Codebook for Member Level Data Sets
Congressional Whip Count Database

C. Lawrence Evans
College of William and Mary
January 9, 2012

The purpose of this codebook is to provide users with the background information necessary to interpret the contents of the member level data sets, especially the whip poll positions. Section I explains the basic structure of these data files and the individual file names. Section II describes the variables in the member level data sets. Section III explains the number and order of the whip poll variables. Section IV describes the values coded for the poll variables. And Section V provides detailed information about the process through which these coding decisions were made.

Users of the database can contact me if they have remaining questions and I will do my best to respond. But please consult the contents of this codebook before downloading and analyzing any of the member level data sets.

I. File Structure and Names

The member level data sets are all in Excel 97-2004 format (xls). Each member data set includes the individual responses for a single polled question. (Details about the date of the count and the text of the polled question are provided in the whip count index elsewhere on this website.) For the member level data sets, the names of the Excel files all have the same basic structure. Each one begins with the whip count identifier, where the first letter is for party (D for Democrats and R for Republicans), the next two entries are the year of the count (55 for 1955, 56 for 1956, and so on), and the next two entries are the count number within the relevant year (01 is the first count, 02 is for the second, and so on). So a file named D7744 would be the 44th count conducted by the Democratic leadership in 1977, and R7904 would be the fourth count conducted by Republicans in 1979. As much as possible the count numbers capture the temporal order in which polls were conducted. Within the Excel file names, and following the count identifier and separated by a period, there generally is a word or two or other entry reflecting the substantive contents of the relevant polled question. These substantive entries were an additional protection against mislabeling and other errors during the process of database construction and can be ignored by users.

II. Variables in the Member Level Datasets

congress   Congress number

stateno   ICPSR state code

statename   State name

district   Congressional district number (97, 98, etc. indicate state-wide
                      districts)

partyno   ICPSR party number (100 indicates Democrat, 200 indicates
                      Republican, 328 indicates independent)

idpoole	ICPSR member identification number (as corrected by Keith Poole and Howard Rosenthal, www.voteview.com)
idicpsr	ICPSR member identification number (uncorrected)
name	Member name
poll1…polln	Member position on whip count, drafts 1 …. n (for details see below)

III. Whip Poll Variables (poll1, poll2, etc.)

The "poll" variables in the member level data sets are for whip count responses. Typically, but not always, there are multiple whip poll drafts per polled question, denoted in the datasets as separate variables, that is, by poll1, poll2, poll3, …. These multiple poll variables are necessary because member positions often change during the course of a single whip count, and as much as possible these position changes are captured in the database.

The maximum number of positions taken by a single member generally determines the number of poll variables for a count. In other words, if the largest number of positions taken by an individual member during a whip count was four, there would be four poll draft variables (poll1, poll2, poll3, and poll4) in the relevant dataset to fully capture all of the member positions for that question.

Fortunately, it was usually feasible to construct the poll variables in a manner that reflects fairly well the temporal order in which the positions were taken. For the most part, then, the position a member takes on poll2 occurred after the position reported in poll1, the entry for poll3 was reported after the one for poll2, and so on. Along those lines, most often the entries reported for a particular poll draft appear to have been tabulated by the leadership at around the same time and reported in the same document or type of document. For example, all the entries for poll1 might be from the same whip pad or notebook, the entries for poll2 might be position updates jotted down next to or over these initial entries, and poll3 might be for further position updates scratched on several (similar) lists of members targeted for late-minute lobbying by the leadership.[1]

That said, users should be cautioned against assuming that the positions for poll1 *always* occurred prior to those in poll2, or poll2 responses *always* occurred before poll3; or that the positions for a particular poll draft were *always* taken at exactly the same time across individual members. Please keep in mind that these data are coded from what initially

---

[1] When, as was usually the case, it was possible to discern the basic temporal order of position changes in the archival materials for a whip count, and these records include position updates or changes for some members, but not for others, then positions for these other members on the subsequent poll draft (capturing the updates) are carried over from the previous draft.

were working congressional staff files, and, not surprisingly, the archival record is often disorganized and fragmentary. Systematic and comprehensive evidence about the timing of member positions was not always available, and in such instances we simply did out best at ascertaining the temporal sequence.

