

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

Adam J. Berinsky
Assistant Professor
Massachusetts Institute of Technology
Department of Political Science

berinsky@mit.edu

April 28, 2004

Version 1.2

****DRAFT VERSION****

For Most Recent Version, See: <http://web.mit.edu/berinsky/www/QCS.pdf>

A previous version of this paper was presented at the Southern Political Science Association Meeting, January 8-11, 2004, New Orleans, LA. For valuable discussion and advice regarding this paper, I would like to thank Steve Ansolabehere, Jake Bowers, Bear Braumoeller, John Brehm, Andrew Gellman, Steve Heeringa, Sunshine Hillygus, Doug Kriner, Jim Lepkowski, Tali Mendelberg, 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. For financial support, I thank Princeton University and MIT.

The decade from 1935 to 1945 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. The relationship of public opinion to government policy during these years is of great importance. Fortunately, a great deal of data concerning the public's views during this time. 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 much information on questions of central importance to political scientists, historians, and policymakers.¹

It is somewhat surprising, then, that these data have largely been overlooked by modern researchers. One reason why scholars have ignored these data is because of the seemingly haphazard manner in which they were collected. From a modern standpoint, the data collection methods seem

¹ For one, these polls have the potential to change our understanding of public opinion concerning war. What we know about public opinion and war, we have learned from failed international interventions – such as Korea and Vietnam – and relatively short-term military excursions – such as Gulf Wars I and II, Somalia, and Afghanistan. Thus, quite paradoxically, systematic studies of the relationship between government and the mass public during wartime conducted in the last 40 years have overlooked the largest and most important international conflict in U.S. history. By attending to World War II, much can be learned about the relationship between the mass public and political elites in a democracy during wartime. These polls contain a plethora of questions concerning public opinion and World War II. For example, between the beginning of World War II in August 1939 and the entry of the U.S. into the war, Gallup repeatedly asked respondents if it was more important to help England or to stay out of the war. By examining these items closely, we can determine the changing structure of support for isolationism in the pre-war period. Furthermore, by comparing questions concerning support for the war from 1941 to 1945 to support for the Korean and Vietnam wars, we can determine why – unlike the wars in Indochina – the public supported the War effort in the 1940s despite mounting casualties. These data are also valuable for scholars of domestic politics. Opinions polls give us an important window into the role played by the mass public in ending the rush of New Deal programs in the late 1930s. For instance, Gallup asked respondents in January 1939, “Do you think the federal government is spending too much money for relief?” and in May, 1939, “Do you think that the government should do away with work relief, (such as the WPA job) and give only home relief?” In July, 1938 Roper asked, “On the whole, do you approve or disapprove of President Roosevelt’s wages and hours legislation?” and in March, 1939, “Do you think that WPA should be continued by the federal government on the same scale as it is now?” Furthermore, scholars interested in the development of racial attitudes can study questions relating to discrimination and policies of racial equality and can determine how the structure of attitudes concerning race have changed over the last 65 years. For example, a poll conducted by NORC in June 1942 contained the following items (among others): “Generally speaking, if a Negro has the same training as a white person, do you think he can do a particular job just as well as a white person?” “Do you think [your employer] should hire Negroes?” “Do you think white students and Negro students should go to the same schools or separate schools?” Another study, conducted by NORC in May 1944 asked policy questions similar to those in the 1942 survey, as well as questions relating to biological racism, such as: “As far as you know, is Negro blood the same as white blood, or is it different in some way?” and “In general, do you think Negroes are as intelligent as white people – that is, can they learn just as well if they are given the same education.” These questions have been used in studies of racial attitudes at the aggregate level by Schuman, et al. (1997). But to uncover changes in the structure of racial attitudes over time, closer attention must be paid to individual-level data from this era.

shoddy. Principles of random selection took a back seat to interviewer discretion and concerns about survey cost. But just because the data were collected in ways that now appear questionable does not mean scholars should ignore the early public opinion polls. Whatever their flaws, these polls provide insight into the beliefs of the mass public at this crucial time in American history.

In this paper I lay out a strategy to properly analyze the public opinion data of the 1930s and 1940. I first describe the quota-control methods of survey research prevalent during this time, contrasting these methods with survey methods used in the present day. I then detail the bias introduced through the use of quota-control techniques. Specifically, the polling methods of the early days of survey research introduced two types of errors: those arising from the establishment of the response quotas by the central office and those arising from interviewer discretion in respondent selection. Next, I describe specific strategies that researchers can employ to correct these problems in data analysis at both the individual and aggregate levels. The limitations of the data collection efforts of the early pollsters may be serious, but they are not insurmountable. The practices of the pollsters introduced certain predictable forms of bias into the survey results. We can therefore tailor our analytic techniques to account for this bias. At the individual level, to correct for the sample selection and interviewer-induced biases introduced by the quota-control sampling method, researchers should introduce statistical controls for: (1) the variables that made up the quota – such as gender, region, and occupation (2) the education level of the respondent, when available, and (3) a series of fixed effects for the interviewer number, where available. To correct for sample selection and interviewer-induced biases at the aggregate level, researchers should use post-stratification weights to bring the distribution of the variables in the sample in line with those in the census. Finally, I use examples from several public opinion studies in the early 1940s to show how the methods of analysis laid out in this paper can enable us to understand the shape and structure of public opinion during this time. With some simple corrections and straightforward analytic

techniques that are self-conscious of the limitations of the data collection, we can extract valuable information about the nature of public preferences during this critical era.

Quota Controlled Sampling: An Introduction

To obtain truly representative samples from populations, surveys should use simple random sampling (SRS), a procedure that ensures each and every element from the population has an equal chance of being selected to be in the sample.² In practice, however, it is infeasible to use such methods.³ Modern opinion poll sampling is therefore conducted using a modified probability method – such as random-digit dialing telephone interviewing or a face-to-face multi-stage area probability sampling design.⁴ While these methods do not meet the “gold standard” of 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).⁵

Such methods of random sampling may be the norm in the U.S. today, but these sampling schemes have not always reigned supreme in survey research.⁶ While probability sampling methods

² Formally, SRS is a sampling scheme with the property that any of the possible subsets of distinct elements from the population of N elements is equally likely to be in the chosen sample.

³ For instance, SRS requires the researcher to draw up a list of all observation units in the population, to be used as the sampling frame. But a full listing of the target population for a survey is rarely available.

⁴ For example, the National Elections Study (NES) is conducted using a multi-stage area probability design. Sampling proceeds through a series of nested stages, with probability methods employed in all stages of sample selection. First, the United States is stratified by geography and size within each of the four Census regions. One primary sampling unit is selected from each stratum using a controlled selection procedure. Within these primary areas, housing unit clusters are stratified by geography and size. Then, these clusters are sampled with the probability of selection proportionate to number of occupied housing units in the area., creating a list of area segments Next, the sampling staff lists all the housing units within each segment and selects a random subset from each segment. Once a household is selected, the interviewer lists all eligible adults residing in that household. In the final stage, to ensure that all people in the household have an equal probability of being chosen, the interviewer uses a selection table to choose randomly the respondent to be interviewed. While this procedure is not SRS – for example, since the number of eligible adults may vary from one household to another, the random selection of a single adult introduces inequality into respondents' selection probability – it approximates a random sample to a large degree.

⁵ In practice, most political scientists do not utilize the methods advanced by Kish to analyze survey data – such as adjusting for clustering – further blurring the distinction between modern methods of survey sampling and the looser methods used in the 1930s and 1940s. I thank John Brehm for making this observation.

⁶ Glenn (1975) reports that by the mid-1950s, all of the major survey organizations had moved to using modified probability samples that were designed to represent the total adult non-institutionalized civilian population. A notable exception to this practice in the modern day is some forms of internet polling. In addition, in Europe many polls use

were well established in government agencies, such as the Census Bureau and the Works Progress Administration, by the late 1930s (Converse 1987), opinion polls in the U.S. before the 1950s were conducted using an entirely different methodology – namely quota-controlled sampling methods.⁷

Pollsters employed this sampling method both for commercial polls – such as those conducted by Roper and Gallup – and public-interest polls – such as 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 certain predetermined proportions of people from particular segments of the population. The researcher controlled this process by dividing the population into a number of mutually exclusive

modified quota-sampling methods to collect data. Many of the arguments that occurred in the surveying community in the 1930s and 1940s in the United States continue abroad (see, for instance, the discussion of polling during the 1992 British election in Lynn and Jowell (1996) and Worcester (1996)). It should be noted, however, that these quota-sampling methods avoid the most egregious errors of the quota-control sample polls discussed below.

⁷ Quota control sampling was used almost exclusively by survey research firms until 1950. Writing in the late 1940s, Meier and Burke (1947-8) identified only one organization that used the probability sampling method – the *Washington Post* – and even then area sampling was used only on an experimental basis. At the same time, there were 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-control sampling. Academic 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 are unknown, the estimates of sampling error of quota samples that were based on sampling theory would be incorrect (Hansen and Hauser 1945; Berhad and Tepping 1969). 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, 1946, 1948; Johnson 1959). Supporters of quota-controlled sampling, on the other hand, were almost all professional pollsters. Though much of their argument for quota-controlled sampling centered on cost, many of these researchers did not trust that area sampling would collect a representative population (see below for more discussion). At the time, this debate became quite heated (see, for example, the exchange between Banks on the one hand and Meier and Burke on the other in “Laboratory Tests of Sampling Techniques: Comment and Rejoinders” (1948). The debate was not simply one of “academics” versus “industry.” As Converse 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” (1987, 126). Some academics even came to the defense of quota-control sampling (Wilks 1940). The debate over quota-control samples was never completely settled. But the fallout from the 1948 election – in which all the major polls that used quota sampling predicted a Dewey victory – hastened the demise of quota-control samples. For example, 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 SSRC study of the election did not place the blame for the failure of the pre-election polls primarily on quota-sample methods. As the investigators conclude, “the use of probability sample will not in any way guarantee that one can predict elections. Their use can only remove one of the many potentially serious sources of inaccuracy. As is shown elsewhere in this report, there are many other sources of error associated with such an operation and these may far outweigh the sampling aspects of the problem” (1949, 117).

sub-populations, or strata – such as geographic region or age groupings – and then allocating the sample among these strata in proportion to their desired size (Stephan and McCarthy 1957).⁸

The goal of this method was ostensibly to obtain a purely descriptively representative sample of citizens in the United States. Thus, the quotas were set to represent the population on the quantities the survey researchers thought would capture politically relevant subgroups, such as gender, age, and occupation.⁹ As Elmo Roper described his enterprise, “Our purpose is to set up an America in microcosm. We want to have each constituent element of the entire population represented in its proper proportion in the sample” (1940a, 326).¹⁰

This desire to achieve representative samples through strict controls represents the largest point of departure between the pollsters of the 1930s and 1940s – such as Gallup and Roper – and modern survey researchers. Gallup and Roper did not believe in the primacy of the law of large numbers – they simply did not trust that chance would ensure that their sample would accurately represent the sentiment of the nation.¹¹ 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 truly representative sample. A quota-controlled

⁸ Specifically, quota-control sampling proceeded in three steps. First, the surveyor apportioned the sample to geographic regions. Next, he selected specific geographic 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. As one contemporaneous observer described the procedure: “what is done is to select several hundred fairly small sampling areas (e.g. cities, counties, etc.) over the United States which may be regarded as a representative sampling by areas of major geographic districts and allocate the sampling to those areas in such a way that the portion of the sample drawn from each area is representative with respect to age, sex, color, and economic status within that area” (Wilks, 1940, p. 263).

