THE POLLS—A REVIEW

PREELECTION SURVEY METHODOLOGY: DETAILS FROM EIGHT POLLING ORGANIZATIONS, 1988 AND 1992

D. STEPHEN VOSS
ANDREW GELMAN
GARY KING

I. Introduction

Before every presidential election, journalists, pollsters, and politicians commission dozens of public opinion polls. Although the primary function of these surveys is to forecast the election winners, they also generate a wealth of political data valuable even after the election.[1] These preelection polls are useful because they are conducted with

D. STEPHEN VOSS is a doctoral candidate in the Department of Government of Harvard University. ANDREW GELMAN is assistant professor in the Department of Statistics in the University of California, Berkeley. GARY KING is professor in the Department of Government of Harvard University. They thank the following individuals for their openness and detailed assistance: David K. Krane and Robert Spanski of Louis Harris and Associates; Martin Frankel, a consultant for Louis Harris and statistian at Baruch College; Kathleen Frankovic, Bala Ramnath, and Marla Kaye of CBS; Sharon Warden of the *Washington Post;* Kristen Conrad of Chilton Research; Dale Kulp and Amy Starer of Marketing Systems Group, responsible for the GENESYS Sampling System; Steve Shaw of Media General; Kathy Walenczyk and Linda Piekarski of Survey Sampling, Inc.; Dan Soulas of ICR Survey Research Group; Burns W. Roper, Peter Case, and Brad Fay of the Roper Organization; Hal Quinley of Yankelovich; Jeff Alderman of ABC; Shari Weber of Gallup; and especially Gallup's Jack Ludwig, who weathered the inherent inefficiency of a trial run. They also thank Mark Lew and Cassie Hartzog for programming assistance; Bradley Palmquist, Dale Kulp, Roger Purves, Bob Groves, and anonymous reviewers for comments and corrections; and the National Science Foundation for grants SBR-9223637 (to Gelman and King), DMS-9457824 (to Gelman), and SBR-932121 (to King). Gary King also thanks the John Simon Guggenheim Memorial Foundation for a fellowship during which time this research was completed. E-mail addresses are Voss, dsvoss@isr.harvard.edu; Gelman, gelman@stat.berkeley.edu; King, gking@harvard.edu.

1. Most of these data are available through the Inter-University Consortium for Political and Social Research (ICPSR) or the Roper Center for Public Opinion Research.

Public Opinion Quarterly Volume 59:98–132 0033-362X/95/5901-0007$2.50

such frequency that they allow researchers to study change in estimates of voter opinion within very narrow time increments (Gelman and King 1993). Additionally, so many are conducted that the cumulative sample size of these polls is large enough to construct aggregate measures of public opinion within small demographic or geographical groupings (Wright, Erikson, and McIver 1985).[2]

These advantages, however, are mitigated by the decentralized origin of the many preelection polls. The surveys are conducted by diverse private enterprises with procedures that differ significantly. Moreover, important methodological detail does not appear in the public record. Codebooks provided by the survey organizations are all incomplete; many are outdated and most are at least partly inaccurate. The most recent treatment in the academic literature, by Brady and Orren (1992), discusses the approach used by three companies but conceals their identities and omits most of the detail.

If our only interest were in estimates of key variables for the entire population, such as national voter support for the Republican nominee, then presumably intensive detail would be unnecessary—we could just use the adjusted data provided by the organizations.[3] However, knowing how the surveys were conducted is crucial for researchers interested in more complex analysis of the data, such as studying subgroups in the population (e.g., support for the Democratic ticket among educated women), analyzing several variables simultaneously (as in a regression analysis), comparing surveys performed by different organizations, or assessing changes over time.

For example, understanding the sampling method that led to particular survey data is necessary to compute correct standard errors for estimated population quantities or to poststratify (i.e., adjust for differences between the sample and known population averages) by additional variables such as religion or income. Knowing the different procedures used by each organization also allows the researcher to estimate a variety of interesting population parameters. For example, surveys that used the "last birthday" method of respondent selection are useful for estimating the response rates of different subgroups of the population. Finally, knowledge of the weighting schemes used by the survey organizations tells us which variables they consider to account for the grossest discrepancies between samples and population.

Our purpose here is to present methodological detail for 1988 and

2. Notably, Wright, Erikson, and McIver (1985) chose to use only one organization's polls, those of CBS, even though the questions of interest, measures of self-reported ideology and partisan affiliation, are common to most of the polls. Presumably, they could have estimated state-by-state variables with greater certainty using polls from other sources as well.

3. Yet some knowledge of the process by which data are generated is always necessary for valid statistical inference (King 1989).

1992 collected from the most visible preelection pollsters, reported without concealing the organization names. This is valuable not only because it saves others the effort of gathering such information but also because it preserves information that otherwise might never appear in the public record. Our experience was that, because of personnel turnover and inadequate documentation, this type of detail quickly disappears from organizational memory. Collecting comparable information for as recent an election year as 1984, for example, is probably impossible. Without this basic information about the process by which survey data were generated, the vast collections of these data at the Inter-University Consortium for Political and Social Research (ICPSR), the Roper Center, and elsewhere are considerably less valuable.

We report on the following organizations and their data collection processes: (1) CBS and the *New York Times* (hereafter cited as CBS); (2) Chilton Research Services, which polls for ABC News and the *Washington Post* (hereafter as Chilton); (3) Gallup Organization; (4) Louis Harris and Associates; (5) Media General, which conducted 1988 polls for the Associated Press; (6) ICR Survey Research Group, which conducted polls for the Associated Press and sometimes ABC News and the *Washington Post* in 1992 (hereafter cited as ICR); (7) Roper Starch Worldwide; and (8) Yankelovich Partners. Our ability to compile this information relied at almost every step upon the openness and professionalism of sources at these organizations; they took exceptional pains to ensure that our queries received prompt and accurate responses and even read earlier versions of this article. As we had the willing cooperation of many knowledgeable individuals, we believe the information from these organizations to be accurate and complete. We were disappointed that, despite numerous phone calls and letters, we were unable to find anyone at the *Los Angeles Times* who was both willing to speak with us and knowledgeable about the process by which their survey data were collected. Unfortunately, this makes analyses of their data considerably less trustworthy.

Our observation while collecting the technical detail from the cooperating organizations was that the various polls actually differ along a limited number of procedures. Therefore, rather than simply listing the methodology of each organization separately, we introduce the main variables across which professional polls tend to differ. This allows us to discuss a variable in the abstract, and then briefly outline the specifics for each organization.

Our discussion is broken into three parts. Section II outlines how each organization generates its list of phone numbers from among those possible for the continental United States. Section III examines the choices required after a list of phone numbers has been generated, such as how to treat busy signals, refusals, and calls answered by

electronic devices, how to decide which household members are eligible to be interviewed, and how to select the respondent from among those eligible. Section IV details the weighting procedures generally used by each polling group to improve the data after their collection. We also analyze data from 50 preelection polls from 1988 in order to demonstrate the importance of weights and the effects of different methods of weighting. Appendix A provides details on how we gathered this information.

Our purpose here is to report the methods used by these survey organizations. Estimates of quantities such as cross-tabs and regression coefficients are not usually trustworthy without knowing how the data were collected and adjusting appropriately. Because these surveys can be used for many purposes, and we are not specifying ahead of time the parameters to be estimated, we cannot make comparative judgments about the different sampling procedures we discuss here.[4]

II. Choosing Phone Numbers

Until several decades ago, telephone polls usually drew phone numbers from those listed in telephone directories; this procedure skewed samples away from households that were more mobile or chose unlisted numbers (Fletcher and Thompson 1974; Roslow and Roslow 1972). It also introduced a bias away from households without telephones, a weakness that is still tolerated (Brick et al. 1994, pp. 8–9). A major reason for the popularity of this unsatisfactory method was that calling a list of random 10-digit numbers was prohibitively expensive—approximately 80 percent of potential phone numbers are not residential connections. Warren Mitofsky, formerly of CBS, seeking a feasible method of random-digit dialing, developed a sampling procedure that reduced the probability of a wasted call by utilizing "phone banks," or clusters of telephone numbers with the same first eight digits (Frankovic 1992, p. 33). (For example, the number 617-555-1212 is part of the bank ranging from 617-555-1200 to 617-555-1299.) Only a fraction of these banks contain residences; if identified, these residential banks can be included in a more restrictive sampling frame that reduces the number of wasted calls (Waksberg 1978).[5]

4. For example, it is difficult to construct unbiased measures of complicated causal effects from survey data selected in ways related to the outcome variable.

5. Because telephone company switching technology used to be quite costly and unwieldy, phone companies filled banks methodically, with houses in a subdivision or companies in an industrial zone sharing similar numbers. Packing numbers in this way minimized the switches a company needed to buy. However, several sources told us that the recent revolution in telecommunications technologies has reversed the incen-

Two primary methods are used to sample functioning residential banks. One, the Mitofsky-Waksberg Method, is a two-stage sampling procedure. First, a survey organization samples from the pool of active central office codes (i.e., three-digit area codes plus three-digit exchanges) used by telephone companies in the contiguous United States. Once a reduced group of exchanges is selected, the survey organization dials a random number in each. If the first number called from a bank produces a nonworking commerical or industrial line, or if no one answers after repeated attempts and the phone company verifies the lack of service for that number, the call is terminated and the entire bank passed over. If the telephone number is a residential line, the bank is added to the list of Primary Sampling Units (PSUs). Thus, in theory each has a probability of inclusion in the list of PSUs that is proportional to the number of accessible residential numbers in the bank of 100. Then, in a second stage, surveys can be conducted by drawing numbers from the PSUs.

