern respondents. Where the census showed that 3 out of 10 residents of the United States lived in the South, Gallup drew only 10 to 15 percent of his sample from that region.[13] The important point to note here is that the Gallup data come from a deliberately skewed sample of that public.[14]

12. In addition to low turnout rates among women, Gallup believed that women would vote in the same manner as their husbands. To use Gallup's words, "How will [women] vote on election day? Just as exactly as they were told the night before" (Gallup and Rae 1940, pp. 233–34).
13. These same demographic discrepancies can be found in the OPOR samples in the early part of the war, when Gallup did the fieldwork for the surveys Cantril designed (in 1942 Cantril created his own survey field organization).
14. For uses of this data as "representative" of public opinion, see, for example, Kennedy (1999) and Casey (2001), who regularly equate opinion polls from this period with the voice of the American public.

NONRANDOM SELECTION WITHIN STRATA: THE CONSEQUENCE OF INTERVIEWER DISCRETION

In addition to these deliberately induced sample imbalances, the practice of quota sampling also introduced a number of unintended errors. The geographic distribution of the sample was controlled from the main office. Once interviewers were sent to specific towns and cities to conduct the surveys, however, interviewer judgment became the guiding force. Apart from having to fulfill their demographic quotas, interviewers were given great discretion to select particular citizens to interview.[15] Moreover, interviewers could simply replace citizens who refused to be interviewed with other respondents who met the requirements of the quota.[16] Since interviewers preferred to work in safer areas and tended to question approachable respondents, the selection of a particular respondent within a given stratum was not random.

The difference between the sample of survey respondents and the mass public is most apparent on measures of educational attainment. Comparisons between census data and AIPO data show that Gallup's respondents were better educated than the mass public.[17] The 1940 census indicated that about 10 percent of the population had at least some college education. But almost 30 percent of a 1940 Gallup poll sample had some college education.[18] This skew is almost certainly a result of the fact that better-educated respondents lived in areas traversed by interviewers and were more willing to be surveyed than citizens with less education.[19] The problems of an overeducated sample were not

15. Even though the marginals on the quota categories may be correctly balanced, the specific population breakdowns within those quota categories may be incorrect. Survey interviewers were instructed to fill their quotas, not to cross their quotas. The fact that the quotas were not integrated led to some strange interviewing practices. As one interviewer recalled, "When we got to the end of the week, we did our best to fill the quotas. We would do 'spot' surveys. We'd drive down the streets trying to spot the one person who would fill the specific quota requirements we had left" (quoted in Moore 1992, p. 65). The net effect of these practices was to skew the joint distribution of the quota variables, a practice that can be accounted for in the analysis of the data.

16. Proponents of area sampling procedures were especially critical of the practice of replacing refusals and hard-to-reach respondents with other respondents. Several studies at that time found that the opinions and behaviors of respondents who were easily reached differed from respondents who were contacted only after several attempts, even though the two groups had similar demographic characteristics (Campbell 1946; Noyes and Hilgard 1946).

17. Gallup was aware of the problems created by the unequal distribution of education within his sample, and in 1948 he weighted his sample by education when reporting election polls to reflect the census estimates of education (Mosteller et al. 1949). This change may have been in part a reaction to criticism raised over his over-prediction of the Republican vote in 1940 and 1944 (Katz 1944).

18. This skew in the education of the sample is typical of the Gallup and OPOR polls I have examined. Roper did not measure education in the early war period, so it is not possible to confirm the education bias. However, given that the same education bias present in the Gallup data is found in Roper data collected in the mid-1940s and that Roper interviewers followed procedures for obtaining respondents similar to those of Gallup, the Roper data from the late 1930s and early 1940s is almost certainly biased toward those with more education.

19. See Cantril et al. (1944) for a discussion of this phenomenon (especially p. 148).

just the concern of Gallup.[20] Each time a survey organization in the 1930s and 1940s measured education, the distribution of that variable was tilted toward those with a college education. Similarly, polls conducted by Gallup and Roper tended to include a greater percentage of people with "professional" occupations than recorded by the census. The skew in these variables is not surprising, given that education and occupation were not quota categories. It is likely that the highly educated and professionals were more willing to be interviewed, and as a result, they comprise a disproportionately large share in these samples.

In addition to attracting particular types of respondents who may not have been fully representative of the population, the discretion given to interviewers may have allowed for inappropriate practices of respondent selection, thereby generating additional sources of error.[21] For instance, the quotas were set on the aggregate numbers, not on the subgroups within those aggregates. As long as the quotas were filled, the composition of the survey across quota categories did not matter. Thus, interviewers for Roper could fill age and gender quotas by interviewing only old men and young women rather than a combination of young and old men and young and old women. In short, the potential for interviewer effects during the interview process is potentially greater than that found in modern survey research. Any systematic differences between interviewers in respondent selection style or interviewing practices could be reflected in the survey responses collected by the polling agencies.[22]

SUMMARY: THE NATURE OF QUOTA-CONTROLLED SURVEY DATA

Due to the existence of these two sources of potential error—nonrepresentative strata and nonrandom selection within strata—some modern survey researchers view survey data from the 1930s and 1940s with great suspicion. Many political scientists who are aware of the limitations of polls conducted before

20. Experimental studies by Hochstim and Smith (1948) and Haner and Meier (1951) found that quota-controlled samples suffered from a larger skew on education than probability samples. Moreover, Haner and Meier found that the education bias was orthogonal to class bias. The skew in the education distribution induced by quota sampling relative to probability sampling was also found in a comparison conducted by the Social Science Research Council in 1946 (reported in Stephan and McCarthy 1957).

21. In order to fill their interview assignments quickly, some interviewers seem to have engaged in the questionable practice of interviewing several respondents at one time. While these practices almost certainly introduced significant bias, evidence suggests that few interviewers engaged in such egregious behavior. The major polling companies actively discouraged their interviewers from these practices. For instance, NORC (1943, p. 646) specifically admonished interviewers not to interview people in groups because "the presence of even one other person may influence the respondent to change his answer."

22. For instance, Cantril's (1944) analysis of an October 1940 AIPO poll found that interviewers collected opinions on the war that were, on balance, in line with their own beliefs. Upon closer examination, Cantril found that this finding was caused by the behavior of interviewers in small towns and rural areas. Cantril argued that this result was driven by the fact that interviewers were better known to their respondents in a small town and may have chosen respondents who thought like they did.

1950 have arrived at a simple solution: they reject the polls out of hand. For example, Converse (1965) concludes that the AIPO and Roper data "were collected by methods long since viewed as shoddy and unrepresentative." Rivers argues that quota sampling is "a methodology that failed" (quoted in Schafer 1999). The dearth of modern research that makes use of the opinion polls from the 1930s and 1940s speaks to the apparent power of the critique raised by Converse and Rivers. Such criticisms may be valid; clearly the early opinion polls have a number of substantial flaws. But this does not mean that the critical information those polls contain should be abandoned. Instead, we should recognize the inherent value of this data while taking into consideration the flaws of the collection procedures. In the next section, I advance some simple methods to draw the best inferences we can from the data.

A Method of Analysis

The central problem introduced by quota-controlled sampling is that many of the survey samples do not represent certain groups in proportion to their share in the population. Though demographic characteristics of respondents do not always predict attitudes or behaviors well, in some circumstances gender, region of residence, and education can be powerful determinants of political predilections. In fact, these variables were used to construct the quotas because pollsters thought they would predict opinion cleavages on politically relevant questions.[23] The mismatch between the demographic characteristics of poll samples and the demographic characteristics of the population, as measured in the census, could therefore lead to a mismatch between the political views expressed in the polls and those that might have been expressed by the population at large.