⁹ For example, Roper created quotas based, in part, on a division of the sample into respondents who were under 40, and respondents who were 40 and older. In Roper’s view, the use of this dividing line was critical because, “The people over forty are more conservative, less apt to look with favor on a new idea or an innovation, and are usually found either more reconciled to their fate in life or more apt to express their irritation at life in railing at the sexual or drinking habits of the young than in railing at any real or fancied inequalities in our economic system” (1940a, 327).

¹⁰ Similarly, the NORC interviewer’s manual explained that “all public opinion research is based on the proven fact that the views of a comparatively small number of persons accurately reflect the opinions of all the people – provided that the sample of persons interviewed represents an exact miniature of the total population” (1944, 57).

¹¹ To use the words of one researcher, they believed that, “because our sample is not large, it must be exact” (Williams, 1942, p. 635).

sample – and only a quota-controlled sample – was the microcosm of America desired by the pollsters.¹²

But the tight control 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. Thus, once in the field, interviewers were given wide discretion in selecting the particular people they chose to interview.¹³ Some interviews took place in people’s houses; others were conducted on street corners. 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.¹⁴

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.¹⁵ This chart represents the compilation of sometimes contradictory information from a variety of sources,

¹² 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 – such as farmers and city people – 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, 98). Gallup was not alone in his belief of the primacy of quota-controlled samples. An excerpt from the NORC interviewer’s manual also underscored the importance of control:

If a universe which is to be sampled is homogeneous, the best procedure is random sampling – a method by which chance alone determines who or what is to be included. People, however, are not homogenous in their characteristics and factors such as age, sex, education, occupation, and income all influence opinion. Therefore a random sample is not satisfactory, and some controls are necessary to assure that most of these factors are present in the same proportion that they exist in the total population. (1944, 58)

¹³ The arbitrary nature of respondent selection is made clear by *The Journal of Educational Psychology*’s description of Roper’s sampling procedures. The journal noted that “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” (1940, 252).

¹⁴ The reduction of interviewer discretion is the largest change in quota sampling procedures from the 1930s to the present. Firms that use quota sampling methods today give strict instructions regarding quasi-random procedures for respondent selection.

¹⁵ Converse (1987) provides an excellent history of the survey research enterprise, for example, but 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.”

because there is no single source for this information. Even so, the chart contains a great deal of information concerning the surveying practices of the time. Clearly, there was a great deal of variation across the organization in their specific polling practices. But there was also a great deal of common ground. All the survey organizations set quotas on both the geographic region and the size of the place where interviewing occurred. In addition, all organizations sought to control the distribution of age, gender, and economic status through the imposition of quotas.¹⁶

Quota Control Sampling: Problems and Concerns

Whatever its appeal in theory, in practice the quota-control sampling practices used by the major survey research firms led to highly biased samples of survey respondents. Two main classes of problems emerged from these sampling procedures: first, the sample was skewed relative to the population because the survey researcher allocated the sample to the different quota categories in particular ways. Second, the discretion given to interviewers in filling their quotas distorted the survey results.

Skewed Samples: Representing Perceived Voters, not Citizens

Contrary to their populist rhetoric, not all pollsters were interested in obtaining descriptively representative samples of Americans. In particular, Gallup designed quotas to produce sample proportions that differed systematically from the population. Because Gallup's livelihood depended

¹⁶ 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's classification – which he termed a "scale of living" – contained 4 economic levels that he conceived as a "sliding scale" (Roper also used a fifth classification for blacks). Roper concludes that "We have found this rather arbitrary definition of income levels fairly satisfactory over a six-year period" (1940b, 272). Similarly, Gallup classified his respondents into 6 categories, ranging from "wealthy" to "on relief." The polling firms set their quotas in relation to these classifications, in a somewhat arbitrary manner. For instance, Roper set his highest group – the "A"s at 7 percent of the sample, and his lowest group – the "D"s at 23 percent of the population. But because the economic indices were estimated by the individual interviewers and were not comparable with census figures, these classifications were by definition arbitrary. These measures were dropped by the survey firms by the 1950s, as the researchers decided that measures of education and income more accurately captured the social class of the respondents (Smith 1987).

on his ability to successfully predict elections, he 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.¹⁷

Southerners, blacks, and women turned out 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.¹⁸ 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 U.S. lived in the South, Gallup drew only 10 to 15 percent of his sample from that region. Figure 1 demonstrates the imbalances in the gender and region of the survey organizations over time.¹⁹

These same biases can be found in the OPOR samples in the early part of the war, when Gallup did the fieldwork for the surveys Cantril designed.²⁰ For example, the 1940 Census found that 50 percent of the U.S. population was female, 10 percent was black, and 31 percent lived in the South. By contrast, the sample of a December, 1940 OPOR poll was only 34 percent female, 3

¹⁷ The deliberate unrepresentativeness of Gallup’s polls is ironic given his numerous academic and non-academic writings trumpeting the ability of opinion polls to reveal the collective voice of all Americans (see Robinson 1999 for discussion of this point).

¹⁸ 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. 233-4).

¹⁹ Another issue that arises in the quota samples is that 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. As Cantril described the practice, “This procedure does not mean, of course, that an attempt is made to distribute one variable properly within another and that within a third variable. That is, the interviewer is not expected to get the proper age distribution among white people in the average income group” (1944, 140). 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).

²⁰ After 1942, Cantril created his own survey field organization.

percent black, and 13 percent Southern.²¹ However, not all survey organizations collected such skewed samples. Roper seems to have been more interested than Gallup in drawing samples to conform to Census population figures. Thus Roper's samples contain the population proportion of Blacks, women, and southerners. This is not to say that Roper's samples are without problems. As discussed below, these samples suffer from bias in non-quota category characteristics. But the important point to note here is that the Gallup and OPOR data that scholars of public opinion have used to represent the political voice of the mass public, in fact, comes from a skewed sample of that public – and a sample that is skewed by design.²²

Unusual Respondents: The Consequence of Interviewer Discretion

As troubling as these deliberately induced sample imbalances may be, the practice of quota sampling also introduced a number of unintended biases. 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.²⁴ Apart from having to fulfill their demographic quotas, interviewers were given great discretion to select particular citizens to interview. In addition, interviewers could simply replace citizens who refused to be interviewed with another respondent who met the requirements of the quota.²⁵ The key problem

²¹ These figures are typical of the OPOR polls I have examined in 1940 and 1941, indicating that the non-representative proportions were deliberate. It should be noted that the OPOR polls (but not the Gallup polls) have separate files that contain an oversample of southern respondents (half of whom are black). However, publicly available reports – such as Cantril and Strunk's 1951 compilation of opinion polls – do not make use of these oversamples.

²² For uses of this data as “representative” of public opinion, see for example, Kennedy, who regularly equates opinion polls from this period with the voice of the American public. For instance, he writes, “By late 1939, few could doubt where American sympathies lay. The mind and conscience of America were decidedly anti-Hitler. A Gallup poll in October found that 84 percent of respondents were pro-Ally and only 2 percent pro-German” (1999, p. 427). For other writers who equate opinion polls from this period with mass public sentiment, see also Casey 2001; Divine 1979; Doenecke and Wilz 1991; Leigh 1976.

²⁴ For this reason, the quota-controlled sample surveys of this period may properly be called “community intercept samples” (Smith 1987).

²⁵ 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 (Noyes and Hilgard 1946; Campbell 1946; for a discussion of the unit non-response problem in the present day, see Brehm 1993).

with this procedure, as Kish notes, is that, unlike probability samples, “selection within the quotas is not randomized selection from frames within strata, but is directed by the judgment of the interviewers” (1965, 564).²⁶ Since interviewers preferred to work in safer areas and tended to question approachable respondents, the “public” they interviewed often differed markedly from the public writ large in important ways (Glenn 1975; Converse 1987). Put another way, while interviewers may have collected a descriptively representative sample of the population with respect to the quota-controlled characteristics, in other respects that sample may not have accurately represented the population.

There is some evidence that the samples of the 1930s and 1940s were skewed toward the upper classes (see Cantril 1944), but the difference between the sample of survey respondents and the mass public writ large 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.²⁷ 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 a college education.²⁸ This skew is almost certainly a result of the fact that education was not a quota controlled category, and better educated respondents were more willing to be interviewed than citizens with less

²⁶ The irony was that the biases created by quota control sampling were induced by the improper exertion of control by the survey researchers. Gallup, Roper, and their peers exercised excessive control over the balance of cases in their sample, but only minimal control over their interviewers. It was this uneven application of control that created the tension that led to the highly biased samples of the 1930s and 1940s. If the pollsters had chosen the opposite strategy, the samples almost certainly would have been less biased than they turned out to be

²⁷ Gallup himself was aware of the problems created by the unequal distribution of education within his sample and, by 1948, he weighted his sample by education to reflect the census estimates of education (Mosteller et al. 1949).

²⁸ This skew in the education levels of the sample is typical of the Gallup and OPOR polls I have examined. Robinson (1999) reports that the under-representation of those with less than a high school education in the Gallup polls continued through the 1940s and into the 1950s. 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 towards those with more education.

education.²⁹ A contemporaneous observer, Rugg argued that the inflation of education levels was largely due to “the tendency on the part of interviewers to select, within each economic category, the more articulate and hence usually better educated respondents” (in Cantril 1944, 148).³⁰

The problems of an overeducated sample were not just the concern of Gallup. As Figure 2 demonstrates, in every case in which education was measured by a survey organization in the 1930s and 1940s, the distribution of that variable was tilted toward those with a college education.³¹ Clearly, the skew on education was a fundamental problem for survey researchers before 1950.³²

In addition to attracting particular types of respondents who may not have been fully representative of the population, the discretion given interviewers may have created other sources of error in surveys conducted before the 1950s. 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. Paul Sheatsley, who conducted interviews for Gallup in 1937, demonstrates a series of pathologies when describing his respondent selection strategy:

“The interview was very simple...I remember in those days the way I would fill my relief quota was to walk around town until I saw a WPA construction gang and I would get them on their lunch hour, three or four

²⁹ 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. While both the quota and probability samples overestimated education levels relative to the census measures, the quota sample contained fewer grade-school educated respondents and more college graduates than the area sample survey (Reported in Stephan and McCarthy 1957)

³⁰ Rugg also argues that some of the skew in education was due to an upward bias in the distribution of the economic status of the sampled cases relative to the population. However, he places greater weight on the non-random selection of respondents as an explanation for the skew.