The second method of PSU selection is called list-assisted sampling (Lepkowski 1988). Unlike the polls of decades ago that simply used telephone directories as their lists, the common list-assisted procedures used today are more sophisticated and introduce fewer selection effects.[6] The main advance over old list-based sampling methods is that pollsters now only use the list to identify their sampling frame; they then implement a system of random-digit dialing within chosen telephone banks. The most common directory used is the Donnelly Quality Index, a comprehensive data base of listed residential telephone numbers in the United States. From this a survey sampler can draw a list of "working banks," defined as phone banks in the contiguous United States with a particular number of listed residential phone numbers. Once the residential banks are identified, they can be stratified based upon geographical criteria. For example, the area code and exchange for a bank tells us its location, and it is possible to further identify geographic locations of telephone area codes.[7]

---

tive—the more a company spreads numbers over different switches, which are much smaller and cheaper than their predecessors of 2 decades ago, the less likely any one is to be overloaded. Although phone number assignment changes incrementally, new policies are eroding the cost efficiency of modern telephone survey methods, a loss only partly compensated for by the increasing number of phone lines in residential units.

6. Brick et al. (1994, p. 4) estimate that a common list-assisted method, one used by the GENESYS Sampling System, excludes 3.7 percent of all telephone households, with a 95 percent confidence interval of 3.0–4.3 percent. They also present evidence that excluded households are similar to included ones on numerous demographic characteristics.

7. Roper includes preelection poll questions in its door-to-door polling, not in telephone surveys. Since its method of respondent location is unique among these organizations, we save the details for App. B below. All the other polls restrict their calling to the District of Columbia and the 48 states excluding Alaska and Hawaii.

A. CBS/NEW YORK TIMES

CBS uses the two-stage Mitofsky-Waksberg Method. In the first stage, CBS stratifies central office codes by the four census regions and size of place.[8] The regions used are the four Census Bureau regions. The "size of place" ranking includes four categories: large city, small city, suburb, and miscellaneous. Classification of a central office on this scale is determined using its latitude and longitude, provided by AT&T. If the exchange's central office is also the central office for more than 25 other exchanges, it is classified as a large city. If it contains 15–25, it is a small city. If it does not fit one of the first two categories but its central office is located within 15 miles of a large city central office, the exchange is considered suburban. The remaining exchanges are placed in the miscellaneous category. Within strata, exchanges are listed numerically.

CBS then selects exchanges for sampling using systematic selection.[9] That is, from a list of all exchanges listed by stratum and then by number, CBS divides the total number of exchanges by the quantity desired to form a sampling interval. A random start is generated within that interval, indicating one exchange; the interval is then repeatedly added to this starting value to generate a list of exchanges of the desired size.

Once the exchanges are selected, four random digits are added to each central office code, and the resulting phone number is called. CBS calls each selected number as many as six or seven times if it is unanswered. Of the 15,000 or so numbers called, around 3,500 produce residences.[10] CBS adds an eight-digit "seed number" to its list of Primary Sampling Units for each successful contact, corresponding to the phone bank for the number called. For example, if interviewers make a successful residential contact at 617-441-0586, then the eight-digit seed 617-441-05 becomes a PSU. Phone numbers for all pools conducted that year are drawn from this list of PSUs.

Once CBS has developed its list of PSUs, the eight-digit seed numbers within each region are distributed serially into about 40 equal-sized batches, in the order they were first generated.[11] That is, the first PSU goes into the first replicate, the second into the second; then when the forty-first is reached it would go into the first replicate again. This assignment is used all year until the sampling of PSUs occurs again.

8. Certain central office codes, such as internal phone company exchanges and exchanges reserved for military use, are excluded from the sampling frame along with unused ones.
9. The quantity varies by year. The last CBS sample selection included 15,487 numbers.
10. The last sample selection produced 3,253 seed numbers.
11. That is, they are not assignd in the order they were confirmed as residential.

When CBS is preparing a particular survey, it allocates the number of PSU batches needed to have a list of phone numbers of the desired size, distributed equally across the four categories (i.e., quota sampling by region). One to three phone numbers will be generated for each PSU, with the quantity per PSU constant for each survey; this decision obviously influences the number of batches required.[12] Across surveys, CBS cycles through the batches, so if the first 12 batches for the South were used in the year's first poll, the next survey will take its southern batches starting with number 13. For each phone number to be generated in a PSU, two random digits are added to the seed number for that bank, providing a 10-digit number. CBS calls every phone number generated for a particular survey. For this reason, the number of observations in a survey depends on the response rate.

B. GENESYS SAMPLING SYSTEM: CHILTON AND ICR

Chilton and **ICR** get a data base of working banks with at least two listed households as part of the GENESYS Sampling System, available from Marketing Systems Group in Fort Washington, PA.[13] The GENESYS system, which is used to generate each list of phone numbers for the two organizations, uses an implicit stratification scheme across the 10 census divisions (New England, Middle Atlantic, East North Central, and so on). Working banks are stratified among these 10 divisions, then further segregated according to whether they serve a Metropolitan Statistical Area (MSA), effectively producing 20 categories. These are ordered in sequence: New England Metro, New England Non-Metro, Middle Atlantic Metro, and so forth. Within each metro stratum the associated working banks are ordered by size of MSA, by whether they serve a central city or suburban county, and finally in numeric sequence. Within the nonmetropolitan strata, banks are ordered in a serpentine geographic fashion for each state.

GENESYS divides the total number of possible phone numbers contained in all working banks (approximately 173 million, 100 per bank) by the desired quantity of phone numbers, thereby determining the required size of the selection interval. The list of eight-digit working banks, with their associated 100 numbers, is sliced into a series of equally sized intervals. Finally, a single phone number is selected at random from all potential numbers making up each interval, resulting in single-stage equal-probability of sampling method (epsem) telephone numbers.

12. The number of batches is not determined using a formal estimation procedure; those administering the survey guess the number needed based upon past experience.
13. GENESYS is a flexible computer software application. This article describes how it is used by Chilton and ICR, not its overall potential.

GENESYS divides phone numbers into replicates for Chilton and ICR. The survey organization specifies the number of replicates; each phone number is assigned randomly to one of the replicates. So if 100 replicates is the goal (commonly the choice), a number is chosen randomly between 1 and 100 for each phone number; this governs its placement. When interviewing, Chilton and ICR will exhaust each replicate before opening another.

C. SURVEY SAMPLING, INC.: MEDIA GENERAL AND YANKELOVICH

Survey Sampling, Inc. (SSI), based in Fairfield, CT, uses a cleaned-up version of the Donnelly Quality Index to generate lists of "working phone banks" for Gallup and Harris and actual lists of phone numbers for **Media General** and **Yankelovich.** A working bank is defined, for this purpose, as a phone bank with at least three listed residential numbers. Here we describe Survey Sampling's method of sampling phone banks; in later sections we will discuss differences between the polling organizations that use these lists.

SSI assigns each telephone exchange to an individual U.S. county. An estimated 70 percent of exchanges fall within a single county's boundaries; the remaining are placed in whichever county contains the highest proportion of listed phone numbers. The counties are listed using the Federal Information Processing Standards (FIPS) codes, and each is given a measure of size equal to its number of estimated telephone households. Finally, each county is given a cumulative measure of size (MOS) equivalent to its MOS plus that of all counties listed before it. The number of estimated telephone households is determined using data reported by Market Statistics, Inc., which provides estimates of households and population at the county level.[14] SSI subtracts the county number of census nontelephone households from the publication's county household estimate to get its estimate of telephone households and therefore its estimated proportion of households with telephones.[15]

SSI adds up the total number of estimated telephone households across all counties in the contiguous United States. It divides this by the desired quantity of telephone numbers for a poll to construct a sampling interval, generates a random number in this interval, and then adds the interval size to this starting value once for each phone number to be included in the sample. For each number, the accompa-

14. Since the 1992 polls, SSI has begun to use Strategic Mapping's Conquest system. The Market Statistics data were published in *Sales and Marketing Management* from Bill Communications, recently renamed *Market Statistics.*

15. Therefore, SSI assumes that almost all households gained or lost since the census have telephones.

nying county is identified—this will be the first county in the listing with a cumulative MOS larger than the value. Every time a county is selected in this way, it is assigned one phone number for the list.

Within each county, working banks are sorted by area code and exchange. The phone numbers for each county are assigned in the following way. Each working bank is given a probability of selection equal to its proportion of the county's listed numbers, using an interval method for which the exchanges and banks are listed numerically. Once the bank is selected, the final two digits are determined randomly. If the number first selected is not eligible, each number in the bank will be checked in sequence, and on to subsequent banks, until an eligible number is located. Each phone number selected must pass two eligibility checks. First, it must not appear on a data base of about 11.2 million business phone numbers maintained by SSI. Second, it must not be marked as already having been called by SSI or a client in the year of the survey.

When Survey Sampling develops the list of telephone numbers, it usually divides them into replicates of 200 (although Yankelovich sometimes purchases lists of 250 numbers each). The number of replicates, therefore, is determined by the quantity of telephone numbers generated for a particular poll. As the numbers are determined, they are assigned across replicates serially. Media General will not use a replicate until the one preceding it is exhausted.

D. GALLUP ORGANIZATION

Gallup purchases lists of working banks from Survey Sampling and then stratifies the working banks by state and then by county. Each phone exchange is assigned to the county that contains a plurality of the exchange's listed numbers.

