The ‘Most Important Problem’ Aggregate Dataset (MIPD) Codebook

Release 1.0

Colton Heffington
Brandon Beomseob Park
Laron K. Williams
University of Missouri

January 2, 2017

Contents

List of Tables

1 Overview

This document discusses the procedures used to change the individual-level MIPD into a variety of aggregate-level datasets containing the weighted percentages for each category. This includes both the "MIPD Aggregate–Release 1.0.dta" and "MIPD Annual–Release 1.0.dta" datasets. In this document, we will discuss weighting procedures, variable descriptions, and provide some basic descriptive statistics. Those interested in the original data collection or question wording can examine the codebook for the individual-level dataset, "MIPD Codebook–Release 1.0.pdf".

2 Reference

2.1 Location

The MIPD dataset can be found in the following locations:

- Laron Williams' personal website: `http://faculty.missouri.edu/~williamslaro/mipdata.html`
- Laron Williams' Harvard Dataverse: `http://dataverse.harvard.edu/dataverse/laronwilliams`
- Replication materials (as well as the Release 1.0) are also available on the *Conflict Management and Peace Science* website.

2.2 Citation

Please use the following citation if you use or reference the MIPD:

Heffington, Colton, Brandon Beomseob Park and Laron K. Williams. Forthcoming. "The 'Most Important Problem' Dataset (MIPD): A New Dataset on American Issue Importance" *Conflict Management and Peace Science*

2.3 Acknowledgements

This project would not be possible without the helpful research assistance of Rachel Dicke, Ted Masthay, Andy Philips, Stella-Leonie Wancke and Murat Yildirim.

3 Introduction

The MIPD provides three methods of calculating measures of problem importance over time:

1. Use the "MIPD–Generate Aggregate Data Sets–Release 1.do" file to transfer the individual-level data ("MIPD–Release 1.0.dta") to an aggregate data set. This is a more flexible option, and the user can choose the temporal domain (i.e., annual, quarterly or monthly), the coding scheme used for the MIP responses (i.e., CAP, MARPOR and Singer), and specific variables for subgroup analysis (e.g., gender or partisan identification). The resulting data set will be either survey/time (containing the MIP percentages for that survey in that time period) or specific temporal domain (i.e., annual, quarterly or monthly). See the section below for more information about the available options.

2. Use the aggregate dataset that we provide (MIPD Aggregate–Release 1.0.dta). The MIP aggregate dataset has aggregated all of these responses into percentages for each code in the three coding schemes for each survey. If one is interested in seeing the aggregate percentages at each survey, then this dataset will be appropriate.

3. Use the annual dataset that we provide (MIPD Annual–Release 1.0.dta). The MIP aggregate dataset has aggregated all of these responses into a single observation for each year (see an in-depth description of these procedures below). If one is just interested in analyzing MIP across time, then this dataset will be appropriate.

The MIPD aggregate dataset is structured so that it is easy to examine MIP measures over time for a variety of coding schemes, including the Comparative Agendas Project (CAP, Baumgartner and Jones 2002). Nonetheless, it should also be noted that there are a few minor, but meaningful, differences in the calculation of the MIP measures herein and those in the CAP. It should be helpful to illustrate the following three differences:

1. While the MIPD includes up to three MIP responses (if available), the MIP measures are only calculated based on the first MIP. The CAP aggregates across all three MIP responses, which requires some normalization (because the sum of the percentages often exceeds 100%).

2. All problems that could not be categorized based on our coding scheme are placed into an "other" category, and all respondents who could not identify a problem, or refused, or stated "don't know" are placed into a "missing" category. In the CAP these categories are combined. The result is that the MIP measures calculated herein are based on the total number of respondents who identified a problem (i.e., excluding the "don't know/refused/missing" category), whereas the CAP measures are based on the total number of respondents in that survey. Because of that difference, the MIP measures in the MIPD will be slightly higher than those in the CAP (depending on the number of respondents in the "don't know" category).

   If you want to make our measures more comparable with those from CAP, select the "include" option, which will include the don't knows/missing responses in the denominator when calculating percentages.

3. The CAP averages all the polls in that year to create the annual measure, meaning that each poll has the same influence on the annual measure. The MIPD uses aggregation weights (see the description below).

3.1 Weighting Procedures

This aggregate dataset uses two different weighting procedures to ensure that the individual-level responses can be effectively used to generate aggregate-level estimates of the percentage of respondents identifying different categories as the "most important problem" facing the country today.