³¹ Not surprisingly, some polling agencies were aware of these problems and tried to correct them. The manual prepared by NORC to train its interviewers specifically pointed to problems regarding the distribution of education: “Only about 10 per cent of the people in the United States ever attended college...Most polling agencies find their interviewers consistently including in their quotas a few too many college graduates and not enough persons with grade school education or less. This consistent bias may well be caused by the unconscious selection by interviewers of too many persons who look pleasant or ‘intelligent.’ To avoid biases of this sort, you will have to steel yourself against turning away from a person just because he looks ill-tempered or ignorant” (1944, 77-78). Even with this admonition, NORC’s interviewers continued to return samples of respondents with higher levels of education than the census estimates.

³² The simple solution to correct this education skew would be to put a quota on education levels. It is not clear why the polling firms did not establish such quotas. Perhaps the researchers were concerned with the high costs of obtaining interviews from those citizens not inclined to participate in the survey.

men sitting around eating their sandwiches and drinking their beer. I'd pull out my questionnaire and say, "do you approve or disapprove of a treaty with Germany" or whatever it was, and then I'd say, 'How about you, and you, and you?" I got four interviews very quickly that way. I'd go to parks, good places to find people ... You couldn't find many A-level people this way, so you'd have to screw up your courage and go through a fancy part of town and try to figure out which house looked the most approachable" (Converse, 1987, 126).

The extreme practices detailed by Sheatsley almost certainly introduced significant bias. For instance, citizens interviewed in the company of their friends may have felt obliged to give similar responses, leading to correlations between respondents. But it is important not to make too much of Sheatsley's story; evidence suggests that few interviewers engaged in such egregious behavior. The major polling companies actively discouraged their interviewers from these practices. For instance, NORC instructed interviewers not to attempt interviews "if you can't talk with a person alone" (1946, 84) and 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" (1943, 646).³³

Though the practice of multiple interviews may not have been a serious problem, Sheatsley's attempt to cut corners to fill his interview assignments underscores an important point. Because interviewers were given great discretion in selecting respondents, the potential for interviewer effects in survey responses is potentially greater than that found in modern survey research; interviewers could have a strong influence over the data collection process at the respondent selection stage as well as during the interview process.³⁴ Any systematic differences between interviewers in

³³ There was also some concern about interviewers making up responses to particular questions, or to entire interviews (see, for example, Crespi 1945-6).

³⁴ The question of standardized interviewing practices is of special concern in the early days of polling. It took some time for the survey organizations to build up their team of interviewers. Consider the experiences of Elmo Roper's organization. In 1951, 41 percent of the interviewers had at least 5 years of experience, and only 10 percent had less than 1 year of experience. But in 1948 half as many interviewers – 19 percent – had over 5 years of experience, and 25 percent had less than one year (Anderson 1952). While comparable numbers do not exist for the late 1930s and early 1940s, if we extrapolate these numbers back to the late 1930s and early 1940s, the average experiences of the survey interviewers would almost certainly be quite minimal.

respondent selection style or interviewing practices could therefore be reflected in the survey responses collected by the polling agencies.³⁵

Summary: The Nature of Quota Controlled Survey Data

Because of the existence of these sources of potential error, modern survey researchers view survey data from the 1930s and 1940s with great suspicion.³⁶ Political scientists who are aware of the limitations with polls conducted before 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).³⁷ Such criticisms may be valid – clearly the early opinion polls have a number of substantial flaws. But just because the polls are flawed does not mean that the critical information those polls contain concerning important events in American history should be abandoned. Instead, we should recognize that this data is valuable, but was collected in a manner that was less than ideal. In the next section, I advance some simple methods to draw the best inferences we can from the data.

A Method of Analysis

While the quota-control survey data from the 1930s and 1940s were collected in ways that appear from a modern vantage point to be haphazard, the errors of the data collection process

³⁵ For instance, Cantril’s analysis of an October 1940 AIPO poll found that survey interviewers collected opinions on the war that were, on balance, more in line with their own beliefs. Interviewers who supported helping England were more likely to collect responses that favored aiding the Allied war cause, while those who opposed international involvement collected responses that favored staying out of the war. 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 “an interviewer is probably better known by his respondents in a small town than in a large city. As a result, he may consciously or unconsciously choose respondents who think as he does himself, or, at least, with whom he is acquainted and who may be most likely to share his own views” (1944, p. 112).

³⁶ The practice of quota-controlled sampling has almost completely disappeared in the U.S. As Sudman succinctly put it, “the advocates of probability sampling met and defeated the defenders of quota sampling” (1966, p. 749).

³⁷ The rejection of these polls by political scientists contrasts with the work of historians, who treat these same polls as the true “voice of the people” of the 1930s and 1940s (See, for example, Casey 2001; Divine 1979; Doenecke and Wilz 1991; Kennedy 1999; Leigh 1976).

introduced certain *predictable* forms of bias into the survey results.³⁸ Because this bias is predictable, we can employ particular methods to correct for the bias. Thus, the key to any analysis of public opinion data from the 1930s and 1940s is to first recognize the sources of bias by carefully considering the sampling practices of the different survey organizations and then employing methods that directly account for these different sources of error. While such a strategy may not perfectly correct the problems with the data, identifying and correcting for predictable forms of systematic bias will allow researchers to best utilize the valuable opinion poll data.

The first step in any analysis is to classify the errors introduced by quota control sampling into two types: “systematic sample selection bias” and “systematic interviewer-induced bias.” Sample selection bias arises through instruction to the field staff from the central survey offices. I consider these errors “systematic sample selection” errors because the collector of the data deliberately introduced the bias in the sample. For instance, as noted above, Gallup instructed his interviewers to bring back samples that were skewed away from women and southerners.

Systematic interviewer-induced bias, on the other hand, is the consequence of the extraordinary interviewer discretion in respondent selection. Implicitly, quota-controls assume that within quota categories “the people that you pick are as good as if they were from a probability sample.” (Leslie Kish, quoted in Frankel and King 1995, p. 80). Such a strategy presumes that the relevant differences between those interviewed and those passed by are captured in the quota categories. But from descriptions of the survey practices of polling’s early days, we know that this assumption is not tenable. To the extent that respondents were selected because they were more “approachable” or lived in safer areas, systematic bias is introduced into the data. The key to interviewer-induced bias is that the respondent selection mechanism used by interviewers ensured

³⁸ I use the term “bias” here in the manner of Groves (1989) in the context of survey research. Groves argues that bias is “the part of the error common to all implementations of a survey design” (1989, 8). It could be argued that inter-interviewer differences in surveying practice could lead to errors of variance, as well as bias, but for the purposes of this paper, I am most interested in sources of bias in surveys.

that the citizens who were interviewed differed in systematic ways from citizens who were not interviewed.

Having identified the sources of bias, researchers should employ methods that directly account for and correct for the bias. The specific methods chosen depend greatly on the level of analysis chosen by the researcher; both individual-level regression-style analysis and aggregate analysis of public opinion demand particular methods of analysis.

Individual-Level Analysis

If we are interested in identifying the predictors of a particular dependent variable in regression-type analysis, we can account for both sample selection and interviewer-induced bias through the use of statistical controls. Consider first the problem of systematic sample selection error. For many of the surveys from this time, we have a sample that does not represent certain groups in proportion to their weight in the population. In the case of the Gallup polls, for example, each survey has too many men relative to the population. If men and women hold different positions on the items measured on the survey, our regression estimates will be biased. The intuition here is similar to the problem of omitted variable bias. In the surveys of the 1930s and 1940s, the population of respondents who were interviewed was systematically different from the population who was not interviewed in ways that were determined by the quotas imposed by the survey organizations. Thus, the probability of inclusion in the sample is conditional on the quota variable. As a result, the independent and dependent variables of interest may be correlated with the quota variables. To avoid this bias, we should include the quota category variables as statistical controls, which will allow us to parse out that portion of the common variance between the independent and dependent variables of interest that is correlated through the quota variables. Thus, if the quota variables are introduced as control variables into analysis, the regression estimates should be less biased than if we did not introduce the controls. Gelman and Carlin counsel precisely this strategy,

arguing that the best strategy to account for known differences between the sample and the population in regression analysis is to “include all the information used in the survey weighting as additional covariates in the regression and perform an unweighted regression” (2002, 290).³⁹ In effect, by including all the quota variables as control variables in multivariate analysis, we can guard against the possibility of omitted variable bias.⁴⁰

Controlling for the quota variables in statistical analysis will not solve all the problems of analysis. The interviewer-induced bias introduced by the surveying practices of the time must also be accounted for. Here the problem is more difficult. Interviewer discretion ensured that the respondents who were in the sample might be systematically different than respondents who were not in the sample. In effect, the survey practices of the 1930s and 1940s created a non-random assignment problem. The probability of being interviewed depended on certain characteristics of the respondent that interviewers found attractive. As a result, not all citizens had an equal chance of being interviewed.

Such differences could have important implications for analysis. We know, for example, from descriptions of interviewer practices that respondents who were interviewed by Gallup, Roper, and NORC were more approachable than respondents who were not interviewed. It is reasonable to

³⁹ To be more precise, Gelman and Carlin state that if a researcher is interested in obtaining an estimate of the predictive effect of variable U on a dependent variable Y , the information used to construct poststratification weights can be incorporated directly into the model through the use of statistical controls. Thus, Gelman and Carlin counsel “regressing Y on (U, X) where X represents the variables used in any weighting and poststratification scheme. If the probability of inclusion is constant or units in the sample are conditional on these X variables, any analysis that conditions on X [in this case, a regression of Y on (U, X)] would yield valid inferences without any need for weighting in the estimates” (2002, 296). Gelman and Carlin counsel the inclusion of the weighting variables in the regression rather than ignoring these variables completely so as to avoid model misspecification. The caveat implicit in Gelman and Carlin’s advice – that X captures all important differences between the sample and the population – is an important warning. Such an assumption can be reasonably made in the context of systematic sample selection bias, but might be more tenuous in the case of interviewer-induced bias. Even in the second case, however, the researcher is well served by making use of as much information concerning the relationship of the sample to the population as possible.

⁴⁰ If either the dependent variable or the independent variable of interest is uncorrelated with the quota variables, we do not need to include the quota variable in the regression to obtain unbiased estimates of the coefficient on the independent variable of interest. However, because we are not always certain of the relationship among the variables it may be advantageous to trade off the loss in efficiency caused by including irrelevant variables to ensure that we do not have omitted variable bias.

presume that these “approachable” respondents were especially interested in and knowledgeable of politics.⁴¹ If we were to estimate the effect of media use on levels of political information through regression analysis, such analysis would result in a biased coefficient on media use because our analyses would be tainted by omitted variable bias. Specifically, the analysis would omit a variable – “approachability” – that reflects interest in politics. Interest in politics, in turn, is positively correlated with both media use and levels of political information. The coefficient on media use would therefore be attenuated toward zero.

Correcting for interviewer-induced bias in our analyses is especially difficult because we do not have information on the people who were not interviewed. We know that the set of citizens who acceded to be interviewed differed from those who did not, but we have no way to account for this non-random assignment to interview through selection bias techniques (Brehm 1993). However, we do sometimes have important information that can be employed in analysis – namely the education level of the respondent and the identity of the interviewer.