Take, for example, the Gallup surveys. We know that there was a large regional split in party attachment in the 1930s. To the extent that southerners are underrepresented in the sample, Democratic identifiers will be underrepresented as well. On political questions where strong party cleavages exist, the Gallup polls will therefore misrepresent the voice of the mass public, writ broadly. Put simply, these polls will give us a picture of public opinion that excludes significant portions of the citizenry. Such polls may represent the "voting public" to a certain degree, but the existing aggregate measures—such

23. For example, Roper created quotas based, in part, on a division of the sample along gender lines. In Roper's view, the use of this dividing line was critical because "there are certain subjects on which [men and women] think quite differently. For one thing, many women are apt to feel that certain subjects are peculiarly 'men's subjects' and, therefore, they prefer not to express an opinion but to insist that their answers be recorded in the 'don't know' column. Furthermore, we have found that on many issues women are less apt to go for the extreme viewpoint than are men, and are more content to select a moderate statement as being closest to their own viewpoint" (1940a, p. 327).

as those reported in Cantril and Strunk (1951)—do a poor job of representing the opinion of *all* citizens in the United States.

Of course, some might argue that the strategy adopted by Gallup is advantageous for the purposes of assessing the public will. From this perspective, public opinion polls *should* represent the engaged public. In many ways, Gallup's strategy was similar to the use of screening questions in modern opinion polls to weed out nonvoters. Gallup simply created his representation of the voting public at the sampling stage, rather than at the analysis stage. However, such a strategy is flawed for two reasons. First, while we might choose to ignore those people who do not participate in the political process without examining the biases in opinion polls, we cannot know what types of sentiment we miss in those polls (Berinsky 2004). Second, even if we believe that we should represent only the engaged public, Gallup's performance in predicting elections during this time suggests that the AIPO sample was not contiguous with the voting public. In every election from 1936 to 1948, Gallup underestimated the Democratic vote share in his final preelection poll, most famously in the 1948 presidential election (Robinson 1999).

The problem of nonrepresentativeness, though potentially troubling, is surmountable. The quota-controlled sample data were collected in ways that appear from a modern vantage point to be slapdash, but, contrary to the suspicions of those who have taken issue with these data, the data were not collected so haphazardly as to make population inferences inherently unreliable. If, for instance, the data were unreliable, repeated quota samples using identical survey questions should show large amounts of instability in point estimates and marginal frequencies. An examination of wartime opinion trends presented in Cantril (1948) demonstrates a remarkable level of over-time stability quite consistent with the stability found in trends derived from modern random probability samples (see also Page and Shapiro 1992).[24] The data collection process employed in the 1930s and 1940s, rather, introduced *predictable* deviations between the characteristics of the sample and that of the population.[25] We can therefore employ methods designed to account for these

24. Baum and Kernell (2001) show that even subgroup trends exhibit striking levels of stability (though partially because the trends in their analysis were filtered and smoothed; see also Berinsky 2006).

25. Furthermore, there are some cases where researchers have found that quota sampling polls and probability sampling polls yield similar results. First, Meier and Burke (1947) compared area sampling, probability sampling, and quota sampling procedures, mimicking the three methods by re-sampling previously collected survey data collected from Iowa City in 1946. Though the sample sizes for these surveys were quite small ($N = 50$), Meier and Burke concluded that quota and area sampling methods would yield essentially similar results (though quota sampling methods tended to overstate home-ownership levels relative to the other methods). A second study was carried out in Great Britain in the early 1950s. Moser and Stuart (1953) conducted an experiment in which multiple samples were collected using quota methods and compared with probability samples of the same population. The researchers found several problems with the quota samples, including the familiar upward skew in the distribution of occupation status and educational

differences in order to make reasonable inferences about the political preferences of the American public.

Aggregate-Level Analysis: A Weighting Solution

The quota-controlled sampling procedures effectively introduced a unit nonresponse problem: certain classes of individuals were either deliberately or inadvertently underrepresented in the samples. When we have detailed information concerning the characteristics of nonrespondents, we can employ procedures that model selection bias to account for the differences between the sample and the population (Breen 1996). But when we only have information about the population relative to the sample—auxiliary information taken from the census—weighting adjustments are typically applied to reduce the bias in survey estimates that nonresponse can cause (Holt and Elliot 1991; Kalton and Flores-Cervantes 2003; Lohr 1999).[26]

I employ a model-based poststratification weighting scheme.[27] Under the model-based approach to sampling, the values of the variables of interest are considered to be random variables. Model-based approaches therefore involve employing a model of the joint probability distribution of the population variables (Thompson 2002). It is necessary to take a model-based approach to draw inferences from quota samples because, as Lohr (1999) notes, we do not know the probability with which individuals were sampled (see Smith 1983 for a discussion of a model-based approach to using of poststratification weights for quota-sampled data).

Though the use of weights to adjust for nonresponse is common, there is controversy about the best way to implement weighting (Bethlehem 2002; Deville and Sarndal 1992; Deville, Sarndal, and Sautory 1993; Gelman 2005; Gelman and Carlin 2002; Kalton and Flores-Cervantes 2003; Little 1993; Lohr 1999). Different researchers may prefer different weighting techniques. For the interested reader, I therefore discuss and implement three solutions recommended by the survey weighting literature: cell weighting, raking, and

achievement. However, though they cautioned that quota sampling was not theoretically sound, they concluded, "Without wanting to belittle the drawbacks of quota sampling, we feel that statisticians have too easily dismissed a technique which often gives good results and has the virtue of economy" (Moser and Stuart 1953, p. 388).

26. The use of weighting adjustments can lower the precision of survey estimates because we trade off a reduction in bias for an increase in variance (Kalton and Flores-Cervantes 2003). Given the compositional imbalance in the quota samples, I believe that this is a worthwhile trade-off.

27. The poststratification weights I employ are different from probability weights, which are known at the time the survey is designed and are used to adjust for nonconstant probability of sample inclusion. It is not possible to employ probability weights for the quota samples I examine here because the individual units have an unknown probability of inclusion.

regression estimation.[28] My preferred method is cell weighting because it is simple to employ and requires minimal assumptions. As I will discuss below, the estimates from the three methods are very similar. Thus, if a researcher prefers a technique other than cell weighting, it is preferable to implement that technique rather than to avoid weighting altogether. Even if the use of weights does not change much our picture of public opinion, their use places inferences drawn from the quota sampling data on a firm statistical foundation.

CELL WEIGHTING

Cell weighting is a simple way to bring the sample proportions in line with auxiliary information—namely, census estimates of the population proportions. We stratify the sample into a number of cells (J), based on the characteristics of the population deemed important (the matrix of X variables). If the distribution of demographic variables in the sample differs from the distribution in the population, cell weights are used to combine the separate cell estimates into a population estimate by giving extra weight to groups underrepresented in the sample and less weight to overrepresented groups. Using cell weighting to adjust for nonresponse requires us to assume that the respondents within a given cell represent the nonrespondents within that cell. That is, we must assume that the data is missing at random (MAR) (Little and Rubin 2002; I discuss possible violations of this assumption below). Under cell weighting, the estimate of the population mean for the quantity of interest is:

$$\hat{\theta} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\theta}_j$$

where N_j is the population size in each poststratification cell j, N is the total population size, $\hat{\theta}$ is the weighted estimate of the mean, and $\hat{\theta}_j$ is the sample mean for poststratification cell j (Lohr 1999).