Gallup allocates a number of calls to each assigned county according to the proportion of the population's households residing in that county. For example, if a particular assigned county had .28 percent of the nation's households and Gallup were planning to generate a list of 2,000 numbers to call, the number of sample households allocated to it would be $.0028 \times 2{,}000 = 5.6$. It would receive five or six households, with a 60 percent chance of receiving six. Each telephone number to be generated comes from one of the assigned county's working banks, with each bank given equal probability of being selected. Gallup adds two random digits to the eight-digit working bank seed.

The phone numbers generated by this method are divided serially into replicates, with around 225 phone numbers in each. Interviewers are given one replicate at a time; the next is released only after the first is exhausted, and the minimum number of replicates possible are

used in generating the sample. Since the phone numbers are distributed into replicates serially, surveys that do not use all replicates have been conducted under an implicit method of systematic stratification. For example, if a survey has 10 replicates but only nine are used, then the survey used the first nine phone numbers of every 10 in the list. See Kalton (1983, pp. 16–19) for a succinct discussion of the implications of systematic stratification.

E. LOUIS HARRIS AND ASSOCIATES

Harris also purchases lists of working banks from Survey Sampling. Harris apportions each state's exchanges (considering the District of Columbia as a separate state) by four categories: (*a*) those from the central city of a Metropolitan Statistical Area, which is any area defined as metropolitan by the Office of Management and Budget (i.e., within actual city limits); (*b*) suburban areas of an MSA (i.e., outside city limits but part of the greater metropolitan region); (*c*) non-MSA counties containing a city or town of at least 2,500 households (Rural-I); and (*d*) non-MSA counties without such a town or city (Rural-II). This produces $49 \times 4 = 196$ potential strata, although seven are empty (e.g., New Jersey has no non-MSA areas), leaving 189 valid strata.[16]

Harris determines the number of calls to make in each of its 189 strata according to the proportion of total U.S. households found there. For example, New York Rural-I has .5066 percent of U.S. households. In a sample of 1,250, the number of sample households allocated to it would be $1{,}250 \times .005066 = 6.33$. It would receive six or seven households, with a 33 percent chance of receiving seven.

Within each stratum, systematic selection of phone numbers is restricted to the working residential phone banks. A measure of size (MOS) for each bank is computed as its quantity of listed telephone numbers multiplied by the ratio of county population to listed county phone numbers. Phone banks are sampled with probability proportional to size, as follows. Harris gives each bank a cumulative measure of size, which is its MOS added to those of all previous banks. Therefore the cumulative MOS for the final working bank also represents a total measure of size for the entire stratum. This total MOS is divided by the number of sample households desired from the stratum, producing "systematic sampling intervals" (e.g., New York Rural-I would have six or seven intervals). A number within the interval is selected

16. Data for this categorization are obtained from Survey Sampling, Inc., and from surveys of buying power by Market Statistics, Inc., reported in the publication *Sales and Marketing Management*. For example, the 1992 polls used data from the August 1991 issue of the magazine.

at random, corresponding to the first bank in this list with a cumulative MOS exceeding the chosen figure. The remaining banks are chosen by adding the size of the interval to this original number and finding the bank corresponding to the new figure. One phone number is generated randomly for each of these selected banks.

Harris then generates five backup numbers for each original number in its list, allowing interviewers to use these secondary numbers only when the callback policy on a primary number has been completed without producing an interview. The back-up numbers are generated as follows. Once the primary bank's county is identified, a cumulative measure of size is calculated for the working banks within that county. This number is divided by five, producing five equally sized intervals from which secondary banks are selected using a method identical to that used for the primary bank, discussed above.

III. Selecting the Respondent

Consider the obstacles faced by an interviewer when calling a selected phone number. The interviewer might get a busy signal, no answer, a nonworking number, a business, an answering machine, or a voice mail system. Furthermore, even if the call gets through to a residence, the person who answers might not be cooperative, might be a minor with no eligible voters at home, might not speak English, or might be an adult ineligible for voting because of failure to register.[17] Each of these possibilities requires a policy on the part of a polling organization, which in turn influences the eventual reliability of the sample. In almost all cases, the professional polling organizations work to get interviews out of these initial numbers rather than simply looking for households that are easier to survey, using repeated "callbacks" to the same numbers made at different times of the day. Doing so lessens the selection effects against people who are harder to reach or interview, which is important since they tend to be substantially different from those who are more acessible (see Potthoff, Manton, and Woodbury 1993).

Many residences that do have an eligible adult home also have other adults in the household, some of whom might be present. Taking the first adult available would bias the sample, since certain types of voters (especially females) are more likely to answer. Therefore, a polling organization needs some means of choosing among the adults in the

17. All six organizations that use telephone interviews terminate a call if the phone number is not residential or if language barriers prevent an interview. These numbers do not receive callbacks.

household. For example, some limit the pool of potential respondents in the household to those present at the time of the call; others do not. Some use a systematic method of respondent selection that targets, say, young males. Others use a method of respondent selection that is effectively random, such as asking for the adult with the most recent birthday or even requesting information on the number of eligible adults in the household and choosing from among these randomly.

We can isolate four decisions regarding the procedure for calling randomly generated phone numbers: (*a*) How many callbacks are made to a number that initially provides no answer, a busy signal, or some mechanical answering mechanism? (*b*) What is the policy if an interviewer initially received a refusal from all adults accessible at a phone number? (*c*) Is a random or systematic system of respondent selection used? (*d*) What happens if a respondent claims not to be a registered voter?

A. CALLBACKS

For **CBS,** the number of callbacks is governed by the time constraint under which a particular poll is conducted. Whereas some polls may take 4 days and allow as many as six or seven callbacks to some numbers, others are conducted immediately after a preelection debate or convention speech and afford very little room for additional calls. When callers get a busy signal, the first callback might be within 10 minutes, since presumably someone is home but on the phone. Otherwise busy signals, phones answered by mechanical devices, and phones that are not answered when first called are all treated the same, with callbacks made later at different times of day, some during working hours and some not.

Chilton and **ICR** have a policy of three callbacks after the original dialing, scheduled at different times of day. For ICR, a busy signal is automatically called back in a half hour. If the number is busy for a second time, the call is treated as unanswered and will be called back at a different time of day.

Gallup considers a replicate exhausted only after interviewers have attempted a minimum of two callbacks for each number. Seemingly promising numbers, such as those with answering machines, may receive an additional callback.

Harris attempts three callbacks at various times over the next 3 days. If the caller receives a busy signal at any time, an additional call back is attempted 15 minutes later.

Media General's policy is to attempt five or six callbacks. If the first call produces a busy signal, one of the callbacks is attempted after 15 minutes. Otherwise, callbacks are at different times of day and are

treated similarly for busy signals, phones answered by mechancial devices, and phones that are not answered when first called.

Yankelovich attempts three callbacks to all numbers, regardless of whether they have a busy signal.

B. REFUSALS

When receiving a refusal, **CBS** generally attempts to call back at a later time with a different interviewer, as per the regular callback policy, but removes the number from its list should the residents refuse to cooperate again.

Chilton and **ICR** do not attempt to call back after an initial refusal.

Gallup does not treat refusals differently from other failed contacts. These numbers receive the requisite two callbacks. If someone refuses but agrees to a later interview, additional callbacks are attempted.

If a designated respondent refuses to cooperate with **Harris,** up to three callbacks are made, but once the household is contacted and someone refuses again, a new number is tried. Also, a new number is called if the reason for refusal is a health concern or if the respondent is abusive. The only exception is if the person indicates willingness to be interviewed at a later time, in which case an additional contact is attempted.

Media General treats refusals the same as other failed contacts but moves on if the same number provides a second refusal.

Roper interviewers, in their door-to-door polling, do not return to households where they have received a blanket refusal, although they might if the person expresses willingness to be interviewed at a later time.

Yankelovich does not call back after receiving a refusal.

C. RESPONDENT SELECTION METHOD

To select respondents at a household, **CBS** uses a grid method proposed by Leslie Kish and adapted by Troldahl and Carter (1964) to increase response rates. The Kish grids used by CBS require two pieces of information: the number of adults in the household (one, two, three, or more than three) and the number of adult women in the household (zero, one, two, three, or more than three). For each combination (say, a three-adult house with two adult women) a particular five-by-four table indicates a particular adult (say, the youngest woman). The adult indicated by any particular cell varies across each Kish table, such that random selection of a particular table for a particular call also ensures random selection among all adults in the house-

hold (Backstrom and Hursh 1963). Also, since all adults living in the residence are considered when choosing the respondent, adults not at home have the same chance of being chosen as those who are (which, of course, requires CBS to call back at a later time and allows the collected data to reflect varied refusal rates across the demographic categories). The Kish tables contain a slight but inherent selection bias among the individuals in large households. For the CBS version this bias applies to households clumped into the "more than three adults" category. In cases with four or more adults, the tables are more likely to indicate younger respondents.

Media General will request the adult currently at home with the most recent birthday. **Chilton** requests either the male or female at home with the most recent birthday, with males more likely to be requested. **ICR** asks for either the male or female adult (sex alternates) in the household with the most recent birthday, and therefore includes those not present. In single-sex households, the adult with the most recent birthday is selected. Using birthdays is a convenient means of selecting among those at home, encourages reasonably high response rates, and is unbiased as long as we can presume that political opinions and demographic characteristics are not highly correlated with date of birth (Salmon and Nichols 1983).

When a call is answered, **Gallup's** interviewers request the youngest male present at the time who is 18 years or older. If no adult males are present, the interviewer asks for the oldest female 18 years or older. Gallup takes the first person willing to respond under this systematic selection system, so if the male is uncooperative but a female in the house is not, the interview would be conducted with the female.