Just as the bias created by quota-control sampling methods can be parsed in sample selection and interviewer-induced bias, the interviewer-induced bias in the polls can be separated into unmeasured and measured effects. At first glance, it might appear that we cannot measure any of the effects of bias. We know, for instance, that survey interviewers tended to traffic in the safer parts of towns and cities. But we do not know just where they went to procure their interviews. We also know that interviewers preferred those citizens who were more receptive to being interviewed, and those who were more interested in politics. However, if we do not have information on the non-respondents, we cannot directly measure the nature of the bias.

⁴¹ We see the same phenomenon in the present day when considering non-response on opinion polls. Brehm (1993) finds that many probability sample surveys contain a disproportionate number of politically engaged respondents. Brehm’s finding for instance, explains why the reported turnout rates in surveys are greater than official measures of aggregate turnout. The problem here is that, unlike Brehm, we do not have any information on the non-respondents, so we cannot employ the selection bias corrections he advocates.

We do, however, have some important information concerning the characteristics of the citizens the interviewers selected to fill their quotas. Specifically, the measures of education on the surveys allow us to partially control for the interviewer-induced bias. 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 make an interviewer more likely to pick one individual from a given demographic group than another. The key for the purposes of analysis 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.⁴² In effect, the measure of education is a trace of the systematic interviewer-induced bias in the data. Therefore, by controlling for education – in addition to including controls for the quota variables – we may be able to control for at least some of the interviewer-induced bias in the sample.⁴³ This solution is admittedly imperfect. It is possible, if not likely, that the low-education respondents selected by interviewers may not be fully representative of the population of low education citizens. But while correcting the samples for the unequal distribution of education may be a problematic fix to the interviewer induced bias problem, without information on the non-respondents, this strategy is the best solution available to the modern researcher. After all, controlling for some of the interviewer induced bias through the use of proxy variables, such as education, is preferable to completely ignoring the bias. We might not be able to eliminate the bias from the regression estimates, but

⁴² These assumptions are not perfectly met by the data. While education was not used as a quota variable during this time, the distribution of economic class was controlled by the survey organizations. To the extent that economic class and education are correlated, one could argue that education was indirectly quotaed through the economic class measure. In practice, such a concern is not justified by the data. While education and class are not orthogonal, the correlation between the two measures is not particularly high. For example, in a OPOR survey from January 1941, a four category measure of education (grade school, some high school, high school graduate, some college or more) is correlated at 0.27 with the four-category measure of class used by Baum and Kernel (2001) (wealthy, average, poor, on relief). Similarly, the correlation between these two variables in a March 1941 OPOR survey is 0.31. Thus, while education and class are related, there is no evidence that either variable is a proxy for the other.

⁴³ It might also be advisable to adjust for income and occupation. As Glenn (1975) notes, however, such adjustments are usually not possible, since the early surveys did not code economic level in a manner that makes comparisons with the census data feasible.

introducing the proxy variable controls will likely reduce that bias.

Another potentially significant source of interviewer-induced bias is inter-individual differences in interviewing practices. Survey interviewers were given almost complete discretion to select respondents; we might presume that there was some variation in the procedures they actually followed. The work done on this question in the 1940s and 1950s is ambiguous (Cantril 1948; Hyman et al. 1956). But absent strong evidence to the contrary, it is wise to control, when possible, for systematic sources of inter-interviewer variation in survey practices. On certain surveys, this strategy is feasible. Some polling organizations recorded interviewer numbers with each respondent. Thus, we can tell which respondents were questioned by the same interviewer. In most cases, we know nothing more about the interviewers than their numerical identifier. But by introducing interviewer-specific dummy variables – interviewer fixed effects – into our analyses, we can capture inter-interviewer variation in undesirable practices.⁴⁴ To the extent that large differences exist and that we have a way to capture the residual of those effects, we can correct for this source of interviewer-induced bias in our analysis.

In sum, at the individual level, to correct for the sample selection and interviewer-induced biases introduced by the quota-control sampling method, researchers should introduce statistical controls for: (1) the variables that made up the quota – such as gender, region, and occupation (2) the education level of the respondent, when available, and (3) a series of fixed effects for the interviewer number, where available. This solution is admittedly imperfect. But while we might not get the “truth” by accounting for these differences in our individual-level analysis, we certainly can learn valuable information about the determinants of opinion during this time. Thus, the strategy proposed here is a far better solution than the alternatives adopted by previous researchers who, have either ignored the problem or discarded the data altogether as worthless.

⁴⁴ An alternative strategy is to model these interviewer effects using Hierarchical Linear Modeling (Byrk and Raudenush 2001).

Aggregate Analysis.

In addition to individual-level analysis, researchers are often interested in the balance of aggregate public opinion on major policy questions. In fact, to date, most of the work done using public opinion polls from the 1930s and 1940s has used aggregate opinion measures. There is reason to think that the uncorrected measures of opinion misrepresent underlying public sentiment during this time. 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.⁴⁵ 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 citizenry. Such polls may represent the “voting public” to a certain degree but the existing aggregate measures – such 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 non-voters. Gallup simply created his representation of the voting public at the sampling stage, rather than 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 (for a more general statement of the importance of looking at

⁴⁵ A caveat is in order here. Southern Democrats tended to be more conservative than their northern counterparts. For instance, they were less supportive of FDR and his policies than other Democrats. Thus, even though the poll results might not be as “Democratic” as they otherwise would be if a true population sample were surveyed, the ideological differences between Democrats in the south and Democrats in the rest of the country may minimize the potential misrepresentation of aggregate opinion. I thank Doug Kriner for this point..

public opinion, defined broadly, see Berinsky 2004). Second, even if we believe that we should represent only the engaged public, Gallup's relatively poor 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 pre-election poll, most famously in the 1948 Presidential election (Mosteller et al. 1949). Thus, accepting the Gallup sample as the “public” seems problematic.

In order to correct the systematic sample selection bias in quota controlled samples, a series of poststratification weights can be employed to bring the sample proportion in line with census estimates of the composition of the population.⁴⁶ Poststratification is a form of weighting adjustment that reduces discrepancies between a sample survey and the population. To employ poststratification weights, the sample is stratified into a number of cells, based on the characteristics of the population the researcher deems important. For example, a researcher interested in studying opinion concerning religious beliefs in the 1940s might stratify her sample by gender and age, creating an estimate of religious sentiment for each of 4 groups; older men, older women, younger men, and younger women. If the distribution of demographic variables in the sample differs from the distribution on the population, poststratification weights can be used to combine the separate cell estimates into a population estimate. Specifically, the researcher can weight her estimates for each cell proportional to the number of units in the cell within the population, divided by the number of units in the sample in this stratum. In this way, the researcher can use information about the distribution of demographic characteristics – gained, for example, from Census data – to adjust skewed distribution of these characteristics within the sample (Gelman and Carlin 2002).

⁴⁶ It is important to be clear about specific definitions of weighting schemes. Poststratification weights are different from probability weights. The latter are known at the time the survey is designed and are used to adjust survey estimates for non-constant probability of inclusion. For example, in random digit dial surveys, individuals with multiple phone lines are more likely to be called than individuals with a single phone line. Probability weights adjust for differences in respondent selection probabilities such as these.

Weights are appropriate for use when the source of systematic bias is orthogonal to the source of random bias. In the case of *systematic sample selection bias*, the assumption of orthogonality is plausible. As discussed above, there is no reason to suspect that the members of the underrepresented groups – such as women and Southerners – who were interviewed are systematically different than the members of those groups who were not interviewed (net of interviewer-induced bias). After all, the sample imbalance exists because the pollsters deliberately drew non-representative samples based on these characteristics. Thus, using the cases of the underrepresented groups who were interviewed to represent those respondents who were not interviewed, through the use of poststratification weights, seems an appropriate strategy.⁴⁷

On the other hand, weighting is not necessarily the appropriate solution for *interviewer-induced bias*. In fact weighting can exacerbate the problems created by this form of bias because weights multiply the influence of respondents who are somehow “unusual.” Thus, in general weighting respondents on *all* measured characteristics is not an appropriate strategy. However, in the last section, I argued that education captures many of the important differences between respondents and non-respondents induced by the practices of survey researchers. Thus, when possible, weighting the data by education levels is an appropriate strategy. As with the individual-level analysis, by using such weights we can take advantage of the residue of interviewer-induced bias – namely the education level of the respondent – to mitigate the bias induced by the sampling procedures employed in early opinion polls.

In theory, therefore, the bias in the aggregate data may be corrected – albeit incompletely – by weighting the data on both education levels and those quota category variables – such as gender, region, and occupation – that can be matched to census data. In practice, however, the use of poststratification weights can be difficult. The gains from poststratification weighting come from

⁴⁷ Glenn (1975) recommends such a strategy (see p. 1-13).

fairly strong assumptions – namely that within each poststratification cell, the respondents can “stand in” for the non-respondents.⁴⁸ To make this assumption as plausible as possible, survey researchers often use a large number of poststratification variables. By crossing education by gender by region, for example, we can account for possible variation in the differences between men and women with a high school education across different parts of the country. If, for example, the gender gap on desired levels of government involvement in the economy is larger in the South than it is in the north, such a poststratification strategy will correct for these differences.

The use of a large number of poststrata comes with problems of its own. As the number of stratification variables increase, the number of weighting cells becomes larger and larger. For instance, consider a simple scheme in which the researcher wants to weight by gender, geographic region, and education. Say that she uses the four-category census definition of region (Northeast, Midwest, South, West) and a four-category breakdown of education (grade school or less, some high school, high school graduate, some college or more). With just this simple strategy, the sample must be broken down into 32 cells (2 gender categories times 4 region categories times 4 education categories). If we then wanted to add weights for the urban/rural distinction; we would need to break the sample into 64 cells. This example underscores the central problem of weighting; the finer the distinctions in the weighting scheme – and therefore the more plausible the homogeneity of opinion in a given cell – the fewer cases in each cell and the more unstable the aggregate estimates

⁴⁸ Actually, as Lohr (1999) notes, a poststratification weighting strategy assumes that, “(1) within each post-stratum each unit selected to be in the sample has the same probability of being a respondent, (2) the response or nonresponse of a unit is independent of the behavior of all other units, and (3) nonrespondents in a post-stratum are like non-respondents” (1999, 269). All of these assumptions lead to a single conclusion. In the context of nonresponse, as Gelman and Carlin note, the critical assumption is that poststratification “assumes a constant probability of inclusion within cell, where inclusion encompasses both selection and response. In the presence of nonresponse, it is desirable to post stratify as finely as possible, so that the implicit assumption of equal probability of inclusion is reasonable within each poststratification cell (with these probabilities being allowed to vary between poststrata)” (2002, 292). This intuition is well suited to the present problem because the sampling practices of the early pollsters can be thought of as creating a unit non-response problem.

derived from the weighted estimator.⁴⁹

The solution to this problem is not to abandon poststratification weights. Instead, the researcher should construct a series of poststratification weights, based on three or four variables.⁵⁰ 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 data.⁵¹ We can, for example, usually construct the gender by education by region weighting system outlined above and still have around 20 cases per cell – the minimum level of support recommended by Lohr (1999).⁵² Certainly poststratification weights cannot fully correct the biases that threaten aggregate data analysis of the surveys of the 1930s and 1940s. As Lohr (1999) warns, weights rarely eliminate all bias resulting from faulty sampling procedures. But again, the general strategy of analysis advanced here is to approximate the best solution to minimize the bias in the data. Such a strategy might not get us the “true” measure of public opinion, but

⁴⁹ When faced with this problem researchers tend to use a “raking adjustment” where they construct weights from marginals using an iterative procedure. For my purposes, raking is clearly inappropriate because, in addition to the poststratification assumptions, raking adjustment weighting requires the additional assumption that “the response probabilities depend only on the row and column and not on the particular cell” (Lohr 1999, 271). For example, we must presume that females in the north have identical characteristics to females in the south. If groups are heterogeneous 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. Given that the quotas were not cross-employed, we need to construct the weights carefully.