A well-known practical limitation of cell weighting is that as the number of stratification variables increases, the number of weighting cells becomes larger. With fewer cases in each cell, the aggregate estimates derived from the weighting estimator are less stable. Here the large sample size of the early opinion polls is advantageous. Because these polls often collected samples of 3,000–5,000 cases, the likelihood that any particular cell in a cross-classification matrix of three or four variables would be devoid of cases is lower than it would be with contemporary survey

28. I also implemented a Bayesian regression modeling approach advanced by Gelman (2005). However, this method is very complicated and difficult to implement. Furthermore, the method yields results similar to those presented here. Given the similarity of the results and the complexity of this method, I omit discussion of this technique from the article. These additional results are, however, available upon request.

data.[29] We can, for example, usually construct a gender by education by region weighting system and still have around 20 cases per cell—the minimum level of support recommended by Lohr (1999).[30] Moreover, cell weighting has several advantages. First, as Lohr notes, cell weighting requires minimal assumptions beyond that the data is MAR (an assumption common to all forms of weighting, including those discussed below). For instance, as Kalton and Flores-Cervantes (2003) note, unlike other methods, cell weighting requires no assumptions regarding the structure of response probabilities across cells. In addition, cell weighting allows the researcher to take advantage of information concerning the joint distribution of weighting variables—information that is available from census data.

RAKING

While cell-by-cell weighting allows researchers to use only a limited amount of auxiliary data owing to sample size issues, raking—also known as iterative proportional fitting (Little and Wu 1991) or rim weighting (Sharot 1986)—allows researchers to incorporate auxiliary information on several dimensions. Raking matches cell counts to the *marginal* distributions of the variables used in the weighting scheme. For example, say we wish to weight by gender and a three-category age variable. Picture a three by two table, with age representing the rows and gender representing the columns. We have the marginal totals for each age group, as well as the proportion of men and women in the sample. The raking algorithm works by first matching the rows to the marginal distribution (in this case, age), and then the columns (gender). This process is repeated until both the rows and columns match their marginal distributions.[31]

29. In the early days of opinion polling, survey researchers drew much larger samples than those used by modern pollsters. Though the sample size varied from poll to poll, the AIPO samples tended to be about 3,000 cases, the NORC samples about 2,500 cases, and Roper samples about 5,000 cases.
30. Elliot (1991) reports that the U.S. Bureau of the Census insists on a minimum of 30 respondents per class.
31. More formally, let w_{ij} represent the weight for each cell in the three by two cross-classification table of interest. The goal is to estimate $\hat{w}_{ij}$ from the marginal proportions in the population, w_i (i = 1, 2, 3) and w_j (j = 1, 2), respectively. This can be achieved with the following algorithm (Little and Wu, 1991):

1. Initialize the weights by setting each equal to n_{ij}/n, which is the sample cell count over the sample size.
2. Calculate $\hat{w}_{ij}^{(1)} = w_i * \frac{\hat{w}_{ij}^{(0)}}{\sum_j \hat{w}_i^{(0)}}$ for each i. Here we are "raking" over the rows.
3. Calculate $\hat{w}_{ij}^{(2)} = w_i * \frac{\hat{w}_{ij}^{(1)}}{\sum_i \hat{w}_j^{(1)}}$ for each j. Here we are "raking" over the columns.
4. Repeat 2 and 3 until $\sum_j \hat{w}_i = w_i$ and $\sum_i \hat{w}_j = w_j$ for each i and j, which is when "convergence" is achieved. Though the process can be slow, generally the raking algorithm converges reliably.
5. The estimate of the weighted mean is then $\hat{\theta} = \sum_i \sum_j w_{ij} \hat{\theta}_{ij}$.

Raking allows for more weighting variables to be included than under cell weighting. Little (1993) demonstrates that if the data have an approximately additive structure, raking is an appropriate response to the small cell problem. However, in the case of data from the 1930s and 1940s, we have less information on the demographic characteristics of the respondents then we typically do today. Thus, this concern is less important than it would be for modern researchers, who typically use raking to adjust samples on seven or more variables. Moreover, raking ignores information available in the joint distribution of the weighting variables. Raking requires that "the response probabilities depend only on the row and column and not on the particular cell" (Lohr 1999, p. 271). For example, we must presume that females in the North do not have systematically different preferences from females in the South. If groups are homogenous in their political behavior, such an assumption is plausible. But the sample design of the polls from this period makes it difficult to adopt such an assumption. The quotas, after all, were set on the aggregate numbers, not on the group-based differences within those aggregates. Other methods are more appropriate if we expect there to be meaningful interactions among the weighting variables, a problem addressed by both cell weighting and the regression weighting method I consider next.

REGRESSION ESTIMATION

An alternative method of including auxiliary information is regression estimation, which uses variables that are correlated with the variable of interest to improve the precision of the estimate of the mean (Lohr 1999). In this method, an ordinary least squares (OLS) regression model is estimated using the sample data.[32] The dependent variable is the quantity of interest—be it vote for the president or support for a given policy. The independent variables are the variables in the data set that can be matched to measured quantities in the population (e.g., census data on age and gender). The population information is then used to estimate the population means of the dependent variable in the regression. Specifically, the mean population values of the independent variables are combined with the coefficients calculated from the sample model to derive fitted value estimates of the population mean of the quantity of interest.

$$\hat{\theta} = \underset{pm}{X}\hat{\beta}$$

32. Strictly speaking, since our dependent variables of interest are often binary (e.g., vote choice), using logit or probit would be more appropriate in those situations than OLS. But OLS allows us to interpret the poststratified estimate of the mean generated from the regression estimation method as a weighted average of the data (see Gelman 2005). I believe this justifies using the linear model in order to ease the interpretation of the results.

where X_{pm} is the vector of population mean values, and $\hat{\beta}$ is the least squares estimator $\hat{\beta} = (X'X)^{-1} X'y$. A disadvantage of regression weights is that a new regression model—and a new set of weights—must be calculated for each dependent variable analyzed. Furthermore, regression weighting requires an assumption of multivariate normality in addition to the MAR assumption required by the other weighting methods.

SUMMARY

In this section, I have outlined three weighting methods that differ in their particulars, but all rest on the same foundations. As noted above, I prefer the cell weighting method for this data because of its simplicity, its minimal assumptions, and because it uses information from the joint distribution of the quota variables. However, as I will demonstrate below, other methods give us essentially the same results—a picture of public opinion in the 1930s and 1940s that can in some cases differ appreciably from the picture gleaned by examining the raw data.

What to Weight On? The Art of Selecting Weighting Variables

Poststratification weighting rests on a firm statistical foundation, but in practice it requires a series of data analytic choices. All three weighting methods rely on auxiliary information to arrive at valid inferences. Kalton and Flores-Cervantes (2003) note that it is important to choose auxiliary variables that predict the response probabilities of the different cells. In other words, we need to be sure that our adjustments capture the fundamental differences between our sample and the full population. To make the case that weighting methods capture and correct differences between the survey samples and the population, I next discuss how auxiliary information from the census can be used to correct the problems introduced by quota sampling.

As discussed above, the distortions introduced by quota sampling can be divided into two types: nonrepresentative strata size and nonrandom selection within strata. Nonrepresentative strata size distortions arose through the instructions the central survey offices of the polling firms gave to the field staff—for example, by setting the quota of female respondents below their true population proportion, the pollsters deliberately skewed their samples. By contrast, nonrandom selection within strata distortions are a result of interviewer discretion in respondent selection. This discretion ensured that the citizens who were interviewed differed in systematic ways from citizens who were not. Though these distortions arise through different processes, both can be addressed with a common solution. By employing auxiliary information

about the population, we can correct for the known differences between the sample and the population.