IV. Values for the Poll Variables:

0   **Not a member** of the House or the relevant party caucus at the time of the whip count; because the Speaker seldom votes or participates in whip counts, he/she is coded as zero unless a specific position is attributed in whip records

1   **Yes** to the question as posed; often, but not always, the position of the relevant leadership

2   **Leaning yes** to the question as posed; or probably yes, or probably for

3   **Undecided**

4   **Leaning no** to the question as posed; or probably no or probably against

5   **No** to the question as posed

6   **Yes, if needed**

8   **Aggregate results are available for the relevant zone or region**, but individual-level positions for a member are absent in the archival record[2]

9   **No response**; usually because of a blank next to the individual's name or member is identified as nonresponsive on a list, but also if an entire zone/regional card or sheet is missing or absent and there is no compelling evidence that responses once existed for these members (see 98 below)

10   **Member expected to be absent** for the vote (sometimes explicitly delineated in the whip count records; this is different from being nonresponsive or unavailable for the whip count)

11   **Other response;** typically a written note describing a position as opposed to one of the standard response categories

---

[2] The reason for the omission appears to vary by time period. For the 1960s, fairly often whip zones from the American south refused to provide the Office of the Democratic Whip with individual-level count results, and instead provided zone aggregates (the number responding yes, undecided, and so on, but not associated with particular names). For the 1970s and 1980s, in contrast, the reason generally is that individual level responses were reported to the office but are simply missing from the relevant archival files.

12   **No comment** (sometimes the whip will explicitly note that a member responded as not wanting to respond or comment)

13   **Ill or out of town at time of the poll** and no position reported (care was taken when distinguishing such situations from ones where the member is reporting back that he/she will not be participating in the roll call

98   **Missing zone sheets or summary sheets**; occasionally, whip results were not available in the archival record for an entire or zone or region and there is evidence elsewhere in the records that these responses did exist at one time. Most likely, the materials were lost or discarded prior to processing and transmittal to the archive. This category was used sparingly. Missing responses were coded as "9" unless there was concrete evidence that the materials in fact existed sometime during the whip counting process

99   **Unclear or ambiguous** and does not fit one of other categories; for instance, the markings may be illegible or a member is listed in columns for multiple positions

V. The Position Coding Process:

All of the member positions in this first release of the Congressional Whip Count Database were independently coded multiple times by me and undergraduate members of my research team. Altogether, the work took place over a number of years and involved nearly 40 student assistants, it primarily centered on three independent teams of approximately ten research assistants each, all of whom received extensive background information about the whip process, the nature of archival research and whip count records, and the coding criteria immediately after joining the project. The coding process outlined in this section, it should be emphasized, grew out of extensive pretests that I conducted with an initial group of three student assistants before the project formally began.

The most systematic and extensive coding work took the form of two independent coding waves. After the two coding waves were completed, all member position differences between them were identified and further analyzed, and these differences were reconciled. Subsequent to this process, a third team, reconsidered all of the data files in light of the relevant archival materials and recommended a small number of additional modifications. Throughout the coding process, the research assistants worked in close and regular consultation with me. When there were ambiguities or other difficulties during the coding process, I made all final decisions and thus am fully responsible for any remaining errors.

More concretely, the coding process went something like this.

At the beginning of the project, I personally contacted or visited a number of libraries and research centers that included the papers of former House whips and thus were likely

sources of records about past whip counts. On site, I searched through the relevant archival files and photocopied all relevant documents. Included, for example, were whip pads (comprised of three-by-eight inch cards on which the initial positions of members often are tabulated during whip counts), House "notebooks" (small pamphlets that include the names of all members and adjacent columns in which whip staff denote their positions), type-written sheets that list the positions of members (usually with separate columns for the main position categories), handwritten notes and sheets with the names of members and associated positions (often these notes were the result of telephone conversations), aggregate summaries that tabulated the ongoing results of whip counts, often by zone or region, and a host of other materials. Essentially, if I found a piece of paper in an archival file that associated a poll position with a member or tabulated whip count results, it was photocopied.[3] For the whip count data in this first release, approximately 10,000 sheets of paper were photocopied and transferred back to my office in Williamsburg, Virginia. (For additional information about the evidentiary base, please consult the memo about archival sources located elsewhere on this website.) I began the coding process by preparing the master list or index of polled questions and separating the archival materials into hard copy files, generally one hard copy file per whip count. As much as possible, in placing archival materials into hard copy files, I maintained the same file structure as had been used by the relevant whip's office, including the order in which individual documents had been placed.