⁵⁰ The “three to four” variable guide is not a magic number, but represents an attempt to tradeoff the detail of the classification scheme (using more variables creates a more powerful weighting scheme) against the limited sample size of the datasets (more variables lead to empty weighting cells). While it is not wise to weight by more than three or four variables at a single time, it is possible to use a series of weighting schemes to estimate aggregate public opinion. Provided that the “core variables” – those quota categories, such as education, gender, and region – that are related to politically relevant variables are included, it is possible to rotate other potentially important variables, such as age and urbanicity into the coding scheme. To the extent that the use of the different weights gives us roughly similar picture of public opinion, we can have more faith that no one extraneous variable drives our results.

⁵¹ 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. The reason for the large samples is that researchers had not yet figured out that they could draw relatively small samples and still get adequate estimates of opinion. In fact, one of the early leaders in public opinion polling, Hadley Cantril, said of the 5,000 person sample that Roper used to predict the 1936 election for *Fortune* magazine, “from the point of view of statistical adequacy its sample was ridiculously tiny” (Katz and Cantril, 1937, 171).

⁵² Elliot (1991) reports that the U.S. Bureau of the Census insists on a minimum of 30 respondents per class.

employing these methods will certainly allow us to do better than we would if we ignored the problems in the data – an approach that has been to this point the dominant strategy in the field.

The one remaining question concerning the aggregate-level analysis is how to account for interviewer-specific effects in analysis. At the aggregate level, unlike the individual level, I argue that it is best to ignore such effects. The central problem here is that while it is possible to determine the deviations in the mean opinion of particular interviewers from each other, it is not possible to determine which interview is giving the “correct” baseline. In other words, while it is possible to adjust aggregate opinion to a uniform interviewer effect, determining which interviewer to use as that standard is extremely difficult.⁵³ In practice, if surveyors all interview approximately the same number of respondents, any interviewer-specific effect should cancel out, provided that there are no large systematic effects. Previous research suggests that such an assumption is tenable (Hyman 1956; Cantril 1944). Cantril, for instance, argues that even in cases where significant interviewer effects were found, “by and large, the over all difference obtained in a figure on a national poll will not vary significantly from a figure adjusted for interviewer bias” (1944, p. 117).

So, in sum, at the aggregate level to correct for the sample selection and interviewer-induced biases researchers should use poststratification weights to bring the distribution of the variables in the sample in line with those in the census. In practice, it is not possible to construct weights based on more than a handful of variables at a time. Thus, researchers should pick their weighting variables carefully. The best strategy is to first pick a series of variables from among education levels and those quota category variables – such as gender, region, and occupation – that can be matched to census data. The researcher should then construct weights based on the distribution of three or

⁵³ Of course, if it were possible to identify those interviewers who generated more representative samples of respondents, it would be advantageous to give greater weight to those cases. For instance, we might want to give greater weight to those interviewers who were more closely supervised and/or had more interviewing experience. But, given the small number of cases selected by interviewers and the lack of relevant measures of interviewer characteristics, such methods are not feasible.

four of these variables. Next, the researcher should construct a second series of weights based on a different combination of variables. Analyses should then be performed using both sets of weights to assess the robustness of the results. To the extent that different weights give approximately the same answer, the researcher can be more confident that her results reflect the underlying balance of public opinion rather than the vagaries of the weighting strategy. Again, while this strategy might not give us the “true” measure of the public will, it will give us results that better reflect the underlying distribution of opinion in the United States in the 1930s and 1940s. In Table 2, I outline the strategies for individual and aggregate-level analysis described in this section.

Example: Public Opinion in the 1940s.

Having outlined appropriate strategies for data analysis, in this section I demonstrate how these strategies can allow us to properly draw inferences concerning the shape and structure of public opinion in the 1930s and 1940s. Throughout this section, I use various datasets from this time to demonstrate how the opinion poll data that has largely been ignored for 50 years can inform our understanding of critical events and trends in American history.

For the purposes of analysis, I will primarily focus on several datasets collected by Hadley Cantril’s OPOR. I consider Cantril’s work here for several reasons. First, Cantril consistently collected measures on many of the variables needed to control for the interviewer-induced systematic biases, such as the education level of the respondent. Second, Cantril asked a variety of questions that were repeated at several points over time. It is therefore possible to assess the effects of sample selection and interviewer-induced bias on particular questions over a series of opinion polls. Third, from 1940 until 1943, Gallup did the fieldwork for Cantril’s polls. Thus, the nature of

the biases found in the Cantril data can be generalized to a broader set of data.⁵⁴ Finally, we know that Cantril polls were seen by FDR, thereby establishing a direct link between the opinion poll data examined here and the course of public policy. In fact, FDR actively sought Cantril's data and asked Cantril to include particular items on his surveys (Cantril 1967). By analyzing the OPOR data, we can see how the biases induced by quota-control sampling played out in the polls actually seen by political elites in this period.

Aggregate-Level Analysis

As argued in the last section, the use of poststratification weights to correct for sample imbalances seems the most appropriate strategy of analysis at the aggregate level. However, weights can also create problems for analysis by giving added importance to individuals who are unusual relative to the population. Before we accept weighted aggregate results as a more accurate representation of “the voice of the people” during the 1930s and 1940s, it is important to use auxiliary information to check to see if the weights can bring our survey estimates closer to the truth on quantities we can measure. This step is critical to the analysis that follows. By using the weights to “correct” measures of opinion, some might argue that I am overstepping the bounds of proper analysis. In effect, I ask “what would public opinion have looked like if we administered the survey to all Americans, both those who participated in the poll and those who did not?”

Fortunately, there exists external data to see if the correction I employ on the survey items where we do not know the true answer get us closer to population characteristics on those quantities where we do have information collected using rigorous probability sampling methods. OPOR surveys often contained questions measuring phone penetration. We can see if we can more accurately estimate levels of phone penetration using the weighted estimates than if we looked at the

⁵⁴ The OPOR data is preferable to the AIPO data for this period because AIPO did not consistently measure the respondent's education until 1944. As noted above, the measure of education is critical for conducting the analyses in the remainder of this paper.

unweighted estimates. If the weights do indeed bring us closer to the Census Bureau estimates – which were generated using a census or a probability sample of the full population – on these variables, we have can greater faith that their use will bring us *closer* to the “truth” on the opinion measures as well.

According to Census Bureau data contained in the *Historical Statistics of the United States, Colonial Times to 1970*, in 1940, 37% of American households had telephone service. In 1941, 39% of households had telephone service. Given the class and education biases in the sampling procedure, we should expect that the phone penetration in the sample of respondents would be greater than the population at large. Indeed, this was the case. In the January 1941 poll, 46.1 percent of the sample reported having telephone service in their home.⁵⁵ This figure exceeds the actual level of phone penetration by seven percent. However, once the sample is weighted by gender, education, and region, the estimated phone penetration rate drops to 39.6 percent – a near exact match with the census number.⁵⁶ Similarly, in a March 1941 poll conducted by OPOR, the estimated phone penetration rate was 44.9 percent. Introducing the weights again drops the estimated penetration rate to the true value of 38.9 percent. Thus, across two independent surveys, the weighting scheme brings us closer to the correct estimate of phone penetration, a quantity we can measure. We can therefore have greater faith that this weighting process also brings us closer to the truth on those quantities we cannot compare to “check data.”⁵⁷***NOTE: I am still searching for data on car

⁵⁵ 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 telephone service can be used as a measure of phone penetration during this time.

⁵⁶ To construct these weights I used a weighting scheme that crossed gender with the 4-category census region variable (Northeast, South, Midwest, West), a 4-category education variable (grade school or less, some high school, high school graduate, some college or more). None of the 32 cells in this sample matrix were empty and only one cell – college educated females in the South – had fewer than 25 cases. As I note below, however, the results reported here were largely robust to respecification using different coding schemes, indicating that these small cell size did not drive the results reported here.

⁵⁷ Furthermore the gains from weighting are, as expected, heavily dependent on the inclusion of the education variable in the weighting scheme. Weighting by region and gender only, lowered the estimated phone penetration rates to 44 percent.

ownership. In January 1941 the sample unweighted data show 55.2% own a car; the weighted, show that 49.3% own a car; In March 1941, Unweighted 54.1 have a car, weighted 47.1 have a car.***

I now turn to the more interesting question concerning the estimation of the distribution of politically relevant variables. It is important to enumerate the expectations regarding the nature of bias in the polls. Simply because we know that the polling practices employed in the 1930s and 1940s introduced sample selection and interviewer-induced systematic bias does not mean that the aggregate measures of opinion will be biased in all cases. Aggregate measures of opinion will be biased only in particular circumstances: (1) when the sample imbalances are related to the dependent variable of interest and (2) when all the biases induced by these sample imbalances work in the same direction (that is, there are no countervailing biases). Thus, weighting might sometimes make a difference in our understanding of politically relevant quantities, and other times make little difference. But, 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.

The January 1941 OPOP survey contained a number of questions concerning interest in and knowledge of world events. Given that the sample of respondents overrepresented the highly educated, men, and non-southerners – all groups that we would expect would be more involved with politics – it is likely that the unweighted marginals overstate the true political interest and knowledge of the American public. In other words, the sample biases – the overrepresentation of particular demographic groups – are related to the dependent variable of interest, here political knowledge. Furthermore, these biases work in the same direction across gender, education, and region. As a result, Cantril's survey overestimates the levels of political knowledge and interest among the mass public.