The use of auxiliary information is an especially powerful way to correct for nonrepresentative strata size distortions. There is no reason to suspect that the members of deliberately underrepresented groups—such as women and southerners—who were interviewed were systematically different from the members of those groups who were not interviewed (conditioning on the interviewer-induced differences addressed below). After all, the sample imbalance exists because the pollsters deliberately drew nonrepresentative samples based on these characteristics. Thus, using the cases of the underrepresented groups who were interviewed to represent those respondents who were not interviewed is appropriate. We can simply use the observable characteristics of the respondents to reweight the data.

Auxiliary information can also be used to correct for distortions arising from nonrandom selection within strata. The latitude given to interviewers in selecting their subjects ensured that the probability of being interviewed depended on respondent characteristics that interviewers found attractive. Thus, within quota categories, those citizens who were interviewed were not necessarily representative of the population in that category, potentially violating the MAR assumption. Correcting for nonrandom selection within strata is difficult because we do not have information on the people who were not interviewed. However, we do sometimes have important information that can be employed in analysis—namely, the education level of the respondent. While interviewers did not explicitly select respondents on the basis of their schooling, education is the best proxy the surveys have for the "observables" that made an interviewer more likely to pick one individual from a given demographic group than another individual. The key is that education (1) is a powerful predictor of who is a desirable interview subject, (2) affects politically relevant variables, (3) was *not* used as a quota control, but (4) was often measured by survey organizations. Therefore, by utilizing auxiliary information on education levels, we can account for at least some of the problems introduced by nonrandom selection within strata. As education measures were not included in every survey in this period, however, a respondent's occupation can serve a similar purpose. For the same reason that interviewers gravitated to highly educated respondents, they tended to interview professionals at the expense of laborers and other members of the workforce. In addition, occupation, like education, was not used as a firm quota variable. Thus, even when education is not available, we can correct for nonrandom selection within strata with auxiliary information on occupation.

The use of auxiliary data—particularly on education and occupation—to account for differences between the sample and the population is imperfect. It is possible that the low-education respondents selected by interviewers were not fully representative of the population of low-education citizens. Thus, controlling for education does not completely solve the nonrandom selection

within strata problem because the use of weights may multiply the influence of respondents who are "unusual." However, education captures many of the important interviewer-induced differences between respondents and nonrespondents. While quota-controlled sampling procedures reduced the probability that certain individuals—women, southerners, nonprofessionals, and those with low education—would be interviewed, no individuals were completely excluded from the sampling scheme. It appears that every individual therefore had some probability—no matter how low—of being included in the survey samples of this era. By using auxiliary information on education and occupation, we can take advantage of the residue of interviewer-induced distortions to correct at least some of the problems caused by nonrandom selection within strata. Controlling for some of the problem through the use of proxy variables, such as education, is preferable to completely ignoring the problem. In essence, by conditioning on the variables that affect the probability that a given individual would be interviewed, we can better fulfill the conditions required by the MAR assumption. Without detailed information on nonrespondents, this strategy is the best solution available to modern researchers. The "check data" on turnout and telephone use presented below suggest that my weighting system moves us closer to an accurate depiction of the full population.

In sum, the use of auxiliary information can mitigate the deficiencies of quota sampling procedures. Thus, in aggregate analysis, researchers should weight the data on education levels and occupation and those quota category variables—such as gender, region, and age—that can be matched to census data (see also Glenn 1975 for a similar strategy). The necessary population counts for the 1940 census are available from the Integrated Public Use Microdata Series (Ruggles et al. 2004). Even when the use of weights leads to only a modest difference in our conclusions, their use nonetheless provides greater confidence that our estimates are not attributable to problematic sample design.

Application to Individual-Level Analysis

Thus far I have outlined a strategy to estimate aggregate statistics, such as population means. Nevertheless, researchers are also interested in estimating more complex relationships among the variables available in the data through individual-level regression analysis. Whether or not to include weights in regression analysis is a source of ongoing controversy in the literature.[33] Several authors caution against the use of weights in this manner (see Lohr 1999, pp. 362–65, for a review). This admonition is especially pertinent here because my weights are poststratification weights, not sampling weights. Winship and Radbill (1994) note that when weights are solely a function of

33. The concern here with the use of weights in regression analysis is distinct from the regression estimation methods discussed above.

observed independent variables that can be included in a regression model—as is the case with our data—unweighted OLS will yield unbiased parameter estimates. Thus, the most straightforward method of dealing with the potential bias created by quota sampling is simply to include the weighting variables as independent variables in the regression model (Gelman 2005; Gelman and Carlin 2002).[34] In this case, the problem is similar to omitted variable bias: the oversampling of certain types of respondents—namely, highly educated white males—may mask the true relationship among other predictors if these variables are not controlled for. In this way, the individual and aggregate analyses are closely related, in that in order to get aggregate estimates, we average over the proportions of different types of respondents present in the population. Just as the cell weighting and the regression estimation methods incorporate information about the joint distribution of the sample, introducing the quota variables—and relevant interaction terms—as independent variables allows us to control for the sample imbalances introduced by the quota sampling methods of the time.

While it is possible to control for the bias in the coefficient estimates, compiling accurate measures of uncertainty is a complicated process. Standard tests of statistical significance assume that the data are drawn through probability sampling. Quota samples, however, rely on interviewer discretion for respondent selection, thereby diverging from strict random sampling. Thus, as Gschwend (2005, p. 89) notes, "it is neither clear according to statistical theory how to compute a standard deviation, nor how to estimate standard errors." In my own analyses, I follow the convention of other scholars who have analyzed the data (Baum and Kernell 2001; Schlozman and Verba 1979; Verba and Schlozman 1977; Weatherford and Sergeyev 2000) and present standard errors for the estimated regression coefficients. In effect, I analyze the data as though it were generated through probability sampling. However, my confidence in the validity of the results does not rely on the statistical tests alone. I also look for convergence in estimation results across different polls conducted by different organizations in similar time periods (see Berinsky 2006). Nevertheless, the question of generating standard errors for estimates—at both the individual and aggregate levels—should be the subject of future work.[35]

34. As Gelman (2005, p. 3) counsels, "In a regression context, the analysis should include as 'X variables' everything that affects sample selection or nonresponse. Or, to be realistic, all variables should be included that have an important effect on sampling or nonresponse, if they also are potentially predictive of the outcome of interest."

35. Stephan and McCarthy (1957) argued that the repeated application of a specified quota sampling procedure could generate an empirical distribution for an estimate that could be used to calculate the variance of that distribution (see chapter 10). For instance, if a particular variable was measured on a number of successive surveys, and we were to assume that these repetitions occurred under essentially similar circumstances—that is, the population, the measuring instrument, and the instruction and training of the interviewers did not change from survey to survey—we could treat each survey as a "draw" from the population and calculate the standard error of the estimate. They conclude that, in general, the variances of quota-sampled estimates were approximately 1.5 times as large as probability-sampled estimates. However Stephan and McCarthy concede that the estimation of sampling variation for quota samples cannot be placed on "the same sound theoretic footing as now exists for well designed and executed probability model samples."

