

# A Coding Methodology for Open-Ended Survey Questions\*

Melanie Goodrich<sup>†</sup>

Department of Politics  
New York University  
19 West 4th Street, Second Floor  
New York, NY 10012-1119

April 25, 2008

---

\*Prepared for delivery at the 2008 New Faces in Political Methodology conference. The author would like to thank Eric Dickson, Pat Egan, Catherine Hafer, Dimitri Landa, Michael Laver, Renan Levine, Jonathan Nagler, Andrew Owen, Shanker Satyanath, and Joshua Tucker for their thoughtful feedback and suggestions. This work is preliminary and the author requests that it not be circulated without her consent.

<sup>†</sup>[www.melaniegoodrich.com](http://www.melaniegoodrich.com)

# 1 Dissertation Project

In my dissertation, I develop a model of party identification that incorporates both identifiers and non-identifiers alike. The theory does this without assuming that the non-identifiers are either apolitical or ambivalent between the major political parties, both of which are the traditional explanations offered by scholars of American partisanship (Campbell et al. 1960, Fiorina 1981). Instead of party identification simply being a choice between the parties in the political system, I propose a process requiring two decisions. First, an individual must evaluate whether the partisan environment addresses those issues that are important to her. In other words, she must be able to identify issues that she consistently factors into her vote choice and that one or more of the parties consistently prioritizes and takes a position on. If the individual cannot do this, she will be not identify with any major party, and will thus be an Independent. If one or more of the parties is discussing issues that are important to the individual, then she will identify with the party which best represents both her priorities and her policy position.

In order to empirically test this theory, I need to be able to identify the issues that are important to individual voters and how much the political parties talk about these issues. To do the former, I use the raw responses to the most important problem (MIP) question from the National Election Studies (NES), which asks: "What do you think are the most important problems facing this country?" This question has been included, in one form or another, in every NES survey conducted in even numbered year since 1960. To ensure that the coding is consistent across years, the NES created a simple coding system in 1960 that categorized every response as referring to one of nine issue groups: (i) agriculture; (ii) economics, business, consumer issues; (iii) foreign affairs and national defense; (iv) government functioning; (v) labor issues; (vi) natural resources; (vii) public order; (viii) racial problems; (ix) social welfare. Each of these categories has subcategories which can

vary by election year, resulting in hundreds of subcategories.

Any attempt to condense all of the potential responses to a question as broad as the MIP question into nine categories chosen at one point in time is certain to result in groupings of issues that are both very general and that do not allow for the emergence of new issues over time. Both of these limitations are very problematic for my research purposes. Given that I am looking for evidence that either supports or refutes the idea that a voter will not identify with a major political party unless one or more of these parties is investing in the issues that are important to the voter, it is very important that I am able to identify all of the issues that voters are concerned about. This includes those issues that are infrequently discussed by politicians or rarely covered by the media. My need for a fine grained measure of the types of issues that concern the American electorate is the motivation for the computer assisted coding algorithm that I present in this paper. The primary innovation of this coding methodology is that **the algorithm allows the categories to originate from the responses themselves**. My goal is to develop a methodology that will effectively code the responses to the MIP question. In the future I plan to generalize my algorithm so that it can be used on responses to other open-ended survey questions.

## 2 Textual Analysis

There are a number of research projects in political science that attempt to quantify the information contained within text documents. The approaches vary based on whether the coding is performed primarily by humans or by machines as well as the characteristics of the document being analyzed. One type of document of interest to political scientists is the manifesto of a political party. The Manifesto Research Group calculated what percentage of the sentences of a given party's manifesto are devoted to addressing each of a

variety of policy categories (Budge et al. 2001, Klingemann et al. 2006). Another type of document is newspaper coverage of an event or an issue. The Kansas Event Data System (KEDS) allows researchers to generate event data by identifying the proper nouns, verbs and direct objects within verb phrases in newspaper coverage of the daily interactions of nation-states (Gerner et al. 1994, Schrodt 1998). The Policy Agendas Project uses the manual coding of media coverage to document the public debate surrounding a variety of issues in American Politics (Baumgartner & Jones 1993). Legislative speeches and bills contain enormous amounts of information and a number of scholars have used such documents to study the political agendas of legislative bodies (Purpura & Hillard 2006, Quinn et al. 2006). Still other researchers have developed methods for dealing for a variety of unstructured natural language text documents such as blogs, candidate web sites, and candidate speeches, to name a few. Hopkins and King developed the ReadMe software package which can process a large number of such documents and code them based on the prior hand coding of a small set of similar documents (Hopkins & King 2008).