Several examples from this survey demonstrate this phenomenon. One half of the respondents to the January 1941 survey were asked, “Have you been following the discussion of the President’s lease-lend bill regarding aid to England and other countries, which Congress is now considering?” In the unweighted sample, 62.1 percent of respondents said that they were following the discussion. Weighting the sample to correct the imbalances in gender, region, and education drops the estimate of the size of the engaged public by almost five percent to 57.5 percent of the sample.⁵⁸ The same pattern of results holds when the topic turns from attention to politics to knowledge of politics. Later on the survey, respondents were asked “How many years has Hitler been in power?” In the unweighted analysis, 47.4 percent give the correct answer of 8 years. But the weighted analysis paints a somewhat different picture; only 41.1 percent give the correct answer.⁵⁹ The effect of using weights for our picture of the levels of political knowledge is even more striking on other questions asked on the survey. Respondents were also asked, “can you name a country where Greece and Italy are fighting?” In unweighted analysis, a majority of 54.7 percent gave the correct answer, Albania. But when the weights are introduced to correct for the sample selection and interviewer-induced biases, only 46.2 percent give the correct answer, a drop of almost nine

⁵⁸ The weights used in this analysis have a large range. For the January, 1941 survey, the lowest weight is 0.34 and the largest is 6.22. This range indicates that the 37 observations represented by the largest weight count 18 times as much as the 182 observations represented by the lowest weight. The range for the March, 1941 survey (weighting is region by education by gender) is somewhat more restricted but is still rather large. The lowest weight for that survey is 0.35 and the largest is 5.47. In practice, such a range of weights is normally undesirable. In the modern era, researchers truncate very high weights so that no single observation has a very large contribution to the overall estimate (Lohr 1999). But in the present day, unlike the 1930s and 1940s, no groups are *deliberately* misrepresented in opinion polls. Considering the skewed nature of the sampling procedures used in the 1930s and 1940s, a more generous standard of tolerance of weights seem appropriate. Even so, a note about the effect of the weights is in order. If the weights are truncated at their upper and lower bounds to ensure that the largest relative weight is 10 times the lowest weight, the estimated effects of the weighting diminish on the information and support for FDR questions, though the direction of effects remains the same. However, the performance of the weights in correcting known quantities diminishes as well. Specifically, using the revised weights increases the estimates of the phone penetration rates from 39 percent – the correct answer – to 44 percent. This result indicates that the use of the more extreme weights to correct for the sample imbalances might be justified.

⁵⁹ Using a less stringent standard for determining whether an answer is “correct” yields somewhat stronger estimates of the degree of bias. I computed a second measure that was scored correct if respondents gave an answer in the range of seven to nine years. Using this less stringent standard, in the unweighted analysis, 59.6 give the correct answer. But correcting the sample biases through poststratification weights drops the proportion of respondents who give the correct answer to 51.1 percent.

percentage points.⁶⁰

All together, these results suggest that the mass public that FDR viewed through the results of polls presented to him by Cantril was overly engaged with the political scene relative to the mass public, writ large.⁶¹ This finding is in itself important, but the results presented here also have important implications also for our understanding of changes in levels of political knowledge over time. Consider, for example, the work of Delli Carpini and Keeter (1996). Delli Carpini and Keeter examine levels of political knowledge in the 1940s relative to the present day and find that levels of political information are roughly the same in the two periods, though they find some evidence of modest increases on particular items over time.⁶² For example, they find that the percentage of the public who can define the meaning of the term “business recession” increased from 52 percent to 57 percent. Given that knowledge at the earlier point in time was measured using the same surveying practices as the OPOR poll analyzed above, it could be that the gains in political knowledge over the last half century are larger than they appear to be from an analysis of the uncorrected marginals.⁶³

It is not simply on questions of political information and interest that the sample biases have an effect on the shape of opinion. An analysis of other issue areas demonstrates the importance of attending to politically relevant divisions in the composition of the sample of poll respondents.

Consider for instance, estimated levels of support for FDR. In the July 1940 OPOR survey

⁶⁰ These results were robust to the use of different weighting schemes. For instance, I also used a weighting strategy that crossed gender, region (Northeast, South, Midwest, West), education (grade school or less, some high school, high school graduate, some college or more) and age (under 40/over 40). The percent of the weighted sample giving the correct answers to the information questions was as follows: Follow discussion of lease-lend (57.8%); Hitler’s tenure in power (41.4%); Greece and Italy are fighting in Albania (46.5).

⁶¹ As the estimation of phone penetration suggested, these findings are heavily dependent on education. Weighting by gender and region only decreases the estimated differences between the weighted and unweighted population. The percent of the weighted sample giving the correct answers to the information questions was as follows: Follow discussion of lease-lend (59.2%); Hitler’s tenure in power (44.8%); Greece and Italy are fighting in Albania (50.5).

⁶² This finding of increases in political knowledge contrasts with other work which suggests that knowledge levels may have declined somewhat over time (Bennett 1988, 1989; Neuman 1986; Smith 1989).

⁶³ The difference between the corrected and uncorrected marginals on the knowledge measures from 1947 are in all likelihood somewhat smaller than the comparable differences from the OPOR poll because in 1944 Gallup began interviewing equal numbers of men and women, thereby correcting the gender imbalance in his sample. However, the bias induced by the deliberate under-representation of Southerners and the incidental under-representation of low educated citizens ensures that the bias will still be in the same direction and will almost certainly be of a substantial level.

respondents were asked, “Would you prefer to see Roosevelt or Willkie win the Presidential Election this year?” In the unweighted sample, 48.9 percent expressed support for FDR.⁶⁴ However, while women are no more supportive of Roosevelt than are men, Southerners and those with low education are more supportive of the President. As a result, correcting the estimates for sample bias through weighting increases projected support for the President by over five points, 54.4 percent.⁶⁵ This result demonstrates that the polls from this may also misrepresent the political predilections of the public, broadly construed.⁶⁶

This is not to say that correcting the sample bias through the use of weighted analysis 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 asked 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 to help England win, even at the risk of getting into the war?”⁶⁷ The uncorrected proportion of people who said that it was more important to help England was 59.5 percent, while the weighted proportion was 59.8 percent.⁶⁸ This confluence of the weighted and unweighted results can be explained by the nature of the cleavages on these issues. Women and low educated respondents – who are

⁶⁴Support for Willkie was 38.9 percent, while 10.5 percent of respondents were undecided.

⁶⁵ At the same time the estimated number of respondents who respond “don’t know” increases from 4.3 to 5.0 percent. It should be noted that the findings concerning support for FDR and desire to support England are robust to the use of weighting schemes that include age, as were the information and engagement findings.

⁶⁶ The level of undecided respondents held steady at 10.2 percent, while support for Willkie dipped to 33.9 percent. The findings concerning the effects of sample bias in estimated levels of support for Roosevelt are not surprising. The Gallup poll consistently underestimated support for the Democratic presidential candidate from 1936 to 1948 (Robinson 1999). The analysis here suggests that this miscalculation was a direct result of the sampling methods employed by Gallup.

⁶⁷ 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 that Germany be defeated, even at the risk of our getting into the war?” The distribution of unweighted responses to this item was: “Keep Germany from winning” 57.6 percent, “Stay out of war” 35.8 percent, and “No Answer” 6.6 percent. The corrected responses are somewhat different, mostly owing to an increase in the percentage of “no answer” responses. Specifically, the response distribution is, “Keep Germany from winning” 55.9 percent, “Stay out of war” 35.7 percent, and “No Answer” 8.5 percent.

⁶⁸ An analysis of the March 1941 OPOR data yields similar results. In the unweighted sample, 70.1 percent say that it more important to “help England win.” In the weighted sample, 68.8 percent support this position.

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. The biases therefore cancel out; the OPOR polls give us the right answer for the wrong reason. The convergence between the weighted and the unweighted estimates on questions of war are not limited to a single item. On the same survey, respondents were also asked “If the British are unable to pay cash for war materials bought in this country, should our government lend war materials to the British, to be paid back in the same goods and materials after the war is over?” In the uncorrected sample, 69.4 percent of the sample either supported lending the materials or giving them to England outright. Correcting for the sample biases does not alter the picture of support for taking an interventionist policy; the weighted support for such a position is 70.2 percent, a virtually identical measure of opinion.

The message of this section therefore is clear. The sample selection and interviewer-induced biases have implication for how we understand the shape of public opinion during this time. In many cases, corrected opinion will not differ significantly from uncorrected opinion. But in some cases, correcting the opinion polls through weights will paint a different picture of public opinion. We cannot simply assume that correcting for these biases through weighting will always give us an answer of a particular stripe. We need to carefully consider the cleavages at work on a given issue and think about how those cleavages might affect the shape of opinion. By analyzing the public opinion polls in this way, we can have greater confidence that the inferences we draw about the public opinion more accurately reflect underlying public sentiment.

Individual Level Analysis

In the description of the data analysis strategy above, I identified two areas of concern in conducting analysis at the individual level. As at the aggregate level, it is important to adopt a strategy that accounts for both interviewer-induced and sample selection bias in the data collection

process. As a solution, I proposed controlling for inter-individual interview effects and introducing statistical controls for the quota-controlled categories and education. The logic of including the quota variables and education as control variables in individual-level analysis is clear; it is important to avoid possible omitted variable bias in the analysis. But the impact of accounting for inter-interview variation in response bias is less straightforward. I take up that question here.

Consider the use of such strategies. If we had information concerning which respondents were interviewed by which interviewers we could directly control for these interviewer effects. In general, the polls from this period do not contain such information, much less information about the characteristics of the particular interviewers. Fortunately, however, there is one poll that is particularly well suited to examine the effects of inter-individual differences in interviewing style.

In May 1942, Cantril ran an experiment to see whether the differences in the way that NORC and Gallup recruited their interviewers had effects on the shape of public opinion.⁶⁹ Cantril's survey team selected the areas to send interviewers in the standard way – seeking to draw a sample of “representative communities.” At this point, under normal circumstances Cantril would have sent a single interviewer from AIPO – or a team of interviewers if the local was large – to perform the interviews. But instead, Cantril matched the interviewers from each organization in each geographic location. Thus, for each interviewer from NORC in a given local, an interviewer from AIPO was present as well. Interviewers in the same local were given identical quotas of 10 respondents (for a description of this experiment, see Cantril 1944, Chapter 6).

Contrary to his expectations, Cantril found no consistent house effects on the answers that

⁶⁹NORC provided extensive training and supervision of novice interviewers. AIPO, on the other hand, secured and supervised the work of its interviewers mostly by mail. There were also significant differences in the composition of the staffs of the two houses. While AIPO and NORC had interviewers with comparable political leanings and levels of education, the ratio of men to women was 7:9 at AIPO and 1:7 at NORC. In addition, the NORC interviewers had slightly higher average levels of economic status (Cantril 1944).

individuals gave to the survey interviewers.⁷⁰ But while Cantril may have not found the consistent differences he was looking for, the design of this survey is advantageous for the study of inter-interview differences in interviewing bias because it allows us to match interviewers conducting surveys in the same geographic locale to see if there are consistent differences in interview styles. Specifically, the design allows us to avoid conflating local effects – the effects of the respondents’ state of residence or size of town on opinion – with interviewer effects.

To parse out the interviewer effects, I followed a simple strategy. I chose three questions asked in the survey concerning foreign policy matters: (1) whether respondents thought that the Japanese people would always want to go to war to make themselves as powerful as possible, (2) whether the German people would always want to go to war to make themselves as powerful as possible and (3) whether respondents felt that the information they were getting about the war was true and accurate.⁷¹ These variables are the only substantive questions asked on the survey ****CHECK FOR OTHERS**** I then specified a simple model of opinion direction that included relevant quota-controlled variables, namely the gender, education, and age of the respondent.⁷² Next, I added controls for the respondents’ town size and state or census region of residence.⁷³ Having

⁷⁰ I confirmed the lack of differences between NORC and AIPO on the opinion questions in my own analysis. It should be noted that there were some significant differences in the samples of the two organizations. In particular, the NORC sample had a higher average socio-economic status than the AIPO sample (though this finding is somewhat complicated by the fact that the two organizations used different systems of economic classification).