Harris also surveys the youngest male at home willing to cooperate, but requests the youngest female if no male is home and willing.

Roper, in its door-to-door survey, does not have a systematic selection method for interviewers when they try a particular residence. However, each interviewer is given a fairly rigid quota system to fill while collecting his or her total of 20 interviews. First, at least 9 of the 20 must be regularly employed in a permanent, but not necessarily full-time, job. Additionally, the interviewer must fill four age/sex groupings, women under 45, women 45 and over, men under 45, and men 45 and over. Some interviewers, determined randomly, are told they must get 10 men and 10 women (with more of each in the younger age group). The rest are told they must get 9 men and 11 women, with the quotas again further divided among the two age groupings. The quotas are determined by Census Bureau information on the population's sex and age composition, so that the entire sample will meet population proportions should all interviewer quotas be met. Since

experienced interviewers know which categories are hardest to find, they probably end up using a systematic system similar to those of Harris and Gallup. While administering surveys, the interviewer will skip a house that does not have any potential respondents because of quota limitations.

Yankelovich, for the National Security surveys in 1988, asked for the youngest adult male who was at home and asked for the youngest adult female if no adult males were home. If the respondent was not a registered voter, the interviewer asked to speak to any registered voter who was at home. For the *Time* Magazine/CNN polls, Yankelovich asks for the adult currently at home whose birthday passed most recently.

D. RESPONDENTS WHO ARE NOT REGISTERED VOTERS

CBS does not terminate interviews with respondents who are not registered voters. Rather, it asks whether respondents are registered, and usually screens unregistered respondents from reported results at the weighting stage, discussed below.

Chilton interviewers ask unregistered voters the demographic questions for purposes of weighting and then terminate the interview.

ICR, on the other hand, conducts interviews with all respondents regardless of registration rates; registration status is recorded in the data and can be used for weighting.

Gallup asks only demographic questions for weighting from respondents who claim not to be registered and then terminates the interview. However, if the respondent lives in one of the four states that does not have registration deadlines, the interview is conducted normally.[18]

Media General terminates interviews with unregistered respondents.

Roper conducts no screen for registration or likelihood to vote.

Harris asks demographic questions from unregistered respondents for purposes of weighting only.

Yankelovich conducts polls of all adult respondents. However, the political questions are not always asked of all respondents, especially the questions of candidate support. Instead, political questions are asked of all registered voters included in the poll. After Labor Day, news reports drawn from the polls are even more restrictive; only "likely voters" are used, defined as those who said they were registered, were "very likely" to vote, and "always" voted in previous elections.

18. In 1988 and 1992 those states were North Dakota, Maine, Minnesota, and Wisconsin.

IV. Adjusting the Data

When attempting to estimate political opinions held in the population using a survey sample, inaccuracies in the estimates can stem from many sources. Some of these, such as flaws in the questionnaire, are beyond the scope of this article but can be quite significant. Two types of error are relevant here: sampling error and nonsampling error. "Sampling error" refers to the fact that, while the polling methodology on average would produce accurate population estimates, the particular survey under discussion might fail to do so merely by chance. We refer to "bias," on the other hand, in cases for which estimates of population quantities based on means would be inaccurate on average even if the survey were repeated numerous times, because of systematic error in the sampling technique. Although these two types of error are theoretically distinct, it can be difficult to separate the two sources of error for a single survey, such as when observing variation in poll estimates around an unknown and changing parameter (e.g., support for a candidate). Therefore, we will refer to a poll as "inaccurate" if it misses the mark for an unknown combination of sampling and nonsampling error.

Pollsters use poststratification to adjust for known differences between the sample and target populations. The most common poststratification technique is called "weighting," a method of adjustment based on assigning a numerical weight to each individual in the sample and then estimating population quantities by weighted averages of the individual responses. Weighting is useful because responses to particular questions of interest tend to correlate with broad demographic categories such as sex, race, region, age, education, or income. When poll data is unrepresentative on one of these categories, the analyst is alerted that the sample also might be inaccurate in its estimation of public opinion or preferences. For example, a survey that undersamples African Americans will almost certainly also underestimate support for Democrats. Therefore, in deriving population estimates from their sample statistics, organizations generally weight their reported results to take into account demographic discrepancies between sample and population.

Polling organizations also sometimes use weighting to address two other concerns with their sample. First, individuals have varying probabilities of being interviewed because of differences among their households; the likelihood of selection increases as a household has more telephone numbers or fewer eligible adults. If these household characteristics correlate with a survey variable, then failing to correct for varying probabilities of selection will bias the estimate. Second, a

greater proportion of people tell interviewers they will vote than actually do so.[19] Since preelection surveys usually are intended to estimate the preferences and opinions of voters, not all adult citizens, pollsters often want some means of determining the probability that a respondent actually will vote. Some of this concern is covered when conducting the interview, as we saw above, since often unregistered respondents are excluded from the sample. Nevertheless, some polling organizations use a more sophisticated method to estimate probability of voting that, on average, should produce poll results that are a more reliable indicator of voter preferences.

Whereas the categories of weighting are quite varied, the general procedure followed with these categories is fairly similar across organizations. To get an estimate from our sample of the population's opinion, we do not give each response equal influence. We need to "weight" the responses of undersampled groups more heavily than those of oversampled groups, thereby estimating the poll responses that would have resulted if the survey respondents had matched the population in their demographic characteristics. Consider a simple case, weighting by sex. In a 1,000-person sample that is going to be adjusted by sex, the weight for women would be the number of women out of 1,000 people in the population divided by the number of women in the sample. Each female response would be weighted by (i.e., multiplied by) this ratio.[20] If, for example, the sample had fewer women than the general population, female responses would be given more weight when estimating the population's characteristics on each question.

Of course, this last example only weighted along one dimension. Samples often are weighted along more than one dimension at a time. For example, if the 1,000-person sample is going to be weighted by race (e.g., black, white, other) and sex simultaneously, ratios would be computed for six categories—black men, white men, men of other races, black women, white women, and women of other races. The ratio estimate for black women then would be the number of black women out of 1,000 people in the population divided by the number of black women in the sample. This sometimes is called a two-dimensional matrix of weights, and the individual groups are called "cells" of the matrix. In theory, one could compute a five- or six-dimensional matrix of weights. However, at some point the number of respondents in each cell would be too small for the weighting to be reliable. Therefore, statisticians with the polling organizations com-

19. CBS, in an internal memo called "What Is the 'Probable Electorate'?" reports that 80 percent of respondents claim to be registered to vote, and 80 percent of those claim that they will vote.

20. Prior to the first weighting stage for any poll, the weight for every respondent is 1.

monly perform several stages of weighting, each using one or two categories. The weight for a respondent derived from the first stage is multiplied by the weight determined by the second, and so on, until final weights for all respondents are achieved. This procedure increases efficiency of the estimates, because each cell has a larger weight, but increases potential bias because of the additional independence assumptions.

Some organizations go beyond this to adjust for two-way classifications, using iterative proportional fitting (Deming and Stephan 1940). This method derives marginal totals for specific population characteristics from Census Bureau reports and then cycles through a series of variable combinations, weighting the data to match these targets. During iteration through this list of variable combinations, the sample distribution converges with the target distribution across variable combinations and the weights converge to 1.

Depending upon the method used, the number of respondents reported after weighting may differ from the actual number in the sample. An obvious example of this is when data are weighted for number of phones in the household; all the weights either are one (if the house had one phone) or less than one (if it had more phones). In this case, many organizations normalize results by multiplying the weights by a constant so they sum to the original sample size. Other organizations report the weighted sample size.

A. WEIGHTING METHODS

CBS uses several stages of weighting. A demographic weight is determined by five steps, based on (*a*) number of adults in the household, such that the weight is the average household size within the U.S. population divided by the number in the respondent's household; (*b*) number of telephone lines in the household, with a weight of 1 for one-line homes and 0.5 for homes with more than one line (regardless of how many more); (*c*) a ratio of households within each of the Census Bureau's four regional categories by the sample number in each region; (*d*) race by sex ratios, where race is divided as black and nonblack; (*e*) age by education ratios, with age divided into four groups, 18–29, 30–44, 45–64, and 65 years and older, and education divided into four groups as well: "not a high school graduate," "a high school graduate," "some college without graduation," and "college graduate." Sometimes educational groups are collapsed within age groupings, particularly when cell sizes are very small or when weighting effects disproportionately influence educational differences within age categories.

Preelection polls conducted in mid-September or later are also

weighted for likelihood of voting to produce what CBS calls the "probable electorate." The surveys include a handful of questions geared toward determining a respondent's probability of voting, all of which are worded to encourage the respondent to admit failure to participate: (1) whether the respondent is registered in his or her current precinct or election district; (2) self-reported likelihood to vote on a 4-point scale; (3) whether the person voted in the last presidential election, followed by a request for the name of the chosen candidate; (4) the year of the last election in which the person voted; (5) the last year in which the person registered to vote; (6) attention paid to the campaign on a 4-point scale; (7) whether the person moved in the past 2 years; and (8) whether the respondent voted in a party primary or caucus.