3.1.1 Population Weights

Beginning in the 1970s, polling agencies began supplying population weights in their surveys so that one can make inferences about the population overall. Table 1 shows the surveys and polling agencies that lacked population weights.

Table 1: Distribution of Missing Population Weights across Survey Houses

Survey House	# Surveys†	Missing Years
American National Election Study (ANES)	7	1966, 1972, 1978-1986
Decima Research	1	1989
National Opinion Research Center (NORC)	1	1944
Gallup	99	1939-1971

Note: †: number of surveys missing weights.

Even among those who provided population weights, there is some variation as to the "representative" value; most of the time the average respondent is weighted 1, but others are weighted 10, 100 (Gallup polls from 1993-1999) 1000 (*Los Angeles Times* and *New York Times* polls), or some other integer (Gallup had a representative value of 2 from 1962 to 1983; a *Washington Post* poll conducted by ICR Survey Research in April 1990 had a mean weight of 181,038). For example, early ANES studies either had population weights (e.g., 1960), no weights (e.g., 1966, 1972, 1978-1986), weights to remove duplicate respondents from panels (e.g., 1974 and 1976), or population weights to downweight oversamples of African Americans (e.g., 1964, 1968, 1970).

Population weights (if available) were used to calculate the survey-specific MIP percentages for each problem category (whether CAP, MARPOR or Singer), MIP$_p$ in the following manner:

$$\text{MIP}_p = \frac{\sum_{i=1}^{n}(w_i \times P_{pi})}{\sum_{i=1}^{n}(w_i \times M_i)} \tag{1}$$

where n is the number of non-missing responses, w_i is the population weight for observation i (for those surveys without population weights, $w_i = 1 \quad \forall \quad i$), P_i is a binary variable coded 1 if respondent i selected problem p as the MIP, and M_i is a binary variable coded 1 if respondent i provided a non-missing response to the MIP. The result is that each MIP$_p$ provides the percentage of respondents identifying that problem as the MIP (out of those who identified a problem).

3.1.2 Aggregation Weights

There is a significant amount of variability in the availability of suveys across time. Indeed, there might be 13 available surveys in 1960, but the preceding and following years may only have 4 each. Furthermore, even in the years where we have a large number of available surveys, the availability across quarters and months might vary. For example, consider Table 2, which shows the distribution of 29 surveys across months in 2007.

As we aggregate to a higher level, we should consider how many surveys come from each of the lower levels. Assume that we are calculating an annual measure of MIP responses by aggregating the data from the four quarters. A simple method would be to average the percentages from the 29 monthly surveys, but this would give greater weight to the months where we have a lot of surveys (such as May or September). A better method—and the one that we implement in the MIPD Aggregate Dataset—is to provide lesser weight to the surveys that come from more common periods and provide greater weight from the surveys that come from less common periods. This method provides equal weight to each of the lower time periods rather than the surveys themselves.

Consider the calculation of quarterly MIP responses for the first quarter of 2007 (Table 2). The simple method would be to calculate percentages for each individual survey and then merely take the mean of those six surveys. Our method weights each of the January surveys by 1/3, the February survey by 1, and the March surveys by 1/2. Essentially this gives equal weight to each month, rather than survey. The same process is used to generate the annual measures (giving weights to each quarter). Since the monthly measure already has a fine-grained temporal sequence, no aggregation weights are provided.

Table 2: Available Surveys in 2007 Across Months

	Available # of Surveys
January	3
February	1
March	2
April	3
May	4
June	1
July	3
August	3
September	4
October	1
November	2
December	2
Total	29

4 Stata Do File to Generate Aggregate Data Sets

The "MIPD–Generate Aggregate Data Sets–Release 1.do" file converts the individual-level data ("MIPD–Release 1.0.dta") into an aggregate data set. This method offers a great deal of flexibility, as there are a number of options to tailor the final dataset to suit one's research interests. We include the following options:

1. **Temporal Domain**: The do file converts the individual dataset (containing over 900,000 observations) over eight decades into a dataset of the user's choosing. More specifically, scholars can create annual, quarterly, or monthly datasets. Keep in mind that surveys earlier than 1980 are rather scarce, so temporal domains beyond annual might have considerable missing data. Or, scholars can select "survey" and the end result will be a dataset with the survey as the unit of analysis.

2. **Coding Schemes**: Scholars can select any combination of the three coding schemes for MIP responses, including CAP, MARPOR and Singer.