Results

Having outlined appropriate strategies for data analysis, in this section I demonstrate how these strategies permit us to draw legitimate inferences concerning the shape and structure of public opinion in the 1930s and 1940s. For the purposes of analysis, I will primarily focus on several data sets collected by George Gallup's AIPO and Hadley Cantril's OPOR. All of the weights discussed below were created using the methods discussed above. The weights generated using the cell weighting method will be made available through a set of recodes available via a link from the Roper Center for Public Opinion Research Web page (http://www.ropercenter.uconn.edu).[36]

AGGREGATE ANALYSIS

The utility of the weighting strategy is demonstrated by calibrating the quota-controlled survey data to known quantities, notably election returns and telephone penetration. Because sampling error is a concern, it would be unrealistic to think that we could predict external quantities exactly with either the weighted or the unweighted estimates. But by examining the effect of the different weighting schemes on the estimates of known quantities, we can gauge how well the weighting process might bring us closer to the truth on opinion poll items.[37]

I first employed the three weighting schemes on Gallup's and Cantril's postelection poll estimates of turnout.[38] Table 2 presents the true turnout levels and the estimated turnout for the first postelection poll conducted in each of these election years.[39] As the table demonstrates, in the elections of 1940, 1942, and

(1957, p. 211). Furthermore, their method does not account for the error introduced by the free reign that interviewers had to select their own respondents. Nonetheless, Stephan and McCarthy's observations might serve as the basis for the development of a method to calculate standard errors for estimates derived from quota-sampled data.

36. These weights are being created as part of the Berinsky/Schickler data reclamation project (see note 1).

37. It would, in theory, also be useful to calibrate the quota sampling polls to probability sample polls in the late 1940s. Unfortunately, with one exception, I have not been able to find probability sampling and quota sampling polls that targeted the same population during the same field period. Furthermore, in that one case—the 1948 NES and the Gallup election poll discussed below—there are no opinion poll questions that overlap aside from the vote choice question.

38. Originally, I planned to use estimates of vote support for Franklin D. Roosevelt as a calibration measure. Upon further reflection, it was clear that the vote choice metric was inappropriate. The weights I use allow us to map the opinion polls collected in the 1930s and 1940s to the full census population of the United States. Thus, a weighted estimate of vote choice would provide the level of support for FDR among the *full population*. There is no reason to expect that this number should match the election numbers (which measure support among those who voted).

39. OPOR did not conduct a postelection poll in 1940. While OPOR did conduct a postelection poll in 1942, that data is not archived at the Roper Center.

Table 2. Aggregate Turnout Analysis

Poll	Actual Turnout	Raw Data	Cell Weight	Raked	Regression Weight
Gallup, November 1940	62.4%	89.2%	85.5%	85.3%	85.5%
Gallup, November 1942	33.9%	63.7%	55.0%	53.9%	54.1%
Gallup, November 1944	55.9%	82.8%	80.2%	79.1%	79.4%
OPOR, December 1944	55.9%	80.9%	77.4%	76.4%	76.5%

1944, pollsters overestimated turnout—not surprising, as their intention was to target likely voters. Applying the weights should bring our estimates closer to the actual rates of turnout. However, the use of weights will almost certainly not bring us to the true turnout levels because of the pervasive problem of vote overreporting in polls (Anderson and Silver 1986; Belli et al. 1999; Burden 2000; Clausen 1969; Traugott and Katosh 1979). Even the best probability samples, such as the National Election Studies (NES) overestimate turnout by as much as 20 percent (Bernstein, Chadha, and Montjoy 2001).

Table 2 presents the weighted results. Though applying the different weighting schemes to this data provides slightly different estimates, the weighted estimates are roughly in the same range as one another.[40] All three methods come closer to the true turnout figure than the unweighted estimates by approximately 5 percentage points, on average.[41] Thus, the use of weights not only puts the estimates of turnout on firm statistical ground but also moves those estimates closer to the actual behavior of the electorate.

Information about household telephone ownership can also be used to gauge the utility of weighting. According to Bureau of the Census (1976) estimates, 33 percent of American households had telephone service in 1936, a figure that had risen to 46 percent by 1945. Polls conducted by Gallup routinely asked respondents if they had a telephone in their household. Given the class and education biases in the sampling procedure, it should not be a surprise that the phone penetration in the unweighted sample of respondents would be greater than the population at large. In fact, these unweighted estimates overestimate phone penetration rates by about 10 to 15 points. For example, in the January 1941 OPOR poll, 47.3 percent of the sample reported having telephone service in their home.[42] This figure exceeds the actual level

40. The cell weighting method uses census data on gender, region, and occupation (or education, when available). The other methods use these data, as well as data on age. I do not use age in the cell weighting because of concerns with cell size. However, I did perform a series of analyses where I rotated age into my weighting scheme and found similar results to those presented here.
41. These results do not change when I expand the universe of polls to *all* postelection polls conducted in November and December of election years (results available from author upon request).
42. The exact wording of the OPOR question was, "Is there a telephone in your home (place where you live)?" While the sample design used by OPOR did not preclude multiple interviews in a single household, the percentage of respondents with telephones can be used as a measure of phone penetration during this time.

Table 3. Aggregate Telephone Penetration Analysis

Poll	Census Estimate	OPOR Raw Data	Cell Weight	Raked	Regression Weight
January 1941	39.3%	47.3%	39.9%	39.5%	40.6%
March 1941	39.3%	46.4%	38.9%	39.1%	39.3%

of phone penetration by 8 percentage points. However, once the sample is weighted by gender, education, and region, the estimated phone penetration rate drops to 39.9 percent—a close match with the census number (see table 3). Similarly, in a March 1941 OPOR poll, the estimated phone penetration rate was 46.4 percent. Introducing the weights again drops the estimated penetration rate closer to the true value. These figures are representative of a general pattern; across a variety of polls conducted by AIPO and OPOR in the 1930s and 1940s, applying the weights brings the estimate of phone penetration closer to the true number.[43] The analysis of the turnout and telephone penetration data is especially important because it indicates that the weighting methods allow us to better estimate the measurable characteristics of the mass public.

My analysis also indicates that weighting can change our view of collective public opinion on a number of issues important in the 1930s and 1940s (see table 4). The January 1941 OPOR survey contained several questions concerning interest in and knowledge of world events. For instance, respondents were asked, "Can you name a country where Greece and Italy are fighting?" Given that the sample of respondents overrepresented the highly educated, men, and nonsoutherners—all groups that tended to be more involved with politics—it is not surprising that the unweighted marginals overstate the true political interest and knowledge of the American public relative to the weighted estimates. We can see how these sample imbalances affect our estimates of the knowledge of the mass public by looking first at the knowledge levels of underrepresented groups relative to the overrepresented groups on the Greece/Italy question. Among respondents with a grade-school education or less, 44.2 percent gave the correct answer, "Albania," as compared with 76.2 percent of those with some college education or more. Fewer females answered the question correctly—40.0 percent, as compared with 62.0 percent of men—and southerners were less likely to answer the question correctly—52.3 percent gave the correct answer as compared with 55.0 percent of nonsoutherners. Since each of the underrepresented groups tended to have less political knowledge, weighting had a substantial impact on the estimates. Thus, while a majority gave the correct answer in the unweighted analysis, the results look substantially different when the weights are introduced to

43. The full weighted results are available from the author upon request.

Table 4. Aggregate OPOR Poll Analysis

Poll	OPOR Raw Data	Cell Weight	Raked	Regression Weighting
January 1941				
Know where Greece and Italy are fighting	54.7%	46.3%	44.1%	45.9%
Follow Lend-Lease?	72.1%	66.0%	66.0%	67.2%
Know how long Hitler has been in power	47.3%	41.2%	41.7%	42.7%
More important to defeat Germany than to keep out of war	61.7%	61.3%	60.8%	63.8%
October 1944				
Trust Russia?	58.9%	55.9%	54.4%	56.4%

correct for sample imbalances. In the weighted analysis, a minority gave the correct answer, no matter which method is used (see table 4). Similar differences are found in the proportion of respondents who said they had been following Lend-Lease, and the proportion that correctly identified Hitler's tenure in power at eight years (see table 4). Finally, sample biases can affect the shape of political attitudes. For example, an October 1944 OPOR survey asked respondents, "Do you think Russia can be trusted to cooperate with us when the war is over?" In the unweighted sample, 58.9 percent said they trust Russia, but applying the weights drops trust of Russia 2.5 to 3.5 percentage points, depending on the weighting method.[44]