The responses to the open-ended MIP question are different from the sorts of documents described above in a number of ways. First, the median length of a response is 12 words which is much shorter than a newspaper article, let alone a party manifesto or a candidate's speech.<sup>1</sup> Laver and his coauthors find that the standard errors on the estimates derived from their computerized textual analysis method get very large when the number of words dips below 1000 (Laver, Benoit & Garry 2003). There is no reason to believe that this finding is a result of the author's research design. Rather, it is an indication of the limitations of statistical textual analysis techniques with respect to documents that

---

<sup>1</sup>The NES has granted me access to a set of raw responses that includes 22,411 responses with 1 or more words. 99% of the responses have 150 words or fewer. The longest response is 862 words. The data set that I am working with is taken from NES surveys conducted between 1984 and 2000, with the exclusion of 1994. The absence of 1994 from my data collection is due to the fact that the facilities that the NES has had access to in the past for the purposes of storing the original questionnaires were not always ideal. The result of this is that some of the original forms were destroyed by raccoons, including the responses collected in 1994.

are relatively short in length. Second, the MIP responses vary considerably with respect to the degree with which they demonstrate proper grammatical structure, which makes using techniques that depend on the ability to recognize patterns within the text, like KEDS, inadequate. Finally, the primary characteristic of the responses to the open-ended MIP question that makes coding them different from previous studies is that it is not obvious ex ante **how many categories are needed** to properly categorize the information within each response nor is it obvious **what these categories should be**.

### 3 Open-Ended Survey Questions

Open-ended survey questions promise a wealth of information to the researcher who can properly process the responses. Unlike close-ended survey questions in which a respondent is asked to select the best answer from a finite list of options, open-ended survey questions do not restrict the respondent's ability to fully express herself. This limitation of close-ended survey responses is not particularly worrisome, for example, when asking a respondent her age because the range of potential responses is known at the time of the survey's design. Thus it is possible to account for all potential responses when creating the set of options from which the respondent will eventually choose her response. However if the surveyor wishes to query the respondent regarding the political issues that are most important to the respondent, the limitations associated with close-ended survey responses begin to impose a structure on the resulting data that may be undesirable. By asking an open-ended question to elicit this information, the surveyor eliminates the possibility of the respondent not being able to properly answer the question because the list of options does not include the issue(s) that are most important to the respondent. The very features that make open-ended questions desirable also make the responses challenging to analyze. In this paper, I work with a set of over 20,000 survey responses. Analyzing

these responses requires that I collapse what are essentially over 20,000 unique categories to a set of considerably fewer categories which are analytically useful. However, it is not immediately evident from the data in its raw form how many categories the responses can or should be broken into, let alone what these categories are. In addition, unlike other documents that political scientists have developed textual analysis methodologies for, such as party manifestos and newspaper articles, open-ended survey responses are very short and often lack proper grammatical structure.

## 4 The Most Important Problem Question

In March of 2008 the American National Election Studies released a report entitled “Problems with ANES Questions Measuring Political Knowledge” (Krosnick et al. 2008). The report addresses a number of problems that the ANES has recently uncovered with respect to open-ended survey questions in their surveys, including the most important problem question. These problems include evidence that interviewers failed to transcribe answers despite explicit instructions to do so, coding instructions that were “problematic” or “incorrect” and the non-existence of written records of the instructions given to coders prior to 2000.<sup>2</sup> Given the findings of this report, it is unlikely that there exists a record of how or why these nine categories were originally chosen for the most important problem question.

Even if such a record existed the categories remain problematic. For example, it is not clear why a coding system for politically relevant issues in American politics that is comprised of only nine categories would include as one of those categories, “Agriculture.” Another problem is rooted in the fact that these nine categories have been the same since they were first formed in 1960, which is presumably why abortion is coded as a

---

<sup>2</sup>The report also notes that the instructions that were given to coders were frequently given orally.

Public Order issue.<sup>3</sup> Thus, even before the release of this NES report, I determined that to fully take advantage of the information contained within these response it would be necessary to recode the answers by going back to the raw responses. My goal is to develop a methodology that will effectively code not only the responses to the most important problem question but also responses to other open-ended survey questions.

