Procedures for Assembling the First and Second Data Sets

Congress, the Press, and Political Accountability

R. Douglas Arnold

The first data set shows how a random sample of twenty-five local newspapers covered a random sample of twenty-five representatives for nearly two years. The data set, which consists of 8,003 articles, contains every news story, editorial, opinion column, letter, and list in a local newspaper that mentioned a local representative between January 1, 1993, and November 8, 1994.

The second data set shows how twelve newspapers — a random sample of six newspapers and six competing papers from the same cities — covered a random sample of six representatives for nearly two years. The data set, which consists of 2,175 articles, contains every news story, editorial, opinion column, letter, and list that mentioned a local representative between January 1, 1993, and November 8, 1994.

Selecting Newspapers

The sample of newspapers for the first data set was selected in 1994. At that time there were 1,567 daily newspapers in the country with combined circulations of 57 million copies.[1] Eighty-eight of these newspapers had electronic archives for 1993 and 1994 that were searchable through the DataTimes division of the Dow Jones News Service or the Nexis service of Reed Elsevier's Lexis-Nexis. My aim was to draw a sample of these eighty-eight newspapers that was a reasonable approximation of the universe of all daily papers. The good news was that the eighty-eight papers included 38 percent of the total daily circulation in the country, despite the fact that they represented only 6 percent of all daily papers.[2] This followed from the fact that a majority of citizens read a newspaper with a daily circulation of more than 100,000 copies, and large newspapers were overrepresented among the eighty-eight papers. The bad news was that smaller newspapers were underrepresented in the electronic archives, and smaller newspapers tend to serve small cities and rural areas. In order to draw a sample of newspapers that was representative of what the average citizen reads, I rank-ordered the 1,567 papers according to circulation, and then grouped the papers into approximate sextiles so that each group represented about one-sixth of the total daily circulation in the country. I then highlighted the eighty-eight archived papers within the various sextiles. Given that the two lowest sextiles contained only seven of the eighty-eight papers, I

---

[1]These figures exclude three papers that do not publish local editions — *USA Today*, the *Christian Science Monitor*, and the *Wall Street Journal*.

[2]The eighty-eight newspapers are distributed among the sextiles as follows: (a) 8 of 11 papers of greater than 500,000 circulation are included, representing 77 percent of the circulation in the highest sextile, (b) 24 of 30 papers of between 250,001 and 500,000 are included, representing 81 percent of the circulation in the fifth sextile, (c) 33 of 73 papers of between 100,001 and 250,000 are included, representing 51 percent of the circulation in this sextile, (d) 16 of 130 papers of between 50,001 and 100,000 are included, representing 15 percent of the circulation in this sextile, (e) 6 of 236 papers of between 25,001 and 50,000 are included, representing 3 percent of the circulation in this sextile, and (f) 1 of 1,087 papers of less than 25,000 is included, representing less than 1 percent of the circulation in the lowest sextile.

combined these two sextiles into a single group. I then randomly selected five papers from each of the five groups.[3] Table A lists the twenty-five newspapers, grouped into the six circulation sextiles.

The twenty-five newspapers were a diverse lot. The sample included large national papers like the *Los Angeles Times* and the *Boston Globe*, mid-sized papers like the *Hartford Courant* and the *Tulsa World*, and small-city papers like the *Rock Hill Herald* (South Carolina) and the *Lewiston Morning Tribune* (Idaho). Although very small newspapers were necessarily absent, the inclusion of a few small papers allows one to determine if small newspapers covered House members differently than large newspapers. The sample also contained various types of papers and not just the most celebrated newspapers in the country. It included tabloids, such as the *Chicago Sun-Times* and *Newsday*, rather than their highbrow competitors, the *Chicago Tribune* and the *New York Times*; it included the upstart *Washington Times* rather than the *Washington Post*. The sample was also geographically diverse, with newspapers from eighteen states and the nation's capital. The newspapers sold nearly seven million copies daily — about 12 percent of the nation's total daily circulation. The sample was not as diverse as one would have obtained if one selected a stratified random sample from the complete list of 1,567 newspapers. Taking that route, however, would have required searching most newspapers manually, which would have necessitated a much smaller sample and a much shorter time frame.

The sample of newspapers for the second data set was chosen by first identifying each newspaper in the first data set for which there was a competing paper from the same city. Eight cities — Boston, Chicago, Los Angeles, Phoenix, San Francisco, Seattle, Tucson, and Washington — had competing papers with electronic archives. I randomly selected six of these papers.[4] The second data set contains twelve papers, six from the first data set and six competing papers (Table B).