The first step in generating the "probable electorate" is to eliminate unregistered voters, a step aided by information CBS gathers on each state's election laws. Many states remove voters from the rolls if they have not voted within a certain time, so each respondent whose last reported vote and last reported registration both precede his or her state's cutoff is weighted zero. The exception is for North Dakota respondents, since their state does not require registration, and any state without a purge law or with no registration deadline before the election. Respondents are also weighted zero if they report moving in the past 2 years and do not report registering since then, although North Dakota residents are again an exception.[21]

The remaining voters are then divided into 12 categories based upon their self-reported past behavior and interest in the campaign. Since young voters (those 18–22 years of age) did not have the opportunity to vote in the last presidential election, they are treated as having done so if, in response to question 4, they report having voted in any election. Each category is given an estimated probability of voting, expressed as a proportion. These probabilities of voting are estimated for each presidential election. The categories and their associated weights for 1988 and 1992 are displayed in table 1.

For each category, the probability of voting is determined using data from two sources: (*a*) CBS postelection studies, which give the proportion of voters in each of the 12 categories that report having voted after the election, and (*b*) Michigan validation studies, which can be used to estimate, out of those claiming to vote, the proportion

21. One difficulty arises for people who report not knowing the last time they registered or voted, since they cannot be assumed not to have done so (as was done prior to 1988). As long as a respondent provided a valid answer for either the last time they registered or the last time they voted, a "don't know" for the question would not exclude the person from inclusion in the data-set of eligible voters. Because the poll contains information about the respondent's participation in the last primary and general election, that information is used to supplement responses about last vote.

Table 1. CBS Probable Electorate Weights, by Category

Category	Voted in Last Election?	Likely to Vote in Coming Election?	Attention to Campaign	Moved in Last 2 Years?	Weight in 1988	Weight Average in 1992
1	Yes	Definitely	A lot	No	.91	.88
2	Yes	Definitely	Some	No	.87	.85
3	Yes	Definitely	A lot	Otherwise	.89	.88
4	Yes	Definitely	Some	Otherwise	.87	.86
5	Yes	Definitely	Otherwise	Any answer	.83	.85
6	No/DK	Definitely	A lot	Any answer	.87	.83
7	No/DK	Definitely	Otherwise	Any answer	.75	.69
8	Yes	Probably	Any answer	No	.68	.67
9	Yes	Probably	Any answer	Otherwise	.68	.67
10	No/DK	Probably	Any answer	Any answer	.45	.36
11	Yes	Otherwise	Any answer	Any answer	.46	.29
12	No/DK	Otherwise	Any answer	Any answer	.03	.11

SOURCE.—CBS internal memo on the Probable Electorate, provided by Kathleen Frankovic.

NOTE.—Weights displayed for 1992 and used by CBS in 1992 are the average of the 1988 and 1992 computed weights. "Otherwise" means any response not already mentioned above. "Any answer" means all possible responses or nonresponses. "DK" means "don't know."

of respondents in roughly analogous categories who actually voted. The two are multiplied together to produce the weight for each category, representing the estimated probability that someone in that category will vote. For 1988, the Michigan properties could not be generated, because the question on whether respondents voted in the last presidential election was absent from Michigan validation studies in the early 1980s; CBS had to use proportions from 1980 with their 1984 postelection survey data to produce the 1988 weights in table 1. In 1992, CBS used information on past voting from the 1988 Michigan validation studies.[22] They combined this with the CBS 1988 postelection survey data to produce estimates of the probability to vote for 1992. However, these probabilities were significantly different from the 1988 values, a problem researchers at CBS attributed to differences in question wording between the two organizations. To solve the gap between 1988 and 1992 estimates, CBS used an average of the two when reporting 1992 survey results. The averaged probabilities are shown in table 1 for 1992.

Once these probabilities of voting are determined, voters can be assigned a probability to vote: zero if apparently ineligible; otherwise, the appropriate weight estimated for each person's category. This number is then multiplied by the demographic weight for a respondent to produce a final weight. Weighted data are normalized to so that the results reflect the number of respondents actually polled.[23]

ABC performs its own analysis on the raw data produced by **Chilton.** ABC's representative, Jeff Alderman, refused to provide methodological detail about this weighting scheme for "proprietary reasons." He did indicate that he "looks at swaths of the electorate" and "throws more respondents out" than CBS, and also said he includes a measure of strength of candidate support for preprimary surveys. Furthermore, ABC does not provide likelihood of voting weights with the Chilton data at the Roper Center. Therefore it is impossible to replicate the ABC analysis reported on news programs using the Roper data.

However, Chilton performs its own preliminary weighting; this is the source of weights reported with Chilton data from the Roper Center. **Chilton** uses a four-dimensional matrix of weights: (*a*) sex by age, where age is broken down as 18–30, 31–44, 45–60, and 60 or older; (*b*) race, using "black" and "other" as the categories; (*c*) education, where three categories are "less than a high school graduate," "a high school graduate," and "college education or more." No weighting is

22. These results are included with National Election Studies (NES) data available from the ICPSR. The NES postelection file for 1980 is indexed as ICPSR 7763. For 1988, it is ICPSR 9196. The validation study apparently was dropped from the 1992 NES survey.
23. However, probable electorate weights ordinarily are not normalized.

performed for likelihood to vote or probability of being contacted. However, this convention does not apply to the polls during debates and presidential conventions.[24] Also, Chilton will collapse cells if, roughly speaking, less than 5 percent of the sample falls into any one of them. Education categories are most likely to be collapsed, followed by age and race.

Gallup divides its weighting into two stages. The first stage includes the various demographic categories. The categories for first-stage weighting are (*a*) sex, (*b*) race, with the categories of black, white, and other, (*c*) region, with four categories similar to those of the Census Bureau, (*d*) age, with categories of 18–24, 25–34, 35–44, 45–54, and 55 or older, and (*e*) education, with categories of "less than high school," "high school graduate, "some college," and "college graduate." The first matrix used is age by sex by education. The second is region by sex by race.

A second stage of weighting takes into account the likelihood to vote. The method varies by year. In 1988, likelihood to vote was measured on a 9-point scale, for which the respondent would receive one point in each of the following self-reported circumstances: (1) If registered to vote or in a state with no registration requirement (i.e., only North Dakota in 1988). (2) If planning to vote. (3) If interested in politics a great deal, a fair amount, or a little. (4) If respondent votes always, nearly always, or part of the time. (5) If respondent has voted in the current precinct before. (6) If respondent has thought about the election a lot or some. (7) If respondent can give a specific response when asked where people from the area vote. (8) If respondent either could name which candidate was supported in the last election or claimed to have voted but could not remember the candidate. (9) If respondent, when asked to give likelihood of voting on a scale from 1 to 10, gives a number of 7 or more.

Any respondent who reports being unregistered automatically gets a zero for this scale unless living in a state with no registration requirements (North Dakota) or with no registration deadlines (in 1988, Maine, Minnesota, and Wisconsin). Also, any respondent who expressed no intention of voting is given a score of zero.[25] In 1992, Gallup

24. The surveys of the ABC vice presidential debate and the two presidential debates had no weights in the files supplied to us by the Roper organization. The first ABC Democratic convention poll had all weights equal to 1. ABC's other Democratic convention poll, their three Republication convention polls, and their politics poll of August 1988 had weights that depended only on sex.

25. This scale is adjusted for people ages 18–19 as follows: 0 = 0 (where = in this footnote and the next means "is recoded as"), 1 = 2, 2 = 3, 3 = 5, 4 = 6, 5 = 7, 6 = 9. It is adjusted for people ages 20–21 as such: 0 = 0, 1 = 1, 2 = 2, 3 = 3, 4 = 5, 5 = 6; 6 = 7, 7 = 9. This is because two questions are not used for those in the latter age range—"frequency of voting" and "vote in last election"—and one additional question is excluded for those even younger—"ever voted in precinct."

used a 7-point scale. Two questions were dropped—"registration" and "interest in politics."[26]

After this score is determined for each respondent, Gallup figures out what score would be the cutoff that would produce the expected turnout.[27] For example, suppose that, to produce a given turnout estimate, everyone who ranks a 7 and one-third of those who rank a 6 would be needed to meet the turnout estimate. In this case, everyone with a 7 would not be adjusted, everyone with below a 6 would be weighted zero, and everyone with a 6 would have their weight adjusted to one-third its first-stage value.

Harris uses five variables for ratio estimation: education, which has five categories (less than high school, high school graduate, some college, college graduate, post graduate); race, which has three categories (black, Hispanic, other); sex; and age, which has 10 categories (18–20, 21–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–64, 65–74, 75 and older). Weighting follows iterative proportional fitting, starting with education by a collapsed set of three age categories (18–29, 30–49, 50+), then sex by collapsed age, then race by collapsed age, and finally age by race (with the black and Hispanic categories collapsed into nonwhite). The weighting procedure then cycles back to education, continuing until the weights converge. Although Harris is considering weighting by number of phones in a household, this had not been instituted yet. Harris also occasionally weights by region, but only when it seems to be imbalanced after the other weighting procedure (undersampling in the South occasionally prompts this).

ICR generates household weights and population weights. For the former, households with one voice line are given 1.0; those with more are given 0.5. These weights are multiplied by ratios from a two-dimensional table, with one dimension corresponding to the nine census divisions and the other by location inside or outside an MSA, so that the results reflect the proportion of households in the population that fall into each category.

Population weights are derived in three steps. Households with adults of both sex are given 1; single-sex households are given 0.5, because, for example, males in male-only households have greater probability of being selected than males in mixed households. Then the weights for respondents in households with additional adults of the same sex are multiplied by 2.0 because they have less chance of being

26. For voters aged 18–19, 0 = 0, 1 = 2, 2 = 4, 3 = 5, 4 = 7. For voters aged 20–21, 0 = 0, 1 = 1, 2 = 3, 3 = 4, 4 = 6, 5 = 7.