In this paper I will describe the computer assisted methodology that I have developed to identify the relevant issues of interest to respondents and to subsequently categorize each response as addressing one or more of these issues. In designing the algorithm, I had to contend with the fact that I initially knew neither what the issue categories were nor how many of them I would need. The primary innovation of the coding methodology presented in this paper is that the algorithm allows the categories to originate from the responses themselves. It would defeat the purpose of asking an open-ended survey question if the coding algorithm for the resulting responses required the analyst to first choose a set of the most politically salient issues and then to code all of the responses as addressing one or more of those issues chosen by the analyst.

Once I have identified the issue categories and I have coded each response for the issue categories it addresses, I am able to show the various issues that have become increasingly important (or unimportant) to the American electorate throughout the last two decades of the twentieth century, and how the level of importance has fluctuated over time. Ultimately I intend to measure the amount of resources each political party dedicates to the problems that members of the electorate care the most about, and this algorithm allows me to identify these problems. The methodology that I present in this paper is not only useful to other researchers as a result of the data generated from using it to code the response to the aforementioned NES question, it also provides guidance to other researchers who wish to unlock the information within open-ended survey responses in

---

<sup>3</sup>The landmark case legalizing abortion in the United States, *Roe v. Wade*, wasn't decided until 1973, 13 years after these categories were defined.

the future.

## 5 Coding Methodology

To identify the priorities of individual members of the American electorate, I have obtained the raw responses to the most important problem question asked in the American National Election Survey between 1984 and 2000.

This data set is comprised of over 20,000 responses. Some responses can be very long. The following are examples of some of the shorter responses:

1. "abortion"
2. "economy, welfare should be checked into better, choices of how tax money is spent."
3. "honey, i really couldn't tell you."
4. "moral problem – general moral decay; drugs; crime; poor education; divorce; alcohol."
5. "our national debt"
6. "our physical responsible and the rest will fall into place"
7. "taxes, eats up your income"
8. "the federal bonus they keep promising to give to wwii veterans"
9. "the life amendment"

One of the biggest advantages of this data is that the responses are open-ended. This means that the respondents do not have to to confine their responses to a small set of issues that have been chosen by the designers of the survey. However, the fact that these responses are open-ended is also one of the biggest challenges in using this data. It

means that before the responses can be broken into issue categories, the issue categories themselves have to be identified.

I have developed the following computer assisted methodology to identify the relevant areas of interest to the voters and to subsequently categorize each response as addressing one or more of these areas.<sup>4</sup>

*1. Create a list of all unique one, two, and three word answers to the MIP question. This will be done across all years. These words and phrases will be used by the analyst to identify **key terms**. A key term is anything that the analyst can identify as a single political concept. Words or phrases that could refer to multiple political concepts cannot be key terms.*

Of the 9 example responses from above, this step would identify “abortion”, “our national debt”, and “the life amendment.”

*2. Using this list, the analyst will create **issue categories** by manually grouping conceptually similar terms together. Each issue category will then be defined by the presence of one or more of the key terms that has been grouped together to form the issue category.*

Of the 9 example responses above, “abortion” and “life amendment” would be grouped together as belonging to the same issue category, and “national debt” would be grouped in a separate category. This process resulted in 71 different issue categories.

*3. Each response will then coded based on whether or not the response contains a key term from each issue category. It is possible for responses to be coded as more than one issue category.*

---

<sup>4</sup>The computerized portions of this methodology were implemented in Python.

*Whenever a key term is found in a response, the words comprising that key term will be removed from the response. In addition, the longest key terms will be searched for first.*

For example, if a respondent mentions “war on drugs”, her response will be coded as addressing the topic of drugs and not war because “war on drugs” is longer than “war” and because once the key term “war on drugs” is detected it cannot also be coded as “war” because it will have been removed.

*4. At this point, there will be some number of responses that cannot be coded. In this data set, there are approximately 1500 responses that fall into this category. The analyst will read each of these responses and first determine if there is a key term in that response that simply wasn't represented in the short responses described in Step 1. If there are one or more key terms, the analyst will assign the word or phrase that comprise the key term(s) to the appropriate existing issue category, or if necessary, create a new issue category. If there are issues mentioned in the response that the analyst can identify as referring to an existing issue category but that cannot be summarized using a key term, the analyst will manually code these responses without using the process described in Step 3. Finally, if there are no identifiable issues mentioned, the analyst will code the response as either “no response” or “un-codeable.”*

Of the 9 example responses given, “the federal bonus they keep promising to give to wwii veterans” contained a key term, “veterans”, that was not identified in Step 1. There was no issue category addressing veterans, so I created a new one. Two more responses, “honey, i really couldn't tell you” and “ our physical responsible and the rest fall into place” couldn't be coded using using the process described in Step 3, and will be coded as “no response” and “un-codeable,” respectively.