This is not to say that correcting the sample imbalances through the use of weights will change our measures of public opinion in all cases. The January 1941 OPOR poll contained a number of questions concerning support for taking an active role in World War II. One question, which Cantril posed several times in this period, asked, "Which of these two things do you think is the more important for the United States to try to do—to keep out of war ourselves or that Germany be defeated, even at the risk of our getting into the war?"[45] The uncorrected proportion of people who said that it was more important to defeat Germany was 61.7 percent, while the weighted proportion was 61.3 percent. This confluence of the weighted and unweighted results can be explained by the nature of the cleavages on these issues. Women and poorly educated respondents—who are underrepresented in the survey—tended to oppose involvement in World War II, while southerners—who are also underrepresented in the sample—were highly supportive of involvement. Specifically, only 54.1 percent

44. On all of the policy questions, I omitted those respondents who said they "don't know" their positions. The results are essentially the same if I include these respondents in the analyses.

45. Half the respondents were asked this question. The other half of respondents were asked, "Which of these two things do you think is the more important for the United States to try to do—to keep out of war ourselves or to help England win, even at the risk of getting into the war?" The unweighted data indicated that 62.9 percent of respondents supported helping England, while the weighted analyses showed 64.3 percent supported that position. The distribution of sentiment by demographic characteristics is very similar to that on the "defeat Germany" question.

of women supported helping Germany, as compared with 65.5 percent of men; 61.5 percent of respondents with a grade-school education supported the war, as compared with 64.1 percent of respondents who had attended at least some college. Conversely, 72.8 percent of southerners took a hawkish stance, as compared with 60.1 percent of nonsoutherners. The countervailing tendencies of the sample imbalances therefore cancel out. In this particular example, the OPOR polls give us the right answer for the wrong reason.

These examples drawn from the OPOR data lead to a more general point concerning the impact of weighting on our inferences. Weighting will have the greatest impact on aggregate public opinion estimates when (1) the sample imbalances are related to the dependent variable of interest and (2) the relationships of the dependent variable to the variables on which there are sample imbalances work in the same direction (that is, there are no countervailing relationships). In the information items discussed above, these conditions hold. But on policy items concerning war, weighting often has little effect on aggregate estimates because the sample imbalances work in opposite directions. Thus, weighting might sometimes make a difference in our understanding of politically relevant quantities, and other times make little difference. Even in the case of support for war, however, the weighted opinion estimate is clearly preferable since it does not depend on such errors canceling one another out. Given our theoretical expectations, it is important to determine whether such adjustments change how we portray mass public sentiment during the 1930s and 1940s. Without such an investigation, we have no confidence that the aggregate measures of opinion from that time accurately reflect public opinion.

In sum, I recommend using weights to adjust the aggregate estimates of opinion from the 1930s and 1940s. The use of weights may not always significantly alter our estimates of public opinion, but using weights allows us to make valid statements about the opinions of the U.S. population, and—as the analysis of the information questions demonstrates—sometimes the use of weights can greatly change our picture of attitudes and capacities of the mass public. Though I prefer to use the cell weights, using different weighting schemes, as this section demonstrates, gives us similar pictures of public opinion. Researchers employing different weighting methods will come to similar pictures regarding the shape of public opinion from this era—conclusions that are sometimes different than those reached using the raw survey marginals. Thus, through the use of weights, we can have greater confidence that the inferences we draw about public opinion more accurately reflect underlying public sentiment.

INDIVIDUAL-LEVEL ANALYSIS

As a further demonstration of the utility of the methods advanced in this article, I conducted individual-level analysis of voting behavior using concurrent quota and area sampled data. In November 1948 the Survey Research Center (SRC) at the University of Michigan conducted a postelection survey using

area sampling methods.[46] At the same time, Gallup carried out a survey using the quota sampling methods typical of this time. It is therefore possible to directly compare the determinants of the decision to cast a vote to see if variables of interest have the same multivariate relationships to turnout in the two surveys.[47] Of particular interest are the effects of variables for which Gallup did not set strict quotas, but which are known to affect turnout; namely the respondent's age and education.

I recoded the variables to ensure comparability across the two data sets. Education is entered as a series of dummy variables indicating whether the respondent has a grade school education, a high school education, or a college education. Age was also measured by a series of dummy variables that quantified age in 10-year increments (e.g., 25–34, 35–44, . . . 65 and older).[48] Finally, I included the respondent's gender and—to control for the regional quota in the Gallup data—the regional location in which the interview was conducted.[49]

The results of the probit models are presented in table 5. To ease interpretation of the results, I present in figure 1 the predicted probability of voting at different age and education levels for an "average" respondent.[50] While the Gallup results predict a higher baseline turnout than the SRC study, the dynamics of the effects age and education are remarkably similar across the two studies.[51] These results therefore increase confidence that valid individual-level inferences can be made from the quota-sampled data.

46. According to the codebook, the sample was selected by area sampling. The 662 interviews came from 32 sample points—27 widely scattered counties plus 5 of the 12 largest metropolitan areas, each of which included several counties. In each sample point, several communities were selected. Within each sample city or town, sample blocks were stratified and selected at random. A random sample of dwelling units was taken from these blocks, and one person from each household was interviewed.

47. Unfortunately, there are no attitudinal variables that I can compare across the two data sets.

48. Calibrating the measures across the two surveys necessitated restricting the analysis to respondents who were 25 and older. The 1948 SRC study was originally intended to be a study of foreign policy attitudes, not an election study. As a result, some respondents who were under 21 were included in the sample. However, given that age was recorded only as a categorical variable (with the lowest category being "18–24"), it is not possible to separate out respondents who did not vote because they were ineligible to vote from those respondents who simply did not vote. It should be noted, however, that including all respondents in the analysis does not change the results reported below for the 25 years and older samples.

49. The SRC study did not record the state or region in which the interview took place. However, the comparison results do not change appreciably if I remove the region variables from the Gallup analysis.

50. The "average" respondent is one who is male, lives in the Northeast, has a grade school education, and is 35–44 years old.

51. There are two likely reasons for the higher baseline turnout in the Gallup study. First, while education allows us to account for some of the factors that led respondents to be selected by an interviewer—such as their political interest or generally approachability—it cannot control for all those factors. Thus, the Gallup sample is almost certainly a more politically engaged sample than the SRC sample. Second, there are important differences in the way the two polling operations asked their turnout question. The SRC asked, "In this election about half the people voted and half of them didn't. Did you vote?" Gallup, on the other hand, simply asked, "Did you happen to vote in the presidential election on November 2?" The gentler introduction used by SRC may reduce the amount of vote overreporting due to social desirability concerns (see Bernstein et al. 2001 for relevant discussion; Belli et al 1999 for counterpoint).

Table 5. 1948 Turnout: Survey Research Center (SRC) versus AIPO

	SRC		AIPO	
Variable	Coefficient	(SE)	Coefficient	(SE)
Constant	–0.13	(0.14)	0.10	(0.09)
Age: 35–44	0.18	(0.15)	0.37	(0.08)**
Age: 45–54	0.49	(0.17)**	0.52	(0.09)**
Age: 55–64	0.34	(0.18)*	0.52	(0.09)**
Age: 65+	0.13	(0.19)	0.32	(0.10)**
High school education	0.42	(0.12)**	0.36	(0.07)**
College education	0.62	(0.18)**	0.71	(0.09)**
Male	0.21	(0.11)*	0.28	(0.06)**
Region: Midwest			0.08	(0.07)
Region: South			–0.33	(0.09)**
Region: West			0.08	(0.09)
N/Log Likelihood	591/–362.32		2662/–1263.81	

* $p < .10$.
** $p < .05$.