27. The expected turnout is not determined systematically. Instead, four Gallup employees involved with the polling are asked their predictions, which then are averaged to produce the estimate.

selected than those who are the only adult of their sex in the household; others are multiplied by 1.0. Finally, ICR uses a four-stage, iterative sample balancing, in which the previously weighted data are used to generate ratios, for which the categories are (1) census region (four categories), (2) age by sex (7 × 2), (3) educational attainment (three categories), and (4) race (two categories).

The iteration ends when weights converge. Age is categorized as 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75 or older. The educational attainment categories are "less than high school," "high school graduate or some college/technical school," and "college graduate or more." The race categories are black and other. After each iteration, extreme weights on the low or high end are "trimmed" by being set to the tenth and ninetieth percentile values, respectively.

Once the two are computed, the household weight is divided by the respondent selection weight drawn from population figures. Finally, the household data set is again weighted using two-factor ratios for census division and metro status.

Media General weights by sex and then inspects other demographic categories to ensure that they are reasonably close to population proportions; weighting is occasionally used if the sample is off the mark. In the two 1988 polls in our possession from Media General, one is weighted by sex only and the other by sex and race (black, white, Hispanic, or other).

Roper uses an iterative sample balancing program, with weights drawn from three stages: (*a*) sex by age, using four categories (18–29, 30–44, 45–59, 60 and older); (*b*) region, using the nine census divisions; and (*c*) race, using three categories (white, black, other). While these are the routine weights, other categories are inspected and used as weights if significantly off the mark (although this probably did not happen with the preelection polls in 1988 and 1992). Roper does not weight by size of household or probability of being contacted. It reports the weighted sample size.

Yankelovich generally weights by five variables: region (nine census divisions), gender, race (Hispanic, black, white, and other), age (18–24, 25–29, 30–39, 40–49, 50–64, and 65 or older), education (high school and under, some college, college graduate, some postgraduate study), and marital status (never married, married, divorced or separated, and widowed). These variables are included in a sample balancing program to compute the weights.

B. EFFECTS OF WEIGHTING

Poll results reported in the news are almost always weighted averages. The weights, intended to adjust for unrepresentative samples and, for

some polls, probability of voting, usually accompany the survey data when purchased from ICPSR or the Roper Center. This section shows that, far from being trivial adjustments to the data, these weights often have a significant impact on population estimates derived from the surveys. For this reason, understanding weighting methods is an integral component of survey data analysis.

Using individual-level 1988 survey responses and weights reported by eight of the organizations discussed here, we compare weighted and unweighted population estimates for particular questions.[28] The target populations for the surveys vary; some include the voting-age population and others restrict the survey to registered voters (see Sec. III.D). To keep the different polls comparable in our analysis, we restrict our attention in these comparisons to registered voters—that is, for surveys that do not screen out unregistered voters, we consider only the respondents who claimed to be registered; for the few surveys that did not ask about registration, we consider only the respondents who claimed to be likely to vote. Also, although CBS includes a weight for likelihood to vote with their Roper data when applicable, the reported results only include the weight used in all polls, a combination of demographic and probability of selection adjustments.

Figure 1 displays the effect of weighting on the question of greatest interest at the time of the polls—estimated support for George Bush (or the Bush/Quayle ticket, in the later stages of the campaign). Figure 1*a* displays the change in estimated support for Bush produced by weighting. Each letter on the figure represents a different national survey from the 1988 presidential election campaign; the letters code the survey organization, and the capitalized letters correspond to surveys that were weighted only by sex.[29] For each symbol, the difference between the weighted and unweighted average support for the Republican ticket is plotted against the date of the survey. Therefore, surveys with a positive value along the vertical axis would have underestimated Bush support (presuming that the weighted results are more accurate); those with a negative value would have overestimated it. Except for the outlying Gallup and CBS polls on the top of the graph, weighting generally reduces the estimated support for Bush. If the surveys had been reported unweighted, the Republican ticket on average would

28. The individual-level poll data were purchased from the Roper Center for research reported in Gelman and King (1993). All survey organizations discussed in this article, with the exception of ICR, are included. Sample sizes range from 500 to 2,000. We also include national polls by the *Los Angeles Times* for which we have individual-level responses and weights; all our knowledge of their polling methods comes from the codebooks for their polls supplied to us by the Roper Center.

29. Surveys without weighting adjustments were not included in the figure.

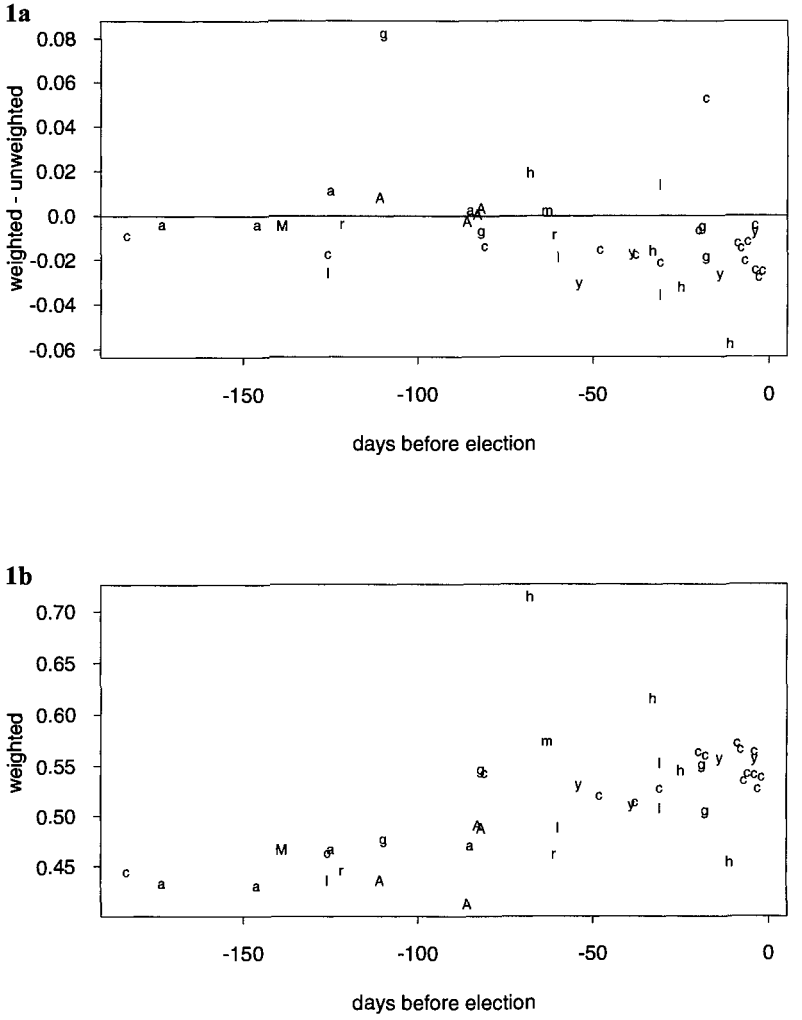

Figure 1. Fig. 1*a*, Effect of weighting on proportion who support Bush. Fig. 1*b*, Proportion who support Bush over time. *a* = ABC/*Washington Post*/Chilton; *c* = CBS; *g* = Gallup; *h* = Harris; *l* = *Los Angeles Times; m* = Media General/AP; *r* = Roper; *y* = Yankelovich. Capital letters correspond to polls whose weights depend only on sex. Polls with no weights or with all weights equal are not included. All figures are for registered voters (or likely voters when registration was not asked). All polls are from the 1988 presidential election campaign, with details presented in Gelman and King (1993).

have appeared 2–4 percent more popular during the campaign. The systematic nature of the weighting indicates that these differences are not merely correcting for random error.

The graph has some interesting yet subtle patterns. First, the different organizations' weights have different effects. For example, the Chilton weights generally have almost no effect; their averages are almost the same, whether weighted or unweighted. Second, the oversampling of Bush supporters seems greatest in the final weeks of the campaign. Perhaps the pollsters are more hurried then, with less time for callbacks and thus more weighting required to adjust for incompleteness in the design.

Figure 1*b* displays the weighted average support for Bush over time, showing that the different organizations all seem to adjust to about the same place. The figure also shows why it is necessary to use information from more than one source for a time-series study of preferences; no single polling organization has a series of polls extending all the way from the beginning to the end of the campaign.

Figures 2*a* and 2*b* present similar figures for the estimated proportion of registered voters who are women, a variable chosen because it is used explicitly in all weighting schemes and because the population parameter does not change much. Figure 2*a* shows a consistent pattern, with the unweighted proportion for women about 2 percent higher than the weighted, meaning that women were consistently oversampled. Also, there are again consistent differences among survey organizations, presumably due to differences in the respondent to selection method—Gallup, Yankelovich, Harris, and Roper have adjustments consistently near zero, while CBS requires more drastic weighting. Figure 2*b* shows that, after adjustment, the surveys differ little, except that the CBS values are about 2 percent higher than the others, suggesting that even higher adjustments might be appropriate.

The need for greater adjustment with CBS, however, does not imply that data from this organization are somehow less reliable than those from others. Rather, this reliance on postsample adjustment probably stems in large part from CBS's strict use of the Kish grid. Since their respondent selection method does not allow an interviewer to switch to other members in the household if the indicated person is uncooperative or unavailable, the CBS results before weighting reflect any unequal rates of accessibility among demographic categories. By contrast, compromises made at the selection level by other organizations smooth out differences between sample and population, and in ways that are not observed directly. Of course, some of the CBS difference also might be a result of its unique weighting scheme, such as by number of telephones in the household.