*5. Once the issue categories have been updated, repeat Step 3.*

6. Once all of the responses have been assigned to at least one issue category, or categorized as a “no response” response or a “un-codeable” response, the remaining words in all responses should be listed together and manually examined to determine whether any alternative spellings of key terms should be added to the issue categories. In addition, an attempt will also be made to identify any words that have a substantive meaning such that they should have been identified as a key term (or as a constituent of a key term), but were previously excluded from the issue category definitions because the word was not in any of the short responses.

7. Once the issue categories have once again been updated, repeat Step 3.

## 6 Preliminary Findings

I have completed Step 3, and am in the process of reading all of the uncoded responses as prescribed in Step 4. I have already identified 71 different issue categories. Below is an alphabetized list of those 71 issue categories. The ones in bold are among the top nine most frequently mentioned issue categories in at least one of the election years between 1984 and 2000:

- Abortion
- AIDS/STDS
- **Arms/Weapons**
- Big Government
- **Budget**
- Business
- Campaign Finance Reform
- Central America
- **Children (not schools)**
- Communism
- Cost of Living
- **Crime (not drugs)**
- **Defense/Military**
- Discrimination/Race
- **Drugs**
- Drunk Driving
- **Economy**
- **Education**
- **Employment**
- Energy
- **Environment**
- Equality
- Families (not values)
- Far East
- Farming
- Guns

- **Healthcare**
- **Housing**
- **Hunger**
- Immigration
- Impeachment
- Imports/Exports
- Insurance
- International Relations
- Iran
- Iraq
- Japan
- Justice System
- Kuwait
- Media
- Medicare/Medicade
- **Middle East**
- **Morality**
- Native Americans
- Nicaragua
- **Nuclear**
- Oil/Gas
- Overpopulation
- Persian Gulf
- Pornography
- **Poverty**
- President
- Puerto Rico
- Religion
- **Russia**
- Saudi Arabia
- Seniors (not Social Security)
- Sexual Preference
- **Social Security**
- Somalia
- South America
- Space Program
- Star Wars
- **Taxes**
- Teen Pregnancy
- Terrorism
- Unions
- Wages
- **War/Peace**
- Wealth Inequality
- **Welfare**

In Table 1, I list the top nine issue categories that are mentioned by respondents in each election year.<sup>5</sup>

[Table 1 about here.]

There are several things to note about the relative frequency with which certain issue categories are mentioned by respondents. First, the issue category *Russia* is the 8th most frequently mentioned issue category in 1984, but it never again appears in the top 9. In 1986 and 1988, *Russia* is ranked 21st and 24th, respectively, in respondent mentions, and

---

<sup>5</sup>It is important to note that because I have not completed all of the steps of the algorithm, the issue categories that are most frequently mentioned may change once the algorithm has been fully followed.

during the 1990's it is never ranked higher than 47th. That mentions of *Russia* should begin becoming less frequent in 1986 makes sense given that the Reykjavik Summit took place in October of 1986. This meeting between President Reagan and Soviet Premier Gorbachev marks when relations began to improve between the two super powers. Another noteworthy category is *Drugs*. First Lady Nancy Reagan began championing the "Just Say No" anti-drug campaign during her husband's presidency. In 1984, *Drugs* is ranked 32nd, but beginning in 1986 this issue category is frequently in the top nine, and subsequently it is never ranked lower than 10th, which makes sense given the events of the time. Finally, *Healthcare* doesn't break into the top nine until the 1992 presidential election of Bill Clinton, at which point it remains in the top nine for the remainder of the decade. The entry of *Healthcare* into the top nine coincides with the period during which First Lady Hillary Clinton chaired the Task Force on National Health Care Reform.

Recall that one of the nine ANES categories is Agriculture. I have identified a similar issue category, which I have named *Farming*, which never appears in the top nine. In fact, during the 1980s, the highest this issue category is ever ranked is 14th, and during the 1990's the highest it is ever ranked is 38th. This would indicate that a coding scheme with only nine categories that allocates an entire category to agricultural issues is problematic.