Conclusion

In this article I advance methods to analyze survey data that have been largely ignored for over 50 years. Through the use of simple corrections to known sampling problems, the quota-controlled survey data of the 1930s and 1940s can be used to assess the origins and direction of mass public sentiment of some of the most important political questions of the twentieth century. Unlike data from the present day that may exhibit flaws similar to the polls collected before 1950—such as Internet polls that collect information from haphazard respondent pools—we cannot simply advise researchers to collect their data in ways that better conform to probability sampling. When dealing with the polls from the past, we can either try to ameliorate the systematic biases that arose from the sampling techniques of the time, or we can ignore this valuable data.

While the solutions advanced here may not fully correct the problems induced by the data collection process, by appropriately accounting and adjusting for demographic imbalances caused by faulty sample design, the data can yield great benefits. Given the public opinion data collected by Gallup, Roper, Cantril, and NORC available through the Roper Center for Public Opinion Research, it is possible to undertake a revolution in the field of political behavior. There are a number of critical questions that can be addressed by researchers given the available survey data and the set of methods described here to draw inferences from those data. By bringing modern tools and theories of political behavior to these data, we can explore many questions of great historical significance.

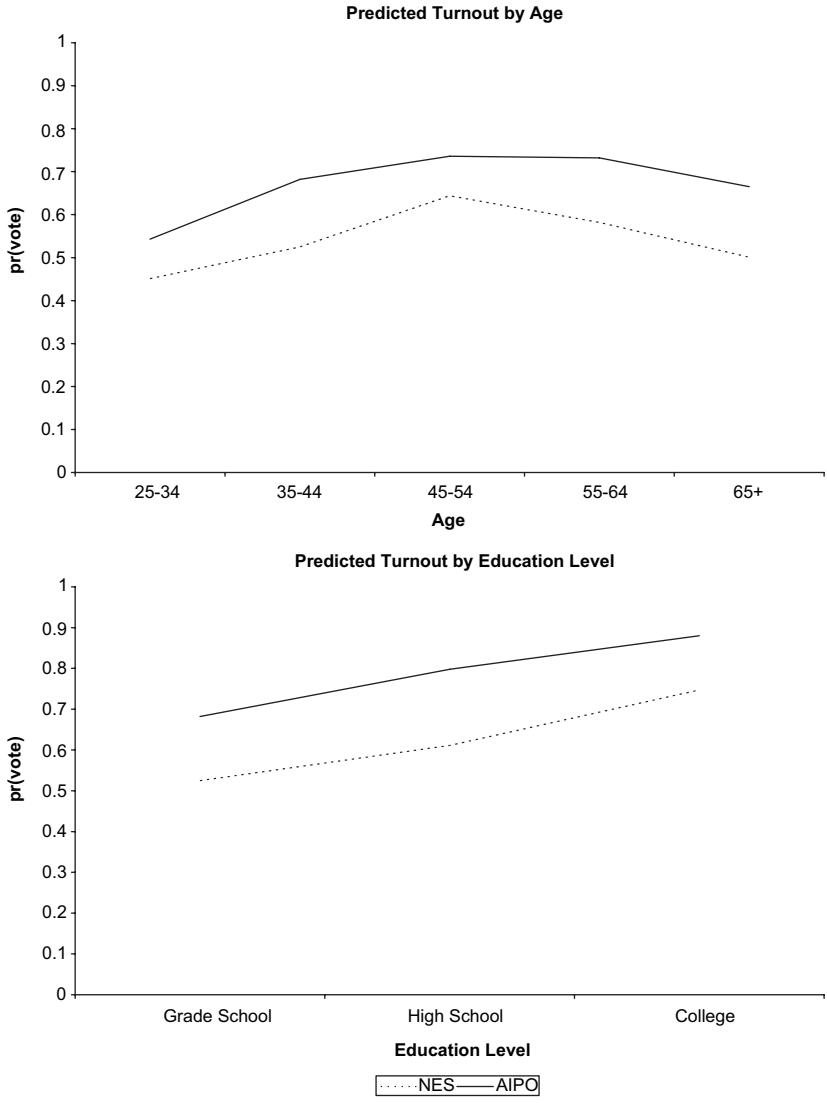

Figure 1. Effect of age and education on predicted turnout, 1948.

References

Anderson, Barbara A., and Brian D. Silver. 1986. "Measurement and Mismeasurement of the Validity of Self-Reported Vote." *American Journal of Political Science* 30:771–85.

Banks, Seymour, Norman C. Meier, and Cletus J. Burke. 1948. "Laboratory Tests of Sampling Techniques: Comment and Rejoinders." *Public Opinion Quarterly* 12(2):316–24.

Baum, Matthew A., and Samuel Kernell. 2001. "Economic Class and Popular Support for Franklin Roosevelt in War and Peace." *Public Opinion Quarterly* 65:198–229.
Belli, Robert F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63:90–108.
Berinsky, Adam J. 2004. *Silent Voices*. Princeton, NJ: Princeton University Press.
———. 2006. "Assuming the Costs of War: Events, Elites, and American Public Support for Military Conflict." Working Paper, Massachusetts Institute of Technology.
Bernstein, Robert, Anita Chadha, and Robert Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quarterly* 65:22–44.
Bershad, Max A., and Benjamin J. Tepping. 1969. "The Development of Household Sample Surveys." *Journal of the American Statistical Association* 64(328):1134–40.
Bethlehem, Jelke G. 2002. "Weighting Nonresponse Adjustments Based on Auxiliary Information." In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little, pp. 275–288. New York: Wiley.
Breen, Richard. 1996. *Regression Models: Censored, Sample Selected, or Truncated Data.* Thousand Oaks, CA: Sage.
Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8:389–98.
Bureau of the Census. 1976. *Historical Statistics of the United States: Colonial Times to 1970.* Washington, DC: U.S. Department of Commerce.
Caldeira, Gregory A. 1987. "Public Opinion and the U.S. Supreme Court: FDR's Court-Packing Plan." *American Political Science Review* 81:1139–53.
Campbell, Albert A. 1946. "Attitude Surveying in the Department of Agriculture." In *How to Conduct Consumer and Opinion Research: The Sampling Survey in Operation*, ed. Albert Blankenship, pp. 274–285. New York: Harper and Brothers.
Cantril, Hadley. 1948. "Opinion Trends in World War II: Some Guides to Interpretation." *Public Opinion Quarterly* 12(1):30–44.
Cantril, Hadley and Research Associates in the Office of Public Opinion Research, 1944. *Gauging Public Opinion*, Princeton, NJ: Princeton University Press.
Cantril, Hadley, and Mildred Strunk. 1951. *Public Opinion, 1935–1946.* Princeton, NJ: Princeton University Press.
Casey, Steven. 2001. *Cautious Crusade: Franklin D. Roosevelt, American Public Opinion, and the War against Nazi Germany*. New York: Oxford University Press.
Clausen, Aage R. 1969. "Response Validity: Vote Report." *Public Opinion Quarterly* 32:588–606.
Converse, Jean M. 1987. *Survey Research in the United States: Roots and Emergence, 1890–1960*. Berkeley: University of California Press.
Converse, Philip E. 1965. "Comment on: Public Opinion and the Outbreak of War." *Journal of Conflict Resolution* 9:330–33.
Deville, Jean Claude, and Carl-Erik Sarndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87:376–82.
Deville, Jean Claude, Carl-Erik Sarndal, and Oliver Sautory. 1993. "Generalizing Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88:1013–20.
Elliot, David. 1991. *Weighting for Non-Response: A Survey Researcher's Guide*. London: Office of Population Censuses and Surveys, Social Survey Division.
Erikson, Robert S., Michael B. MacKuen, and James A. Stimson. 2002. *The Macro Polity*. New York: Cambridge University Press.
"The *Fortune* Survey." 1940. *Journal of Educational Sociology* 14(4):250–53.
Gallup, George. 1944. *A Guide to Public Opinion Polls.* Princeton, NJ: Princeton University Press.
Gallup, George, and Saul Forbes Rae. 1940. The Pulse of Democracy, The Public Opinion Poll and How It Works, New York: Simon and Schuster.
Gelman, Andrew. 2005. "Struggles with Survey-Weighting and Regression Modeling." Unpublished manuscript, Department of Statistics and Department of Political Science, Columbia University.
Gelman, Andrew, and John B. Carlin. 2002. "Post-Stratification and Weighting Adjustments." In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little, pp. 289–302. New York: Wiley.