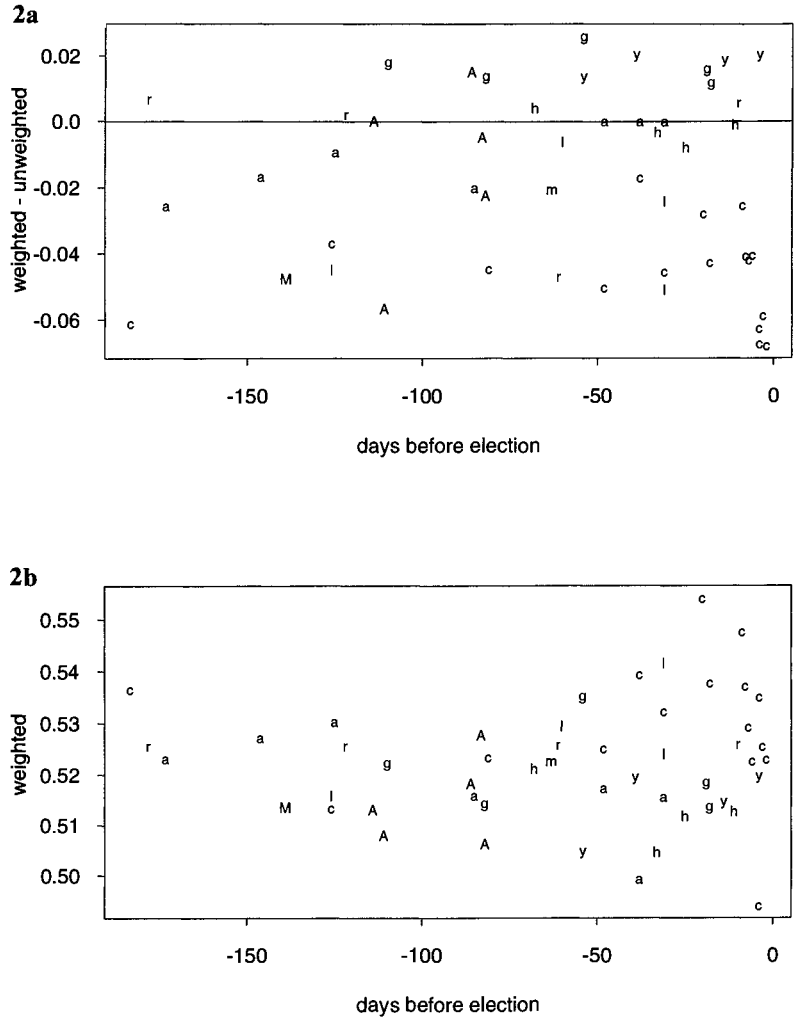

Figure 2. Fig. 2*a*, Effect of weighting on proportion of women. Fig. 2*b*, Proportion of women over time. *a* = ABC/*Washington Post*/ Chilton; *c* = CBS; *g* = Gallup; *h* = Harris; *l* = *Los Angeles Times; m* = Media General/AP; *r* = Roper; *y* = Yankelovich. Capital letters correspond to polls whose weights depend only on sex. Polls with no weights or with all weights equal are not included. All figures are for registered voters (or likely voters when registration was not asked). All polls are from the 1988 presidential election campaign, with details presented in Gelman and King (1993).

Finally, figures 3*a* and 3*b* present the corresponding results for the proportion of registered voters, surveyed in 1988, who claimed to have voted in 1984. Once again, the effect of weighting can be substantial, depending on the survey organization and the particular poll. Even after weighting, however, the surveys and organizations differ dramatically; for example, the CBS and Gallup polls give values about 5 percent lower than Chilton and *Los Angeles Times* polls. Figure 3, therefore, shows both the effects and the limitations of weighting. Any study in which this question is of interest must take account of the survey methods used beyond simply correcting using sampling weights.

Conclusion

Polling organizations generally have three stages of their operation for which policy must be set—random digit dialing, the interview, and adjusting the data—all of which create notable differences among the data reported by each group. To use data from these various sources concurrently (or, arguably, at all), or to engage in complex analysis of these polls, a researcher must have comprehensive knowlege of the methodology used to collect the various survey results. In this article we have attempted to outline the exact methods used in presidential election polls by major American pollsters, in as efficient and informative a format as possible, so that researchers have the information needed to make productive use of this valuable resource.

Our emphasis is on making optimal use of the survey data actually available rather than discussing the relative merits of various data collection schemes. Nevertheless, the detail provided in this article, thanks to the openness of our professional sources, should allow other scholars working in the field of survey methodology to pursue their research armed with specific examples of various methods. In this way, we hope the article helps bridge the gap between sampling theory and sampling practice.

Of course, these methods change over time; even the polls conducted in 1996 are sure to differ somewhat from our descriptions. Nevertheless, as long as the 1988 and 1992 survey data are available this technical detail will be useful. Indeed, public opinion research is the poorer because parallel information for earlier years is not available. It is our hope that the research community and professional pollsters in the future will ensure that details on polling methodology find their way into the public record, so that no more opportunities will be lost and their hard work will remain valuable for future researchers.

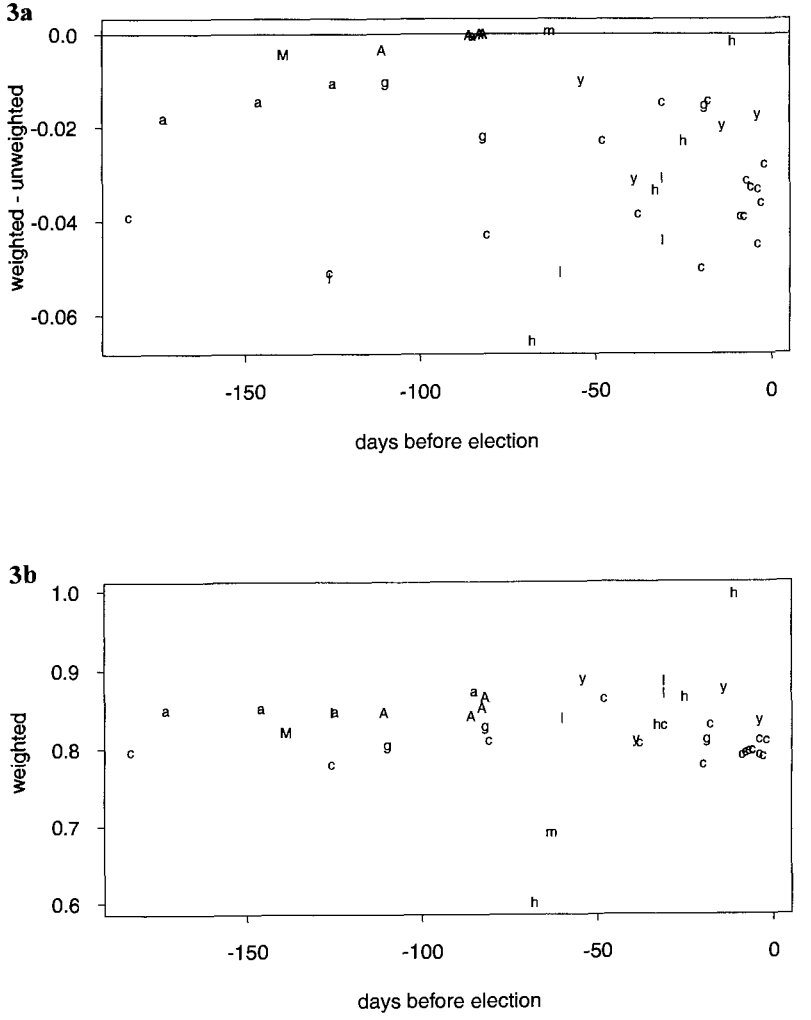

Figure 3. Fig. 3*a*, Effect of weighting on proportion who claimed to vote in 1984. Fig. 3*b*, Proportion who claimed to vote in 1984 over time. *a* = ABC/*Washington Post*/Chilton; *c* = CBS; *g* = Gallup; *h* = Harris; *l* = *Los Angeles Times; m* = Media General/AP; *r* = Roper; *y* = Yankelovich. Capital letters correspond to polls whose weights depend only on sex. Polls with no weights or with all weights equal are not included. All figures are for registered voters (or likely voters when registration was not asked). All polls are from the 1988 presidential election campaign, with details presented in Gelman and King (1993).

Appendix A

Methodology

The point of this article is to report on the process by which survey data were generated so that researchers will be able to use this valuable data in reliable ways. In this Appendix, we describe the process we used to gather this information—nonanonymous, nonconfidential, in-depth surveys. (In order to avoid infinite regress, we do not report the process used to develop this Appendix).

This article reports as fact information that was gathered from highly diverse sources, often under rather chaotic conditions. In no case could we get all details about an organization's methodology from one person or source. Although we feel secure that the article contains few, if any, inaccuracies, the rigor with which we were able to verify it varied from organization to organization. Therefore, although the article does not identify attribution directly, readers should remember that the information reported here is vulnerable to the limitations of human recall as well as the potential for miscommunication between individuals. In keeping with these warnings, this appendix identifies the sources we contacted at each organization and the information we got from each person. They are listed in rough order of contact. We also provide other information that may be useful to researchers.

The documents on methodology distributed by the Roper Center with **Gallup** polls are obsolete. The ones we reviewed all described Gallup's pre-1988 door-to-door polling methods. Although Jack Ludwig did send us a document on methodology, most of the information came directly from phone conversations with him. Shari Weber provided us with documents on the weighting for likelihood to vote. To ensure the accuracy of our information, our description of Gallup's procedure was faxed to Ludwig; he reported no flaws.