3. **Subgroup Percentages**: This option gives users the ability to select one variable to calculate subgroup percentages. For example, selecting the *male* variable for the subgroup analysis will produce separate percentages of each coding scheme requested for males and females. These variables will reflect, for example, the percentage of males selecting "economy".

   It should also be noted that the variable that is created may have a different value than the value it represents: for example, since the negative sign is not allowed in a variable name, the values for *partisan identification* start at 0. The variable labels, however, have the correct value labels.

In addition to *male*, the following variables are available for subgroup analysis (for more information see the MIPD Codebook–Release 1.0.pdf):

- *pid*: while the original partisan identification variable is coded on a seven-point scale, subgoup analyses based on variables with multiple values tend to produce wildly varying percentages based on the particular sample. This variable is coded 0 (Democrat), 1 (independent) and 2 (Republican).
- *ideology*: this variable is coded 1 (liberal), 2 (moderate/neither) and 3 (conservative).
- *income_quartile*: this variable is coded based on household income quartiles: 1 (lower 25th percentile) up to 4 (highest 25th percentile).
- *male*: this is coded 0 (female) and 1 (male).
- *white*: this is coded 0 (non-white) and 1 (white).
- *college*: this is coded 0 (no college degree) and 1 (college degree).
- *approve*: this is the response to the traditional presidential approval question and is coded 0 (disapprove) and 1 (approve).

4. **Denominator**: if the scholar wants the percentages to be calculated with the missing values and don't know responses included in the denominator (similar to the CAP method), then the scholar should select to include those values. If no option is selected, the default is to exlcude those values, so that the percentages reflect the percentage of respondents who identified a problem that selected a particular problem.

5. **Response Selection**: scholars can choose whether to include every single MIP question, regardless of the response options, or exclude those that are "short" (i.e., those where the response is coded into 8-10 broad categories), or closed-ended. If one is interested in using the Singer or CAP coding schemes (where there are multiple categories), we would discourage them from including either short or closed-ended responses. The default is to include all the surveys.

6. **Question Wording**: we have coded about 20 different variations on the "most important problem" question. Some of these may be more appropriate for some empirical analyses than others. Scholars can use the `mipid` variable to select which question wordings to include in the aggregate dataset (see the precise question wordings in the MIPD Codebook–Release 1.0.pdf document). We offer the option (called "MIP") of including any survey that refers to the "most important problem", which excludes the following:

    - 3: "What do you think is the most important problem facing this section of the country today?"
    - 17: "What do you think is the single most urgent problem facing the country today?"
    - 23: "What is the most important problem that you and your family face today?"

- 30: “What do you think is the main problem facing the country today?”
- 31: “What do you think will be the most important problem facing this country in the 21st Century?”

The other option is even more conservative, and only includes the traditional “most important problem” question. It excludes all of those wordings above, in addition to the following:

- 6: “What issue or problem would you say is most important for the next president to address?”
- 12: “What do you personally regard as the most important problem which should be discussed in the coming November election campaign?”
- 20: “What do you think is the most important problem facing Congress today?”
- 24: “Of all the problems facing the nation, which one do you think the president should give most attention to?”
- 99: any other question wordings.

The default is to include all the question wordings.

4.1 Aggregate Data Set: “MIPD Aggregate–Release 1.0”

The aggregate data set that accompanies this data release is generated with the “MIPD–Generate Aggregate Data Sets Release 1.do” with the following options: survey dataset, all three coding schemes, no subgroups, exclude don’t knows/missing, include closed-ended and short (“open+short”), and include all question wordings.

4.2 Variable Descriptions

This section describes some of the identifier variables that accompany the aggregate and annual data sets, and are produced by the do file.

- `studyid`: a string identifying each survey uniquely. These identification tags come from the polling agencies (e.g., Gallup, CBS) themselves and uniquely identify the files on the Roper Center’s iPoll service.

- `prev_ts`: indicates the date of the previous presidential election in Stata’s *dmy* format (e.g. 06nov2012).

- `next_ts`: indicates the date of the next presidential election in Stata’s *dmy* format (e.g. 08nov2016).

- `fw_start`: the date that the field work for the survey began in Stata’s *dmy* format (e.g. 04apr2013).

- `fw_end`: the date that the field work for the survey ended in Stata’s *dmy* format (e.g. 11apr2013).

- `surveyorganization`: string variable containing the survey organization, according to the Roper Center.