⁷¹ The exact wording of the questions are: (1) “Which of the following statements comes closest to describing how you feel on the whole about the people who live in Germany” [Mark “The German people will always want to go to war to make themselves as powerful as possible” as the high response]; (2) “Which of the following statements comes closest to describing how you feel on the whole about the people who live in Japan” [Mark “The Japan people will always want to go to war to make themselves as powerful as possible” as the high response]; (3) “Do you feel that the information you are getting about the war is true and accurate?” [Mark yes as the high answer].

⁷² This model is, admittedly, less than ideal. It would be better to use variables for which we had concrete theoretic expectation regarding the direction and strength of the relationship between the independent and dependent variables. Unfortunately, as noted above, this survey was very short and did not contain the types of variables necessary for in-depth analysis.

⁷³ In the analysis of the interviewer effects, I estimated the regression models using both census region and state fixed effects. Certainly, including state fixed effects is preferable to the region effects. However, when I attempted to use the state effects, the model would not always properly estimate because of excessive multicollinearity. Under these circumstances, I was forced to drop certain interviewer dummy variables. In these instances, I performed the analysis using the region dummy variables instead to capture differences in the responses of people from the same geographic locale. My results were almost identical under both sets of analyses. In the tables that follow, I include the region dummy

specified and estimated such a model, I included fixed effects for the interviewers. I was then able to see if the introduction of the interviewer fixed effects adds any explanatory power to the prediction through the use of an F-test of the joint significance of the interviewer-specific fixed effects.

At first glance, it appears that the introduction of interviewer-specific effects does add explanatory power to the model, indicating that accounting for inter-interviewer differences is important. On the question concerning the Japanese people, the interviewer effects are highly significant regardless of whether geographic residence is measured at the state level ($F(65,699) = 1.63$; $p < 0.002$) or the regional level ($F(88,686) = 1.60$; $p < 0.001$). The interviewer-specific effects are also significant on the question concerning the German people ($F(88,784) = 1.35$; $p < 0.03$) and the information question ($F(88,729) = 2.03$; $p < 0.0001$).⁷⁴

Once the focus shifts from questions of statistical significance to those of substantive significance, however, a somewhat different picture emerges. Table 3 compares the coefficients for the base model without the interviewer specific effects and with the interviewer-specific effects. As the table demonstrates, the interviewer effects have little appreciable effect on the substantive interpretation of the model for the two questions concerning the warlike tendencies of citizens of Axis countries; the coefficients on age, gender, and education remain largely unchanged.⁷⁵ On the question concerning information about the war, the introduction of the interviewer fixed effects affects somewhat the coefficients on the education dummy variables. The effect of a college education compared to a grade school education is somewhat smaller once the interviewer fixed

variables rather than the state dummy variables because in regression analyses I typically use the region dummies to capture geographic effects. However, at certain points, I present the analyses with the state fixed effects to demonstrate the robustness of the results.

⁷⁴ The geographic locale identifiers are used at the level of the census region in these runs.

⁷⁵ The statistical significance level of the coefficients is altered slightly, because the introduction of the interviewer-specific effects “cost” 88 degrees of freedom. I do not include the geographic locale variables for comparison because the meaning of those changes with the introduction of interviewer effects. With the introduction of the interviewer fixed effects, the regional effects are parsed out between region/state and interviewer dummies. What matters for the present purposes, however, is what the introduction of the interviewer dummy does to the coefficients of interest. As Table 3 demonstrates, the effects here are minimal.

effects are added. However, the coefficient is of the same size and still statistically significant.

In sum, this analysis demonstrates that while interviewer-specific effects add explanatory power, they often make little difference in the substantive conclusions we draw from regression-type analysis. So while it is advantageous to include interviewer-specific fixed effects when available, ignoring these effects seems to do little harm to our inferences. Analysis of this survey can guide future work in this area. Since inter-interviewer effects are largely absent in this survey, we can be more comfortable ignoring these effects in those instances where we do not have sufficient information to directly control for the interviewer effects.⁷⁶

Conclusion

To Come.

Some ideas:

- Through the use of simple correction 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 20th century.
- Provided researchers proceed carefully, the methods described in this paper can allow us answer questions that have previously been unanswerable.

⁷⁶ There are other reasons to think that researchers may safely ignore interviewer-specific effects. To determine whether interviewer-specific effects remained constant across different items, I created a dataset in which each observation represented a specific interviewer. I then created a series of variables that represented the coefficient on the interviewer fixed effect for a given question. So, for example, I created a variable called “Germany” that took the value of the coefficient on the interviewer number for the analysis of the “Germans are warlike” question. I created similar variables for the “Japanese are warlike” and the “fair news” questions. I then ran a series of correlation analyses to see whether any of the interviewer effects were consistently large or consistently small across the surveys. I found no such evidence of consistent interviewer effects. On the two questions that were most directly comparable – the “Japanese are warlike” and the “Germans are warlike” items – I found that the interviewer effects had a substantively small and statistically insignificant *negative* correlation of -0.08. None of the other correlations were significant in either a substantive or a statistical sense (The Japan and the fair news items correlated at 0.14, while the Germany and fair news items correlated at 0.05). All told, this analysis provides evidence that the interviewer effects identified here are random across different questions.

- By appropriately adjusting for bias and acknowledging the uncertainty caused by faulty sample design, the data can yield great benefits.
- Given the public opinion data collected by Gallup, Roper, Cantril and NORC, it is possible to undertake a revolution in the field of political behavior. There are a number of interesting 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.

Bibliography

1940. "The *Fortune* Survey." *Journal of Educational Sociology* 14(4): 250-253.
- Alderson, Wroe. 1946. "Trends in Public Opinion Research." in *How to Conduct Consumer and Opinion Research: The Sampling Survey in Operation*. ed. Albert Blankship. New York: Harper and Brothers Publishers.
- Anderson, Dale, 1952. "Roper's Field Interviewing Organization." *Public Opinion Quarterly* 16(2): 263-272.
- Banks, Seymour, Norman C. Meier, and Cletus J. Burke, 1948. "Laboratory Tests of Sampling Techniques: Comment and Rejoinders." *Public Opinion Quarterly* 12(2): 316-324.
- Bershad, Max A. and Benjamin J. Tepping, 1969. "The Development of Household Sample Surveys." *Journal of the American Statistical Association* 64(328): 1134-1140.
- Bryk, Anthony S. and Stephen Raudenbush. 2001. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.
- Bureau of the Census. 1976. *Historical Statistics of the United States: Colonial Times to 1970*. Washington, D. C.: U. S. Department of Commerce.
- 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 Blankship. New York: Harper and Brothers Publishers.
- Cantril, Hadley and Daniel Katz, 1937. "Public Opinion Polls." *Sociometry* 1(1/2): 155-179.
- Casey, Steven. 2001. *Cautious Crusade: Franklin D. Roosevelt, American Public Opinion, and the War Against Nazi Germany*. New York: Oxford University Press.
- Crespi, Leo P. 1945-6. "The Cheater Problem in Polling." *Public Opinion Quarterly* 9:431-445.
- Divine, Robert A. 1979. *The Reluctant Belligerent: American Entry into World War II*. New York: Wiley.
- Doenecke, Justus D. and John E. Wilz. 1991. *From Isolation to War, 1931-1941*. Arlington Heights, Ill.: Harlan Davidson.
- Frankel, Martin and Benjamin King, 1996. "A Conversation with Leslie Kish." *Statistical Science* 11(1): 65-87.
- Gallup, George, 1938. "Government and the Sampling Referendum." *Journal of the American Statistical Association* 33(201): 131-142.
- Gelman, Andrew and John B. Carlin. 2002. "Poststratification and Weighting Adjustments." in *Survey Nonresponse*. Ed. Robert M. Groves et al. New York: John Wiley and Sons.
- Glenn, Norval. 1975. "Trend Studies with Available Survey Data: Opportunities and Pitfalls." In Jesse C. Southwick (ed.), *Survey Data for Trend Analysis*. Williamstown, MA: The Roper Public Opinion Research Center in cooperation with the Social Science Research Council.
- Haner, Charles F. and Norman C. Meier, 1951. "The Adaptability of Area-Probability Sampling to Public Opinion Measurement." *Public Opinion Quarterly* 15(2): 335-352.
- Hansen, Morris H. and Philip M. Hauser, 1944. "On Sampling in Market Surveys." *Journal of Marketing* 9(1): 26-31.

- Hansen, Morris H. and Philip M. Hauser, 1945. "Area Sampling - Some Principles of Design." *Public Opinion Quarterly* 9(2): 183-193.
- 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 Blankship. New York: Harper and Brothers Publishers.
- 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." *Communication Monographs* 64 (June): 161-179.
- Johnson, Palmer O., 1959. "Development of the Sample Survey as a Scientific Methodology." *Journal of Experimental Education* 27 (March): 167-176.
- Katz, Daniel, 1942. "Do Interviewers Bias Poll Results?" *Public Opinion Quarterly* 6(2): 248-268.
- Katz, Daniel. 1946. "The Surveys Division of OWI: Governmental Use of Research for Informational Problems." in *How to Conduct Consumer and Opinion Research: The Sampling Survey in Operation*. ed. Albert Blankship. New York: Harper and Brothers Publishers.
- Kennedy, David M. 1999. *Freedom from Fear: The American People in Depression and War, 1929-1945*. New York: Oxford University Press.
- Kemsley, W. F. F., 1960. "Interviewer Variability and a Budget Survey." *Applied Statistics* 9(2): 122-128.
- Leigh, Michael. 1976. *Mobilizing Consent: Public Opinion and American Foreign Policy, 1937-1947*. Westport, CT: Greenwood Press.
- 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, 1947-1948. "Laboratory Tests of Sampling Techniques." *Public Opinion Quarterly* 11(4): 586-593.
- Menefee, Selden, 1944. "Recruiting an Opinion Field Staff." *Public Opinion Quarterly* 8(2): 262-269.
- Moore, Carroll S., Jr., 1946. [Review of the book, *Interviewing for NORC*]. *Public Opinion Quarterly* 10(2): 102-103.
- Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, and David B Truman, eds. *The Pre-Election Polls of 1948* (New York, NY: Social Science Research Council, 1949).
- 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 Blankship. New York: Harper and Brothers Publishers.
- 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, part1): 325-334.
- Roper, Elmo, 1940b. "The Public Opinion Polls: Dr. Jekyll or Mr. Hyde? Classifying Respondents by Economic Status." *Public Opinion Quarterly* 4(2): 270-272.
- Schuman, Howard, Charlotte Steeh, Lawrence Bobo, and Maria Krysan. 1997. *Racial Attitudes in America: Trends and Interpretations, Revised Edition*. Cambridge, MA: Harvard University Press.
- Southwick, Jessie C. 1974 *Survey Data for Trend Analysis*. Williamstown, MA: The Roper Public Opinion Research Center in cooperation with the Social Science Research Council.
- Smith, Tom W. "The Art of Asking Questions, 1936-1985." *Public Opinion Quarterly* 51(S): S95-S108.
- Stephan, Frederick F. and Philip J. McCarthy, eds. *Sampling Opinions* (New York, NY: John Wiley & Sons, Inc., 1963).
- Sudman, Seymour, 1966. "Probability Sampling with Quotas." *American Statistical Association Journal* 20 (September): 749-771.
- Warner, Lucien, 1939. "The Reliability of Public Opinion Surveys." *Public Opinion Quarterly* 3(3): 376-390.
- Wilks, S. S., 1940. "Representative Sampling and Poll Reliability." *Public Opinion Quarterly* 4(2): 261-269.
- Williams, Douglas, 1942. "Basic Instructions for Interviewers." *Public Opinion Quarterly* 6(4): 634-641.
- 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.