Most of our description of the **Harris** procedure for random-digit dialing and interviewing comes from a 19-page memo on the Louis Harris and Associates National Telephone Sample provided by executive vice president David Krane and dated April 1993. Nevertheless, numerous more specific details came from conversations with him directly. Also, the document contained little on weighting; most of that information was provided by Robert Spanski and Harris's academic consultant, statistician Martin Frankel. We faxed Krane a copy of our description of Harris's procedure, and he concurred with our description.

The documents on methodology distributed by the Roper Center with **CBS** polls also are obsolete. Most of our information on CBS polling came from conversations with Kathleen Frankovic and Bala Ramnath. Frankovic generally provided the information on interviewing; Ramnath outlined weighting. They contributed equally to the information on random-digit dialing. We also were provided with a series of CBS internal memos on the Probable Electorate weighting procedure. Our description of CBS methodology was reviewed over the phone with Frankovic and Ramnath on a few occasions. We also faxed Frankovic and Marla Kaye a copy of our description; they helped us repair one error and otherwise added to the clarity of our description.

For the **Washington Post** surveys, Sharon Warden from the **Post** gave us basic information, but the bulk of our detail comes from Chilton (Kristen Conrad), which actually conducts the poll, and Marketing Systems Group (Amy Starer and Dale Kulp), which designs and maintains the GENESYS Sampling System used by Chilton. Conrad provided the detail on interviewing. She also described the weighting, supplemented by Warden. Finally, Starer provided GENESYS's Random-Digit Dialing methodology, described in a 39-page document called "GENESYS Sampling Systems Methodology" and supplemented by telephone conversations with her. We faxed Marketing Systems Group president Dale Kulp our description of the GENESYS system, and he made a few clarifications.

Most of the **Roper** information came directly from Burns W. Roper. Additional details, including those on weighting, came from Peter Case. Roper later faxed us a four-page document on the Roper methodology, which matched the information we had received verbally. We faxed Case and Brad Fay a copy of our description of Roper methodology; they clarified our phrasing in a few cases.

The **1988 Associated Press** information on interviewing and weighting was provided in brief conversations with Steve Shaw of Media General, who conducted the polling for that year. The information on how Media General's phone number lists were generated came from a fax sent by Survey Sampling, Inc., which required follow-up questions answered by Kathy Walenczyk of SSI.

Our source for the **1992 Associated Press** information was Dan Soulas of ICR (although much of the random-digit dialing information came from Marketing Systems Group). Upon his request, we sent him our description of Chilton's methodology; he then outlined to us the differences between the two organizations. He also sent two documents, one called "EXCEL: National Telephone Omnibus Study," which allowed us to check numerous details against our Chilton description, and one called "EXCEL Weighting Process." He answered several follow-up questions.

Information on the *Time* Magazine/CNN polls taken by **Yankelovich Partners** was made available after an earlier draft of this article was submitted to *Public Opinion Quarterly*. All information on Yankelovich methods came from a short telephone conversation with Hal Quinley. The information on how Yankelovich's phone number lists were generated came from a fax sent by Survey Sampling, Inc., which required follow-up questions answered by Linda Piekarski of SSI.

Of the organizations we approached for this project, the *Los Angeles Times* was the only one that did not assist us.

In addition to the precautions to ensure accuracy reported above, we mailed drafts of this article to each primary source, announcing our intention to publish the details, and asked them to make one final review of our description and report any errors that slipped through the writing process. Almost every source responded to this additional request for assistance.

Appendix B

Methodology of Roper Polls

Roper does not use telephone polling for presidential election surveys. It includes preelection poll questions in its face-to-face surveys, conducted with the following procedure.

1. Counties are divided up into the nine census divisions, with Alaska and Hawaii excluded; these divisions are ordered from east to west.
2. Within these geographic areas, counties are divided up by "density," or degree of urbanization, according to whether the Office of Management and Budget describes them as metropolitan, producing 18 strata.
3. Roper decides how many Primary Sampling Units (PSUs) out of 100 will be distributed to each stratum, with each given a number of PSUs equivalent to this proportion of the contiguous U.S. population (numbers are rounded to the nearest integer).
4. Within a particular stratum, counties are listed from largest to smallest. Each county is given a measure of size (MOS) corresponding to the number of households it contains, as recorded by the Census Bureau. It also then is given a cumulative measure of size, equivalent to its number of households plus the number of households contained in all counties listed before it. The cumulative MOS for the last county in the stratum, then, is equal to the number of households in the entire stratum.
5. The total number of households in the stratum is divided by the number of PSUs indicated by step 3 above, creating a sampling interval. For example, if 10 PSUs are needed from a stratum with 5,000 households, then the interval is 500. (In practice, the number of households in each stratum is much larger than 5,000.).
6. A random number within the sampling interval is generated. In our example, this would be a number between 1 and 500. The county with the household indicated is chosen. So if the random number generated were 250, whatever county had the 250th household would be selected (i.e., the first county with an MOS larger than 250 would be picked).
7. The quantity of the sampling interval then is added to the random number. In our example, this would give us 250 + 500 = 750. This process is repeated once for each additional PSU needed, giving us 1,250, then 1,750, and so on. In each case, the county with the corresponding household is assigned a PSU. Note that, although 100 PSUs are selected in this way, only 97 counties are selected, since some populous counties get more than one PSU.
8. Within a selected county, block groups are listed numerically by census tract, which in turn are listed numerically within places and county subdivisions.[30] The arrangement of places and subdivisions is largest to small-

30. In 1988, many households were not contained in block groups. They were however, contained in Census Enumeration Districts, which were treated as block groups for this procedure.

est. The first block group is selected using an interval method parallel to that used to select the county; the second block group is one-half the county size from the first. (Because they were ranked in size order, one block group will tend to be from the larger towns and cities, and the other from smaller places and rural areas.)

9. Although the distribution of PSUs across counties is conducted once per census, the selection of block groups is repeated frequently, as block groups get used up in repeated surveys.
10. When the year's first survey is being prepared, two blocks within the block group are selected using an interval method parallel to that described above. The interviewer responsible for a block group is assigned the two blocks, one a block for interviews conducted at any time and one for interviews conducted only at night or on the weekends. He or she is given a map for each block, with the starting location chosen arbitrarily (usually the upper-right-hand corner) and the route for each block indicated on a map. The route is chosen systematically by the Roper organization. In subsequent surveys, however, the interviewer picks up where he or she left off in each block, rather than returning to the same corner of the map. When a block is exhausted, an adjacent block is selected.
11. Each interviewer is assigned to find 20 respondents, producing an approximate sample of 2,000 people. In practice this varies slightly; for example, responses sometimes must be discarded because of coding problems.
12. The interviewer has to fill two quotas in getting these 20 respondents, as described at the end of Sec. III.C.
13. Interviewers proceed along the route until all quotas are filled, trying every residence in a particular structure before moving to the next dwelling (such as an apartment building, starting with the first floor and then moving up).

The remaining Roper details are included in the text of this article.

References

Backstrom, Charles H., and Gerald D. Hursh. 1963. *Survey Research.* Chicago: Northwestern University Press.

Brady, Henry E., and Gary R. Orren. 1992. "Polling Pitfalls: Source of Error in Public Opinion Surveys." In *Media Polls in American Politics,* ed. Thomas E. Mann and Gary R. Orren. Washington, DC: Brookings Institution.

Brick, J. Michael, Joseph Waksberg, Dale Kulp, and Amy Starer. 1994. "Bias in List-Assisted Telephone Samples." Paper presented at the conference of American Association of Public Opinion Research, Boston, May 12–15.

Deming, W. E., and F. F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *Annals of Mathematical Statistics* 11:427–44.

Fletcher, James E., and Harry B. Thompson. 1974. "Telephone Directory Samples and Random Number Generation." *Journal of Broadcasting* 18:187–91.

Frankovic, Kathleen A. 1992. "Technology and the Changing Landscape of Media Polls." In *Media Polls in American Politics,* ed. Thomas E. Mann and Gary R. Orren. Washington, DC: Brookings Institution.

Gelman, Andrew, and Gary King. 1993. "Why Are American Presidential Election

Campaign Polls So Variable When Votes Are So Predictable?" *British Journal of Political Science* 23:409–52.
Kalton, Graham. 1983. *Introduction to Survey Sampling*. Sage University Paper 35. Beverly Hills, CA: Sage.
King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press.
Lepkowski, J. 1988. "Telephone Sampling Methods in the United States." In *Telephone Survey Methodology*, ed. Robert M. Groves et al. New York: Wiley.
Potthoff, R. F., K. G. Manton, and M. A. Woodbury. 1993. "Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks." *Journal of the American Statistical Association* 88:1197–1207.
Roslow, Sydney, and Laurence Roslow. 1972. "Unlisted Phone Subscribers Are Different." *Journal of Advertising Research* 10:204–7.
Salmon, Charles T., and John Spicer Nichols. 1983. "The Next-Birthday Method of Respondent Selection." *Public Opinion Quarterly* 47:270–76.
Troldahl, V. C., and R. E. Carter. 1964. "Random Selection of Respondents within Households in Phone Surveys." *Journal of Marketing Research* 1:71–76.
Waksberg, Joseph. 1978. "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association* 73:40–46.
Wright, Gerald C., Robert S. Erikson, and John P. McIver. 1985. "Measuring State Partisanship and Ideology with Survey Data." *Journal of Politics* 47:469–89.