Selecting Representatives

Prior to selecting the twenty-five newspapers, I first identified the primary circulation area for each of the eighty-eight papers, and then identified the congressional districts within each circulation area. Matching newspaper circulation with congressional districts is more art than science because newspapers do not disclose circulation data by district. The matching was done with two maps, one identifying the locations

---

[3]The original sample included the *Bangor Daily News* and the *Evansville Courier*, both from the third sextile. After using their computerized archives for a while, I realized that the two data sets did not actually begin until the middle of 1993. I then substituted the next two random choices from this sextile, the *Norfolk Ledger Star* and the *Bloomington Pantagraph*, both of which included coverage from the start of 1993.

[4]The *Arizona Daily Star*, *Boston Herald*, *Chicago Tribune*, *San Francisco Examiner*, *Seattle Post-Intelligencer*, and *Washington Post*. The two competing papers that were not selected were the *Arizona Republic* (Phoenix) and the *Los Angeles Daily News*.

of a state's newspapers, the other identifying the locations of a state's congressional districts. For each newspaper, I identified from one to fifteen districts as being within the primary circulation area by using four decision rules: (a) include each district that includes any part of a newspaper's home city; (b) include each district that includes a significant portion of a city's suburbs; (c) avoid matching a suburban district with a city's newspaper if the suburban area has its own newspaper included among the eighty-eight papers;[5] (d) avoid crossing state lines unless a metropolitan area is heavily concentrated in a neighboring state.[6] These four rules identified 213 districts as within the primary circulation areas of the eighty-eight newspapers; ninety-one districts were within the primary circulation areas of the twenty-five selected papers.[7] I then selected randomly one representative for each of the twenty-five newspapers. Table A lists the twenty-five representatives and the newspapers that covered them.

The twenty-five House members were reasonably representative of the whole House. The match in party and seniority was especially good. Fourteen representatives in the sample were Democrats (56 percent), just shy of their actual percentage in the House (60 percent). The median representative in the sample was elected in 1986, as was the median member of the House. Two representatives in the sample, James Bilbray and Larry LaRocco, ran for reelection and lost, exactly matching the percentage for the whole House. Retiring members were underrepresented (only Romano Mazzoli), women were underrepresented (only Barbara Kennelly), and black members were overrepresented (Ronald Dellums, Louis Stokes, and Albert Wynn). Two of the twenty-five richest districts in the country made the list (CA24 and NY3); none of the twenty-five poorest districts did.

Selecting Articles

Computerized text searching allows one to examine directly a newspaper's archives, searching for every mention of a representative's name. One is not dependent on some indexer noticing a representative's

---

[5]This restriction applied only to the two largest metropolitan regions, New York and Los Angeles; it is the principal reason why my list matching congressional districts with newspapers is shorter than lists published elsewhere (e.g., *Congressional Districts in the 1990s* (Washington: Congressional Quarterly, 1993)). The restriction removed twelve districts in California from the two Los Angeles papers, assigning them to the *Orange County Register*, the Riverside *Press-Enterprise*, and the *San Diego Union Tribune*. It removed five districts in New Jersey and five districts on Long Island from New York City's paper, assigning them exclusively to the *Bergen Record*, the *Newark Star-Ledger*, and Long Island's *Newsday*. The purpose of the restriction was to ensure that my study focused on representatives that editors were most likely to consider newsworthy, allowing editors in large metropolitan areas to assume that their suburban competitors would take primary responsibility for more distant representatives.

[6]The five exceptions are the metropolitan areas surrounding the river-front cities of Cincinnati, Kansas City, Louisville, St. Louis, and Washington, the home cities for seven of the eighty-eight newspapers.

[7]The number of congressional districts listed in Table A totals 92 because it counts twice Arizona's second district, located within the primary circulation areas of both the Phoenix and Tucson papers.

name and considering it important enough to include in an official index. Computerized searching, however, requires some skill. Not only is there no index, there is no simple command that can locate every article that mentions a particular legislator. The basic problem is that legislators are known by variants of their names, and they share their various names with others. Representative Baker may be referred to as Richard H. Baker, Richard Baker, Rich Baker, Rick Baker, Dick Baker, or Representative Baker (and in a list as Baker, Richard H.; Baker, Richard; Baker, R.H.; Baker, R.; Baker, D.; or Baker); he should not be confused with Dick Baker who just won the fifth-grade spelling bee, Rick Baker who scored twenty points in high school basketball, Richard Baker who was a pall bearer at a local funeral, or Baker, Richards, Smith, and Wollensky who represent a firm filing for bankruptcy.