- `survey`: numeric codes for the various survey organizations (see Table 3).

- `sponsororganization`: string variable containing the sponsoring organization, according to the Roper Center. If there is no sponsoring organization, we list the survey organization here.

- `sponsor`: numeric codes for the various sponsoring organizations (see Table 4).

- `oversamplefull`: full description of the oversample (if requested) provided by the Roper Center.

- `oversample`: binary variable coded 1 if the survey contained an oversample. Population weights are available in all of the surveys containing an oversample.

- `samplesize`: continuous variable that provides the overall sample size (not effective size).

- `sample`: variable denoting the sample used in the survey. The vast majority of respondents were selected from a national sample of adults (98.5%), three surveys are of registered voters, and the remainder are from national adult samples with recalls or youth components. For aggregate analysis, scholars will want to exclude those surveys that are based on registered voters. Though some surveys are denoted *National Adult + Youth*, the youths (<18) are excluded from the dataset.

- `question`: this string variable provides the exact question wording of the MIP question (excluding preamble material).

- `mipid`: this numeric variable groups the questions based on their wording into two dozen categories (see the "MIPD Codebook–Release 1.0.pdf" for more details).

- `short`: these surveys are coded 1 if the "most important problem" question *did* allow open responses, but the responses were grouped into 8-12 broad categories. For example, any mention related to "Soviet Union, World Peace, War or Defense" was coded into one category in the CBS/*New York Times* April 1981 poll.

  - 68 surveys total were coded as *short*, predominantly Gallup, CBS News/*New York Times*, *Los Angeles Times*, Chilton Research Services, and Schulman, Ronca and Bucuvalas.

- `openended`: these surveys are coded 0 if the "most important problem" question *did not* allow open responses, but instead respondents were asked to identify which one out of 6-10 problems was the "most important". One should be cautious about using closed or "short" questions (see below) in either of the more fine-grained coding schemes (Singer and CAP) because the broad categories are not appropriate.

  - The following *Los Angeles Times* surveys were coded as *closed*: December 1985, July 1986, March 1987, July 1988, April 1989, July 1989, November 1989, December 1990, January 1991. One CBS/*New York Times* survey (January 1980) was coded as *closed*.

- `weightavail`: binary variable coded 1 if the population weight was available. If a weight is not available, then all of the respondents are treated as equally likely to be chosen from the population.

- `interviewmethod`: this variable provides the interview method used in the survey (see Table 5).

- **MIP percentages**:
  - The MIP percentage variables all following a similar naming convention based on the abbrevatiation of the coding scheme (“cap”, “marpor” or “singer”) followed by the numeric code.

    For example, the percentage of those mentioning the “economy” (value 1) in the Singer scheme is singer1”. The numeric codes can be found in the varible labels or in the individual-level codebook (“MIPD Codebook–Release 1.0.pdf”).
  - If subgroups are used, then these values are listed after an underscore. Keep in mind that the value of the subgroup may not match the value of the subgroup variable, so pay attention to the variable label. For example, the percentage of females (coded 0 in male”) selecting the economy in the Singer scheme is found in singer1_1”.

4.3 Annual Data Set: "MIPD Annual–Release 1.0"

The annual data set that accompanies this data release is generated with the "MIPD–Generate Aggregate Data Sets Release 1.do" with the following options: annual data set, all three coding schemes, no subgroups, exclude don't knows/missing, exclude closed-ended and short ("open+short"), and only include the traditional "most important problem" question wording ("MIP").

4.4 Variable Descriptions

This section describes some of the identifier variables that accompany the annual data set (and other, non-"survey" temporal domains).

- `tsy`: a time domain variable (in Stata format). This is the year-quarter in the quarterly data set (`tsq`) and the year-month in the monthly data set (`tsm`). These variables uniquely identify every single observation in the data. Keep in mind, however, that if there were no surveys available in that time period then the MIP percentages will all be missing.

- `numsurveys_y`: this variable counts the number of surveys that are available in that year (or quarter for `numsurveys_q` and month for `numsurveys_m`). If this variable equals 0, then there are no surveys available during that time period and all of the MIP percentages variables will be missing.

- `nummiss_y`: this variable counts the number of months that *do not* have available surveys in that year (or quarter for `nummiss_y`). If this variable equals 12 (or 4 for the quarterly data), then there are no surveys available during that time period and all of the MIP percentages variables will be missing. Since there is no aggregation weight for the monthly dataset, this variable is not available.