TABLE 1: QUOTA-CONTROL 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 ⁷⁷	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 states contribution to the total vote in the previous presidential election ⁷⁸	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 ⁷⁹
Roper	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 ⁸⁰	Gender Age 1938-1943: 21-39/40+ 1943-1961: 21-34/35-49/50+ Economic Class (A-D, Negro) Occupation	
NORC	Census Region: Cases assigned to nine regions in proportion to census numbers	Interviewing locations chosen based on size of city	Gender ⁸¹ 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” (1942, 640).

⁷⁷ AIPO did the field work for the OPOR surveys from 1940 to 1943.

⁷⁸ The information concerning the allocation of cases to the states comes from the SSRC report on the 1948 pre-election polls. 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 were assigned to the South.

⁷⁹ Cantril 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” (1944, 148-9).

⁸⁰ *Journal of Educational Psychology*, 14:4, 1940.

⁸¹ 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).

TABLE 2: DATA ANALYSIS SOLUTIONS

Level of Analysis	Systematic Sample Selection Bias	Systematic Interviewer-Induced Bias	
		<i>Approachability</i>	<i>Interviewer Effects</i>
Individual Level	Include quota variables as controls in analysis	Include Education as control in analysis	Include fixed effects for interviewer number
Aggregate Level	Use poststratification weights for education and quota-controlled variables (based on Census data)		Ignore in analysis

TABLE 3: INTERVIEWER FIXED EFFECT ANALYSIS**The German People Are Warlike**

Variable	Without Interviewer Effects Coefficient (SE)	With Interviewer Effects Coefficient (SE)
Age (in years)	0.01 (0.00)*	0.01 (0.00)*
Female	0.24 (0.10)**	0.20 (0.09)**
Education: Some HS	-0.19 (0.17)	-0.14 (0.14)
Education: HS Grad	0.09 (0.16)	0.06 (0.14)
Education: Some College	-0.32 (0.18)*	-0.33 (0.15)**
Education: College Grad	-0.31 (0.16)**	-0.27 (0.13)**
N/ Log Likelihood	754 / -451.69	792 / -533.12

The Japanese People Are Warlike

Variable	Without Interviewer Effects Coefficient (SE)	With Interviewer Effects Coefficient (SE)
Age (in years)	0.01 (0.00)*	0.01 (0.00)*
Female	-0.04 (0.10)	-0.06 (0.11)
Education: Some HS	-0.03 (0.15)	0.08 (0.17)
Education: HS Grad	-0.19 (0.15)	-0.17 (0.17)
Education: Some College	-0.15 (0.16)	-0.09 (0.19)
Education: College Grad	-0.08 (0.14)	0.02 (0.16)
N/ Log Likelihood	854 / -471.20	734 / -396.45

Feel That Information about the War is Fair and Accurate

Variable	Without Interviewer Effects Coefficient (SE)	With Interviewer Effects Coefficient (SE)
Age (in years)	0.00 (0.00)	0.00 (0.00)
Female	0.29 (0.09)***	0.31 (0.10)***
Education: Some HS	-0.25 (0.14)*	-0.30 (0.16)*
Education: HS Grad	-0.27 (0.14)**	-0.36 (0.16)**
Education: Some College	-0.15 (0.15)	-0.08 (0.17)
Education: College Grad	-0.43 (0.13)***	-0.29 (0.16)*
N/ Log Likelihood	833 / -556.93	781 / -458.07

***=p<.01; **=p<.05; *=p<.10

**FIGURE 1:
SAMPLE IMBALANCES: REGION AND GENDER**

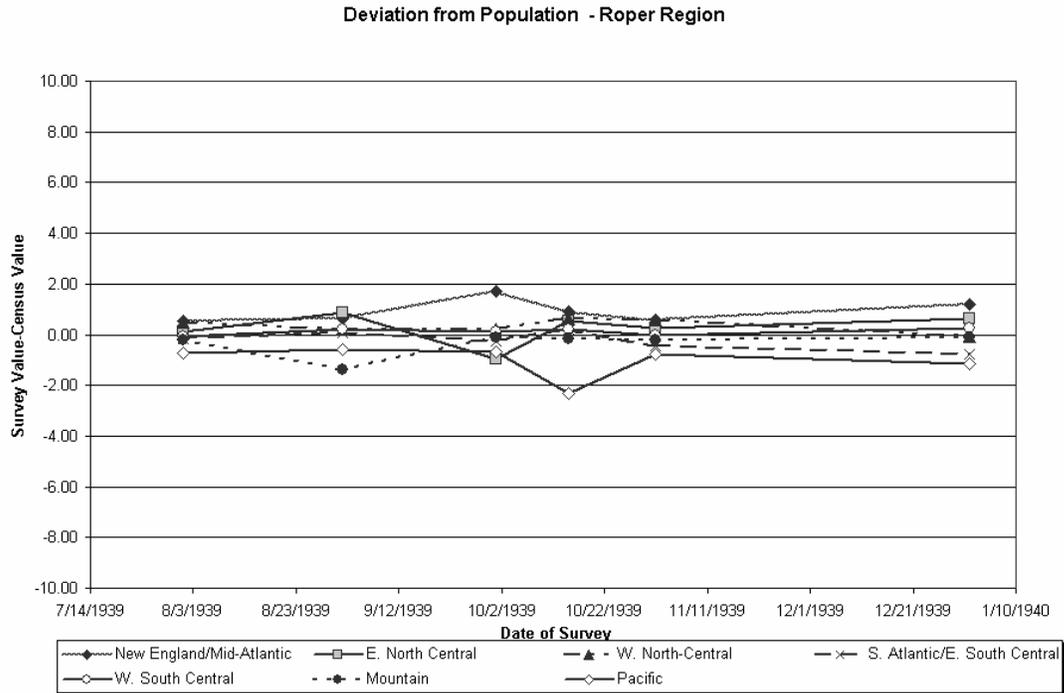
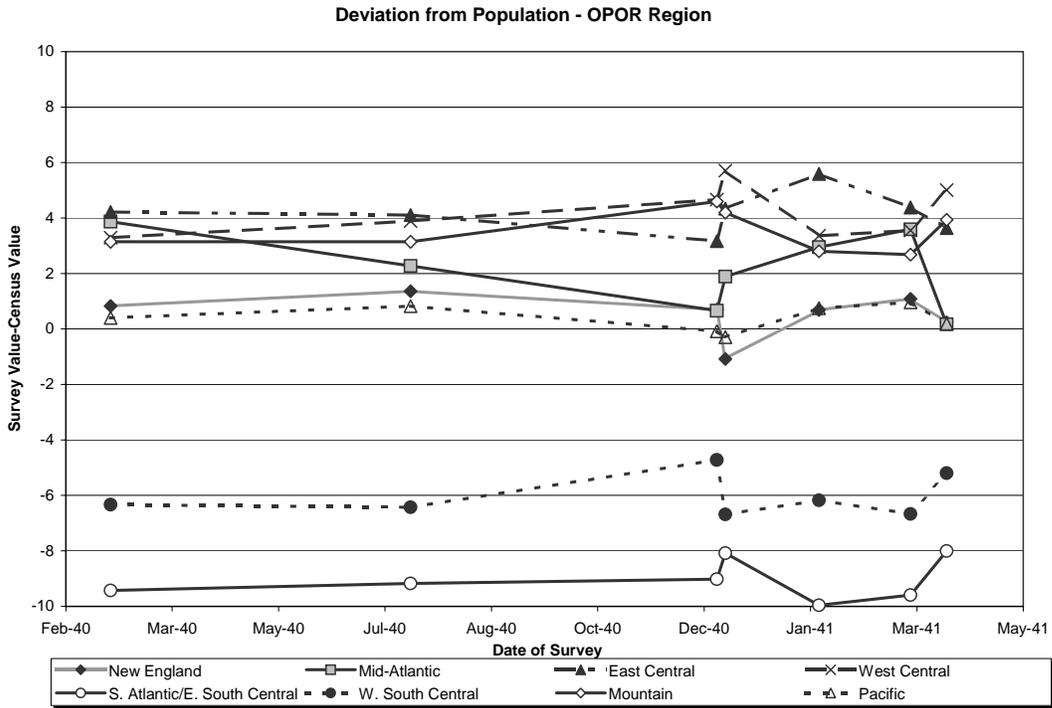
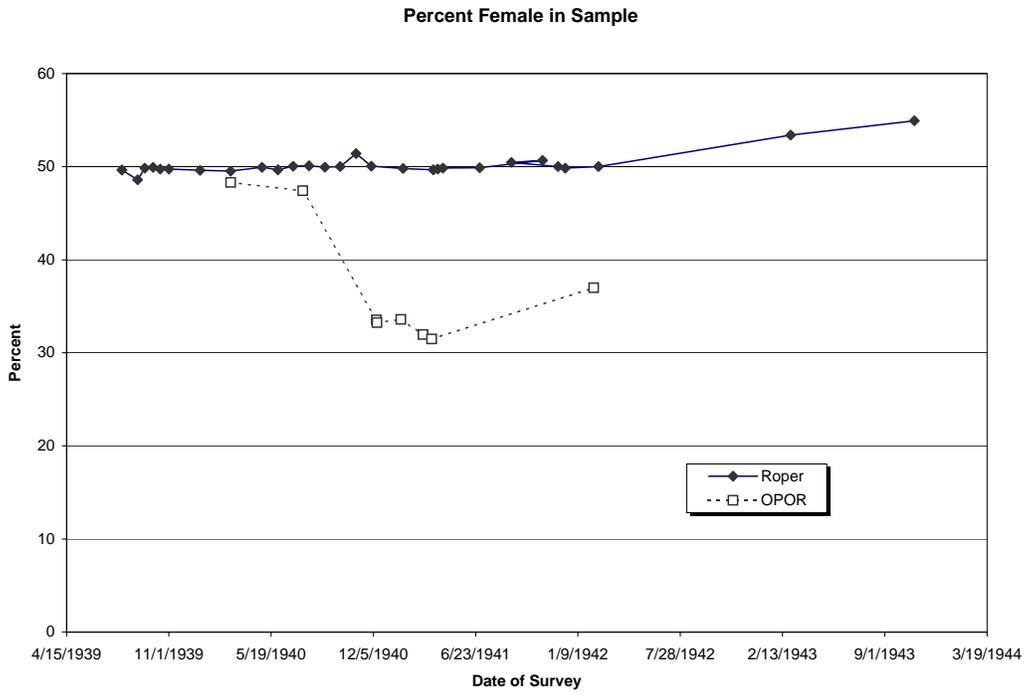


FIGURE 1 (CONTINUED)



**FIGURE 2:
SAMPLE IMBALANCES: EDUCATION**