I began the project by experimenting with various search commands in order to devise efficient search routines for locating articles about legislators. I eventually developed a two-stage process. In the first stage, I searched a newspaper's archives for any mention of a representative's last name occurring within three words of any mention of either his first name or any plausible nickname. I then downloaded electronic versions of these articles, printed them with the representative's last name in boldfaced capitals, and manually removed those articles that did not actually mention the representative. In the second stage, I searched a newspaper's archives for any articles that mentioned a representative's last name *without* mentioning his first name or nickname within three words of his last name.[8] Of course, most of the citations from this second search should be errors, so I simply scanned them on the computer screen, looking for genuine references to the representative. If I found substantial numbers of genuine citations, typically because a newspaper publishes roll-call lists without first names, I downloaded the genuine articles, letters, or lists. If I found only a few genuine citations, I did not bother to download them. Most of the occasional references were in letters to the editor that referred simply to Representative Smith; these occasional references never amounted to more than two percent of a representative's total citations.

The regional editions of some newspapers and the successive timed editions of others presented different challenges. Newspapers did not archive the complete text of each different edition they produced. They archived the complete text of one edition, and any articles that may have appeared in a different format in some other edition. The *Houston Chronicle*, for example, archived its "Two Star" edition, and then included any articles from the later "Three Star" or "Four Star" editions that were not identical to the ones in the earlier edition. For a late breaking story, we might find a short story with a quotation from Representative Archer in the early edition, that was then supplanted by longer stories in the next two editions. Since I assume that few Houstonians read multiple editions of the daily newspaper, I removed all articles that were essentially early versions of a day's final story. The *San Diego Union-Tribune*

[8]Actually, I used a variety of methods to conduct the second-stage searches, especially for representatives with common names (Baker, King, Quinn) or with names that are common words (archer, baker, stokes). Since I was mostly looking for lists of roll-call votes, and such lists generally included several area representatives, I searched for a representative's last name within 100 words of the last name of the representative in the adjacent district.

published as an all-day newspaper, with up to nine editions over a twenty-four-hour period. For newspapers of this type, I considered articles to be duplicates if they were repeated at any time over a complete twenty-four-hour cycle (i.e., even on the next day); I kept the longest version and discarded the rest.

Newspapers with multiple regional editions created their own problems. The *Los Angeles Times*, for example, published separate daily editions entitled Home, Orange County, San Fernando Valley, San Diego County, and Ventura County, as well as special sections for Glendale, Long Beach, San Gabriel Valley, Southeast, and Westside. The same basic article might appear in several editions but with local twists to meet local interests. A long article that focused on Representative Beilenson's efforts on immigration might appear in two slightly different versions in the Ventura and Valley editions, both of which circulated in his district, and then appear in other editions as a shorter article, giving less detail about Beilenson. Since I assume that few Californians read multiple editions, I removed from the sample all articles that were essentially duplicates of the longest version of a day's particular story. I retained the longest version because it was usually the one that circulated within the representative's own congressional district.

The search of twenty-five newspapers from January 1, 1993 to November 8, 1994 found 8,258 articles using the first search method. My assistants and I removed 455 articles because they were essentially duplicates of articles published in other editions and 96 articles because they mentioned someone other than the selected representatives.[9] This sample was then supplemented by the 296 articles obtained with the second search method. The final sample size turned out to be 8,003 articles.

The procedures used for finding articles in the second data set were identical to the procedures used for the first. The second data set consists of 2,175 articles — 1,053 from the six original papers and 1,122 from the six comparison papers. The second data set contains 2,064 articles obtained from the first search method. and 111 from the second method.

Coding Articles

Before assembling the first data set, my graduate assistant and I conducted a pilot study. We first developed a scheme for coding newspaper articles for information relevant to political accountability. Then we searched the 1993 archives of four local newspapers for any articles that mentioned four local representatives. Over the summer we went through a dozen or so cycles, first attempting to code some

[9]The newspapers with the most severe rates of duplicates were the *San Diego Union-Tribune* (26%), *Louisville Courier-Journal* (22%), *Los Angeles Times* (15%), *Houston Chronicle* (12%), *Rock Hill Herald* (7%), *Newsday* (6%), *Phoenix Gazette* (5%), and *Seattle Times* (5%). Twenty-three newspapers had at least one duplicate. The representatives with the most wrong citations were King (19%), Archer (5%), Baker (3%), Spratt (3%), and Quinn (2%). Twelve representatives had at least one wrong citation.

articles, then revising the coding scheme, then coding some more articles with the new scheme, and so on, until we were comfortable that we had developed a reliable coding scheme that several research assistants could use to code a large collection of newspaper articles. The complete coding instructions are reproduced in two documents: (a) Coding Newspaper Articles in the First and Second Data Sets, and (b) Issue and Bill Codes in the First and Second Data Sets.