- **MIP percentages**:
  - The MIP percentage variables all following a similar naming convention based on the abbrevatiation of the coding scheme ("cap", "marpor" or "singer") followed by the numeric code. For example, the percentage of those mentioning the "economy" (value 1) in the Singer scheme is `singer1`.
  - If subgroups are used, then these values are listed after an underscore. Keep in mind that the value of the subgroup may not match the value of the subgroup variable, so pay attention to the variable label. For example, the percentage of females (coded 0 in `male`) selecting the economy in the Singer scheme is found in `singer1_1`.
  - All these variables include the suffix `_perc`.

Descriptive Statistics

Table 3: Distribution of Available Data across Survey Houses

Value		Survey House	# Surveys	N	MIP $N^{\dagger}$
1	=	ABC News	5	5,306	5,121
2	=	ABC/*Washington Post*	10	13,437	12,863
3	=	ANES	20	37,072	29,214
4	=	Associated Press	1	1,499	970
5	=	CBS News	73	78,857	72,985
6	=	CBS News/*New York Times*	118	148,512	133,241
7	=	CCFR	1	1,507	1,461
8	=	Chilton Research Services	23	35,564	33,290
9	=	Decima Research	1	1,000	977
10	=	Harris Interactive	1	1,003	921
12	=	International Communications Research	3	4,977	4,656
13	=	Ipsos	5	5,036	2,429
14	=	*Los Angeles Times*	36	63,068	59,027
16	=	NORC	1	2,564	2,344
17	=	*New York Times*	4	6,071	5,606
18	=	Princeton Survey Research Associates	49	69,318	42,275
20	=	Schulman, Ronca & Bucuvalas	2	2,011	1,902
21	=	Social Science Research Solutions	6	6,168	5,971
22	=	Stony Brook University	6	4,992	4,683
23	=	The Gallup Organization	313	438,989	409,041
25	=	*The Washington Post*	4	3,662	3,544
26	=	Yankelovich Partners, Inc.	4	4,559	4,201

Note: †: number of non-missing MIP observations.

Table 4: Distribution of Available Data across Sponsoring Organization

Value		Survey House	# Surveys	N	MIP $N^\dagger$
1	=	60 Minutes/Vanity Fair	5	5,267	5,108
2	=	ABC News	7	7,696	7,449
3	=	ABC News/Nightline	1	537	516
4	=	ABC News/*Washington Post*	30	46,074	43,309
5	=	ANES	20	37,072	29,214
6	=	Associated Press	6	6,535	3,399
7	=	Bloomberg News	1	1,321	1,309
8	=	CBS News	72	77,722	71,879
9	=	CBS News/*New York Times*	120	150,548	135,210
10	=	CNN/Knight Ridder	1	1,387	1,344
11	=	CNN/*USA Today*	18	18,926	17,775
12	=	Kaiser Family Foundation	12	14,793	7,144
13	=	Kaiser/Agency	1	1,216	543
14	=	Kaiser/Harvard	3	3,695	2,711
15	=	Kaiser/Harvard/*Washington Post*	2	3,343	2,646
16	=	Kaiser/NPR/Harvard	1	1,557	1,366
17	=	*Los Angeles Times*	35	61,747	57,718
19	=	McLean's	1	1,000	977
21	=	NCSC	1	1,502	699
22	=	NORC	1	2,564	2,344
23	=	*New York Times*	4	6,071	5,606
24	=	Newsweek	2	3,715	3,680
25	=	Pew	27	40,961	26,502
27	=	Stony Brook University	6	4,992	4,683
28	=	The Gallup Organization	289	409,161	380,603
29	=	Time Magazine	2	2,011	1,902
30	=	Time Magazine/CNN	5	5,562	5,122
31	=	Times Mirror Center	5	10,451	8,438
32	=	Times Mirror/Harvard	1	1,021	979
33	=	*USA Today*	3	3,063	3,003
34	=	*The Washington Post*	4	3,662	3,544

Note: †: number of non-missing MIP observations.

Table 5: Interview Methods

Value		Method	# Surveys	N	MIP $N^\dagger$
1	=	Telephone: landline	356	456,422	398,218
2	=	Telephone: landline + cell	126	138,737	129,960
3	=	Face-to-Face	197	326,313	298,157
4	=	Face-to-Face + Telephone	7	13,700	10,387

Note: †: number of non-missing MIP observations.