## 7 Future Work

By coding the raw responses to the most important problem question asked by the NES, I will be able to determine what issues are important to individual voters. This is one of the primary elements required to empirically test my theory of the formation of party identification, which I present in my dissertation. I propose that before an individual chooses which political party to identify with, she must first determine whether the partisan environment addresses the issues that are most important to her. If neither of the

political parties consistently prioritizes and takes a position on the issues that the individual consistently factors into her vote choice, she will not identify with either political party. Only if she observes enough of an emphasis on issues that she cares about by the political parties will the individual choose to identify with the party that best represents both her priorities and her policy positions.

Both the algorithm that I propose for coding open-ended survey questions in this paper and the resultant coding of the most important problem question are significant contributions to the literature.<sup>6</sup> The algorithm itself can be used for the coding of other open-ended survey responses which researchers have previously been unable to make use of because of the resources required to extract analytically useful information from such responses. With respect to the coding of the most important problem question in particular, there are a number of substantively interesting questions that this could be used for. In my dissertation I am using this coding to determine whether or not the priority each political party places on the issues that are important to an individual predicts the party identification of the individual. This coding could also be used by scholars to investigate whether the allocation of congressional funds is correlated with the electorate's interest in certain issues. The ambiguity of the current ANES coding scheme makes these, and other similar questions, unanswerable with the responses to this question.

---

<sup>6</sup>There are a number of things that I need to do before this algorithm can truly be useful to other researchers. One of the most important is showing how the resulting coding of the MIP question improves on the existing coding performed by the NES. Another is developing one or more robustness checks to demonstrate that this algorithm produces results that are consistent with what one would expect to find using other coding techniques.

## References

- Baumgartner, Frank R. & Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. The University of Chicago Press.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara & Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. Oxford University Press.
- Campbell, Angus, Philip E. Converse, Warren E. Miller & Donald E. Stokes. 1960. *The American Voter*. Chicago: University of Chicago Press.
- Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. New Haven: Yale University Press.
- Gerner, Deborah J., Philip A. Schrodtt, Ronald A. Francisco & Judith L. Weddle. 1994. "Machine Coding of Event Data Using Regional and International Sources." *International Studies Quarterly* 38:91-119.
- Hopkins, Daniel & Gary King. 2008. "Extracting Systematic Social Science Meaning from Text."
- Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge & Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford University Press.
- Krosnick, Jon A., Arthur Lupia, Matthew DeBell & Sarrell Donakowski. 2008. Problems with ANES Questions Measuring Political Knowledge. Technical report American National Election Studies.
- Laver, Michael, Kenneth Benoit & John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2):311-331.

Purpura, Stephen & Dustin Hillard. 2006. Automated Classification of Congressional Legislation. Technical report John F. Kennedy School of Government.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin & Dragomir R. Radev. 2006. "An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th U.S. Senate." Working Paper - July 18th, 2006.

\*<http://www.people.fas.harvard.edu/~kquinn/papers/TopicsMethodDavis.pdf>

Schrodt, Philip A. 1998. "KEDS: Kansas Event Data System. Version 1.0."

Table 1: Top Issues Categories

| 1984                | 1986              | 1988                   | 1990                   |
|---------------------|-------------------|------------------------|------------------------|
| 1 Budget            | Employment        | Budget                 | Economy                |
| 2 Employment        | Economy           | Drugs                  | War/Peace              |
| 3 Economy           | Budget            | Employment             | Budget                 |
| 4 Nuclear           | Nuclear           | Environment            | Environment            |
| 5 War/Peace         | Drugs             | Housing                | Employment             |
| 6 Defense/Military  | War/Peace         | Economy                | Drugs                  |
| 7 Taxes             | Arms/Weapons      | Education              | Housing                |
| 8 Russia            | Defense/Military  | Crime (not drugs)      | Middle East            |
| 9 Hunger            | Education         | Defense/Military       | Education              |
| 1992                | 1996              | 1998                   | 2000                   |
| 1 Employment        | Crime (not drugs) | Education              | Education              |
| 2 Economy           | Welfare           | Employment             | Healthcare             |
| 3 Budget            | Budget            | Economy                | Social Security        |
| 4 Healthcare        | Education         | Healthcare             | Children (not schools) |
| 5 Education         | Employment        | Crime (not drugs)      | Employment             |
| 6 Housing           | Healthcare        | Children (not schools) | Economy                |
| 7 Environment       | Drugs             | Social Security        | Crime (not drugs)      |
| 8 Crime (not drugs) | Environment       | Poverty                | Defense/Military       |
| 9 Welfare           | Economy           | Morality               | Drugs                  |