Now consider the first coding wave. Here, I assigned whip counts to individual members of my research team, usually in sequential batches of four or five files/counts. The research team was comprised of William and Mary undergraduates and the selection process was highly competitive. Usually, about seven students applied for each available position on the team and I was able to select research assistants who were academically talented, careful and systematic in their work habits, and on the whole remarkably devoted to the project. After joining the team, my research assistants also were given detailed instructions about how to code poll responses, the kinds of complexities and ambiguities that can arise during the coding process, and other relevant background information about the project. Over the course of the project, about two-dozen of them were able to coauthor papers with me (all based on the project) that were presented (usually by the students) at annual meetings of the American, Southern, Midwest, and Western Political Science Associations. In other words, the team members were involved

---

[3] There is one exception. The raw archival files for House Democrats, 1977-1986, feature an immense quantity of information about member positions, with the relevant archival files often including whip pads and multiple copies of type-written position sheets compiled as the whip process progressed. Unfortunately, it was not possible to photocopy and code all of the whip pads in these files because of time constraints, and because, here, photocopying would have required that I permanently disassemble the pads, which would have been inconsistent with basic norms of archival research. However, based on an analysis of a sample of the whip files for this portion of the dataset, it is clear that the positions denoted in the pads almost always are included in the initial type-written position list for that whip count. For the most part, any differences occur when no position is recorded for a member on a pad, but there is a position on the first typewritten sheet. As a result, I am confident that my inability to photocopy and code all of the whip pads for this portion of the project does not have significant consequences for the accuracy or comprehensiveness of the data.

in diverse aspects of the broader research project and understood the goals of the research, the underlying topic, and the contents of whip archives extremely well.

Along with a batch of whip count assignments, the student coders were also given a hard copy of the relevant archival materials. I also provided the coders with a shell coding form in Excel format that, among other variables, listed the members of that party who were in the chamber during the relevant Congress and the dates within that Congress when they entered or left the chamber. In most Congresses, several members are elected sometime after the opening day or resign or die prior to final adjournment – critical information if we are to know precisely which members were subject to a whip count on a particular date.

The research assistant would then take a first cut at coding the whip counts in the assigned batch. Throughout the coding process, at any one point in time, my research team generally included about ten students and I encouraged them to consult their colleagues about any questions or coding difficulties they might have. They also were encouraged to speak with me about these questions. After a batch assignment (again, usually comprised of four or five counts) was completed by a research assistant, this individual met with me, usually for around 30 minutes but sometimes for an hour or more, and we discussed each archival file and the associated count data set in detail, considering possible position ambiguities and making any necessary judgment calls. Most important was settling on the number of independent poll drafts (poll1, poll2, and so on in the member level data sets) and also the proper temporal order of the poll drafts and positions. After these decisions were made, the research assistant entered the necessary modifications and transferred the completed Wave 1 Excel file to me via a protected folder on the William and Mary computer network. The students also drafted for my consideration memos that detailed all judgment calls that we made during the coding process. This full process was utilized for all of the hardcopy files and associated whip counts.

After this first wave was completed, the entire process was redone from scratch, and with a second team of research assistants, selected through the same basic procedure as the initial group. A few students from the first team also were on the second team, but they were not assigned whip counts or archival records during the second coding wave that they had considered during wave 1. As much as possible, I ensured that the two coding waves were independent of one another. Obviously, since the coders in both waves regularly consulted with me about their work, there was some possibility that what I had learned about a whip count or archival file during wave 1 might inform what I said to the wave 2 coder for that count. It should be emphasized, however, that the sheer quantity of archival materials and counts was immense and, without formally consulting the results of the first coding wave, I was generally unable to remember them with any precision.

After completion of the second coding wave, as each assignment batch was completed, the research assistants and I compared the results with the analogous data sets that had been produced during wave 1 and identified all differences. Unfortunately, conducting standard tests of inter-coder reliability across wave 1 and wave 2 was problematic

because there occasionally were differences in the number and structure of poll drafts across the two coding waves. But when there was agreement on the number of drafts and the basic structure of a file, inter-coder reliability across the two waves generally was very high, over 95 percent.

After juxtaposing the two versions of a data set for a whip count, the wave 2 coder and I then reconciled any differences between them in light of the archival materials and legislative context and decided what most accurately reflected the underlying evidence. These reconciled files were once again transferred to me for inclusion in the database.

As a final check, after completing and reconciling the two coding waves, all of the member level data sets were compared yet a third time against the archival materials by me and a third independent cohort of research assistants. This final check resulted in a small number of additional corrections, ensuring that the data are as accurate as possible.

As work on my book manuscript based on the project continues, I may make some modifications in certain of the data sets, but at this point I am confident about the accuracy of the poll variables and am convinced that any remaining changes will be minor. Of course, any updates will be detailed in this website as they are made. Moreover, if users of the database suspect or find any errors I hope that they will contact me with that information so that I can make any necessary corrections.