The coding of the articles in the first data set took place during the summer of 1995, by which time I had drawn a sample, assembled 8,003 articles from twenty-five newspapers, and hired three full-time assistants to code them. My graduate assistant and I spent more than a week training the three undergraduate assistants to search for the kinds of information in which I was interested and to use the elaborate coding scheme that we had developed. All five of us then did several days of practice coding. We alternated coding identical articles by ourselves with group meetings in which we compared our individual code sheets and discussed why we diverged on particular variables. Training ended when everyone was comfortable that we shared a common view of each variable and when it was clear that we were coding the articles similarly. From that point on, the three coders worked alone, coding batches of articles on paper forms and then entering their results into a networked database.

In order to guard against the coders becoming too familiar with each newspaper's coverage, and in order to replicate the haphazard way in which most people read newspapers, I did not assign long runs of consecutive stories. Every story had an identification number, so the sample was easily divided into half according to whether story numbers were odd or even. Coders spent the first half of the summer coding odd-numbered articles — essentially every other story in a chronological listing.[10] A typical assignment might be twenty odd-numbered stories about Representative Archer during early 1993, followed by thirty odd-numbered stories about Representative Baker during late 1993, followed by twenty-five odd-numbered stories about Representative Crapo during the 1994 campaign, and so on. After finishing all the odd-numbered stories, the coders received similar assignments for the even-numbered stories. Each coder read and coded articles from each of the twenty-five newspapers.

I arranged the coding assignments so that 4.5 percent of all articles would be coded twice. I told the coders on day one that their assignments would overlap, but I made daily assignments in a way that made it virtually impossible for the coders to know which articles would be (or had been) assigned to a second coder. Coding some articles twice served three purposes. It provided strong incentives for the coders to do their best work. It allowed me to check their work week-by-week to make sure that they continued to use similar standards throughout the project. It provided a good sample of articles for calculating the degree of intercoder reliability.

---

[10]The identification numbers place the articles from search one in chronological order followed by the articles in search two in chronological order; missing numbers indicate articles that we removed because they were duplicates or because they failed to mention the representative.

Coders employed sixty-eight variables to summarize the content of each story. Two variables — the representative's identification number and the story's identification number — uniquely identified each story; these variables were sufficiently important that my graduate assistant checked and corrected any entry errors for these variables. A third variable required that coders make an educated guess; modest disagreements were therefore the norm. For each of the remaining sixty-five variables, I calculated the frequency with which pairs of coders disagreed over the 23,530 individual decisions (362 double-coded articles times 65 variables).

Comparing the coding decisions for all double-coded articles reveals a high degree of intercoder reliability. The coders disagreed on only 6.4 percent of all decisions; the disagreement rate for the median variable was 3.9 percent. As one might expect, disagreement was minimal for variables that summarized simple facts. The coders disagreed 1.1 percent of the time on whether a story was accompanied by a photo of the representative, 1.4 percent on whether a story mentioned NAFTA, and 1.9 percent on whether a representative was identified as the chair or ranking minority member of a committee. They disagreed more widely on matters that required judgment and for which there were several acceptable codes. They disagreed 28.2 percent of the time on what was the most prominent policy issue mentioned in a story. This rate was not surprising given that coders could select a first, second, and third most prominent issue, and they sometimes disagreed on which one deserved top billing. More frequently, they disagreed on which of the 214 possible issue codes best described the most prominent issue; their menu included many similar codes, including thirty-two codes for various types of governmental expenditures and thirty-four codes for various aspects of defense and foreign policy. The real news was not that they disagreed so much, but that they disagreed so little.[11]

The final data set contains only one version of the double-coded articles. I retained all articles for the coder who appeared to be most accurate (coder two), deleted all double-coded articles for the coder who appeared least accurate (coder three), and retained the double-coded articles for coder one whenever coder three was the other coder. The final data set contains 3,226 articles coded by coder one (40 percent), 3,690 by coder two (46 percent), and 1,087 by coder three (14 percent).

The procedures used to code articles in the second data set were identical to the procedures used for the first. Most of the coding for the second set took place after the first data set was complete.

---

[11]Other measures of intercoder reliability exist that are more sophisticated than the rate of disagreement. Some measures use as a benchmark the amount of agreement that one should expect purely by chance while others take into account the extent of disagreement (miss by an inch or miss by a mile). I was sufficiently pleased with the overall agreement among the coders that I did not calculate a full repertoire of alternative measures. I did spend several days reading the double-coded articles along with the codes provided by each coder to determine how and why they diverged on specific variables. My sense was that most of the disagreements were relatively minor and that they were not likely to interfere with my basic analysis.