Glenn, Norval. 1975. "Trend Studies with Available Survey Data: Opportunities and Pitfalls." In *Survey Data for Trend Analysis*, ed. Jesse C. Southwick, pp. 6–48. Williamstown, MA: Roper Public Opinion Research Center in cooperation with the Social Science Research Council.
Gschwend, Thomas. 2005. "Analyzing Quota Sample Data and the Peer-Review Process." *French Politics*. 3(1):88–91.
Haner, Charles F., and Norman C. Meier. 1951. "The Adaptability of Area-Probability Sampling to Public Opinion Measurement." *Public Opinion Quarterly* 15(2):335–52.
Hansen, Morris H., and Philip M. Hauser. 1945. "Area Sampling: Some Principles of Design." *Public Opinion Quarterly* 9(2):183–93.
Hauser, Philip M., and Morris H. Hansen. 1946. "Sample Surveys in Census Work." In *How to Conduct Consumer and Opinion Research: The Sampling Survey in Operation*, ed. Albert Blankenship, pp. 251–258. New York: Harper and Brothers.
Hochstim, Joseph R., and Dilman M. K. Smith. 1948. "Area Sampling or Quota Control? Three Sampling Experiments." *Public Opinion Quarterly* 12(1):73–80.
Hogan, Michael J. 1997. "George Gallup and the Rhetoric of Scientific Democracy." C*ommunication Monographs* 64:161–79.
Holt, Robert D., and David Elliot. 1991. "Methods of Weighting for Unit Non-Response." *Statistician* 40(3):333–42.
Johnson, Palmer O. 1959. "Development of the Sample Survey as a Scientific Methodology." *Journal of Experimental Education* 27:167–76.
Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81–97.
Katz, Daniel. 1944. "The Polls and the 1944 Election." *Public Opinion Quarterly* 8:468–87.
Katz, Daniel and Hadley Cantril. 1937. "Public Opinion Polls' Sociometry. 1:155–179.
Kennedy, David M. 1999. *Freedom from Fear: The American People in Depression and War, 1929–1945*. New York: Oxford University Press.
Kish, Leslie. 1965. *Survey Sampling.* New York: John Wiley and Sons.
Little, Roderick J. A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88:1001–12.
Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data.* 2d ed. New York: Wiley.
Little, Roderick J. A., and Mei-Miau Wu. 1991. "Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ." *Journal of the American Statistical Association* 86:87–95.
Lohr, Sharon L. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Duxbury.
Lynn, Peter, and Robert Jowell. 1996. "How Might Opinion Polls Be Improved? The Case for Probability Sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 15(1):21–28.
Meier, Norman C., and Cletus J. Burke. 1947. "Laboratory Tests of Sampling Techniques." *Public Opinion Quarterly* 11(4):586–93.
Moore, David W. 1992. *The Superpollsters: How They Measure and Manipulate Public Opinion in America.* New York: Four Walls Eight Windows.
Moser, C. A., and A. Stuart. 1953. "An Experimental Study of Quota Sampling." *Journal of the Royal Statistical Society: Series A (General)* 116(4):349–405.
Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, and David B Truman, eds. 1949. *The Pre-Election Polls of 1948*. New York: Social Science Research Council.
National Opinion Research Center (NORC). 1943. *Interviewer's Manual*. Denver, CO: NORC.
Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97(4):558–625.
Noyes, Charles E., and Ernest R. Hilgard. 1946. "Surveys of Consumer Requirements." In *How to Conduct Consumer and Opinion Research: The Sampling Survey in Operation*, ed. Albert Blankenship, pp. 259–273. New York: Harper and Brothers.
Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago Press.
Robinson, Daniel J. 1999. *The Measure of Democracy: Polling, Market Research, and Public Life, 1930–1945*. Toronto: University of Toronto Press.

Roper, Elmo. 1940a. "Sampling Public Opinion." *Journal of the American Statistical Association* 35(210, part 1):325–34.
———. 1940b. "The Public Opinion Polls: Dr. Jekyll or Mr. Hyde? Classifying Respondents by Economic Status." *Public Opinion Quarterly* 4(2):270–72.
Ruggles, Steven, et al. 2004. Integrated Public Use Microdata Series: Version 3.0. Minneapolis, MN: Minnesota Population Center.
Schafer, Sarah. 1999. "To Politically Connect, and Profitably Collect." *Washington Post*, December 13, p. F10.
Schlozman, Kay Lehman, and Sidney Verba. 1979. *Injury to Insult: Unemployment, Class, and Political Response*. Cambridge, MA: Harvard University Press.
Sharot, T. 1986. "Weighting Survey Results." *Journal of the Market Research Society* 28:269–84.
Smith, Tom W. 1983. "On the Validity of Inferences from Non-Random Samples." *Journal of the Royal Statistical Society: Series A (General)* 146(4):394–403.
———. 1987. "The Art of Asking Questions, 1936–1985." *Public Opinion Quarterly* 51:S95–S108.
Stephan, Frederick F., and Philip J. McCarthy, eds. 1957. *Sampling Opinions.* New York: John Wiley and Sons.
Sudman, Seymour. 1966. "Probability Sampling with Quotas." *American Statistical Association Journal* 20:749–71.
Thompson, Steven K. 2002. *Sampling*. New York: John Wiley and Sons.
Traugott, Michael W., and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behavior." *Public Opinion Quarterly* 43:359–77.
Verba, Sidney, and Kay Lehman Schlozman. 1977. "Unemployment, Class Consciousness, and Radical Politics." *Journal of Politics* 39:291–323.
Voss, D. Steven, Andrew Gelman, and Gary King. 1995. "Pre-Election Survey Methodology: Details from Nine Polling Organizations, 1988 and 1992." *Public Opinion Quarterly* 59:98–132.
Warner, Lucien. 1939. "The Reliability of Public Opinion Surveys." *Public Opinion Quarterly* 3(3):376–90.
Weatherford, M. Stephen, and Boris Sergeyev. 2000. "Thinking about Economic Interests: Class and Recession in the New Deal." *Political Behavior* 22(4):311–39.
Wilks, S. S. 1940. "Representative Sampling and Poll Reliability." *Public Opinion Quarterly* 4(2):261–69.
Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23(2):230–57.
Worcester, Robert. 1996. "Political Polling: 95% Expertise and 5% Luck." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 15(1):5–20.