Using http://sda.berkeley.edu for data analysis

Note: These tips refer to the "classic interface", which you access from the upper left corner of the screen. The current interface is the one we used in class, which contains the same information but all on the same screen. Pick the one that works best for you. If you are having trouble, try the classic interface, which is simpler.

This document provides an overview of instructions for using SDA to produce tabulations and cross-tabulations of survey items in the 2004 National Election Studies. If at all possible, to make things easier, use the 2004 NES dataset for your group projects.

The first section presents general tips and advice.

The second describes how to produce a tabulation and cross-tabulation.

The third describes how to recode variables (survey items) to produce cleaner results. --- *This last section is included for your information, but is not necessary to complete the homework. It discusses examples from the 2004 election, but could be generalized to our work with the 2008.*

The fourth describes how to produce a derived variable, such as an index of ethnocentrism or moral traditionalism.

I. General tips

1. Before analyzing data, always open up the codebook in a new browser window.
2. Remember to follow convention: **"dependent"** variables go on the row, "**independent"** variables go on the column.
3. Always calculate percentages down the columns, on the independent variables.
4. When you run your cross-tabulation, always check the box marked "question wording" so that you can verify that you correctly chose your intended survey item.
5. The responses to open-ended questions ("Is there anything in particular you like about George W. Bush?") are grouped into 'master codes' found in the codebook appendices. Until you are more familiar with the NES, ignore these questions.
6. SDA includes *all* the variables included within the survey. Some may not be readily interpretable until you look at the question wording. And most, if not all, require some sort of recoding to make them meaningful.

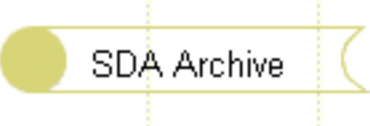
7. For the survey weights, leave the selected variable alone. We'll discuss what the weights mean later in class.
8. Be careful when using questions with a "Branching" format, in which respondents answer a question, such as "How angry did John Kerry make you feel?", only after they answered "Yes" to a question asking them "Did John Kerry make you feel angry?". If you using the question "How angry did John Kerry make you feel?", you will only be using responses from those persons who answered "Yes" to the first question; think about how you would want to handle responses to persons who said "No."

II. Producing a Tabulation

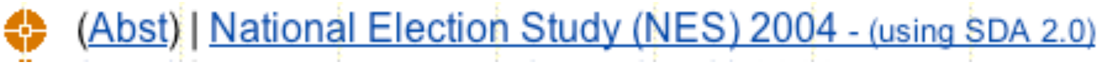
In this example, we will produce a frequency tabulation of responses to a survey item asking whether the phrase "intelligent" describes John Kerry.

GO TO http://sda.berkeley.edu

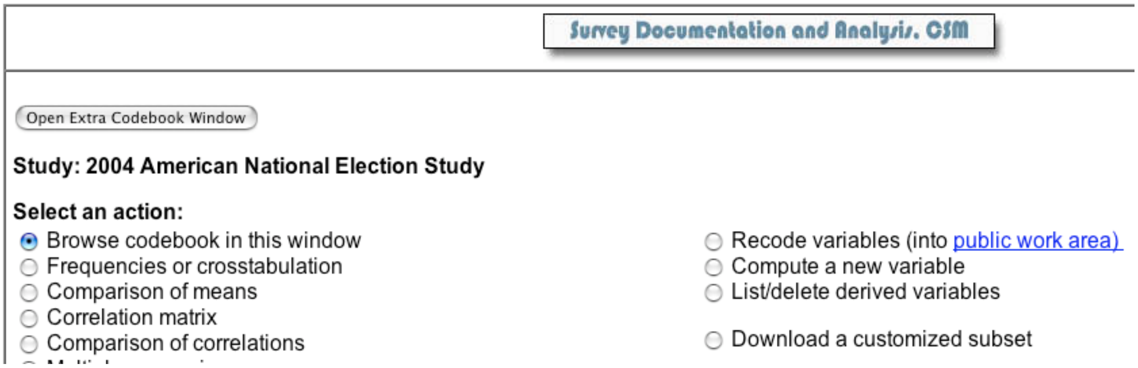
AT THE TOP, CLICK ON "SDA ARCHIVE":



THEN CLICK ON THE NES 2004 LINK:


(Abst) | National Election Study (NES) 2004 - (using SDA 2.0)

BEFORE YOU DO ANYTHING ELSE, CLICK ON 'Open Extra Codebook Window".


Survey Documentation and Analysis. CSM
Open Extra Codebook Window

Study: 2004 American National Election Study

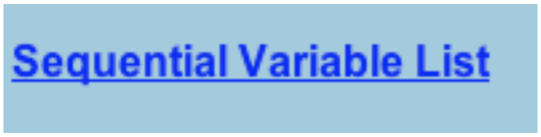
Select an action:

● Browse codebook in this window
○ Frequencies or crosstabulation
○ Comparison of means
○ Correlation matrix
○ Comparison of correlations

○ Recode variables (into public work area)
○ Compute a new variable
○ List/delete derived variables

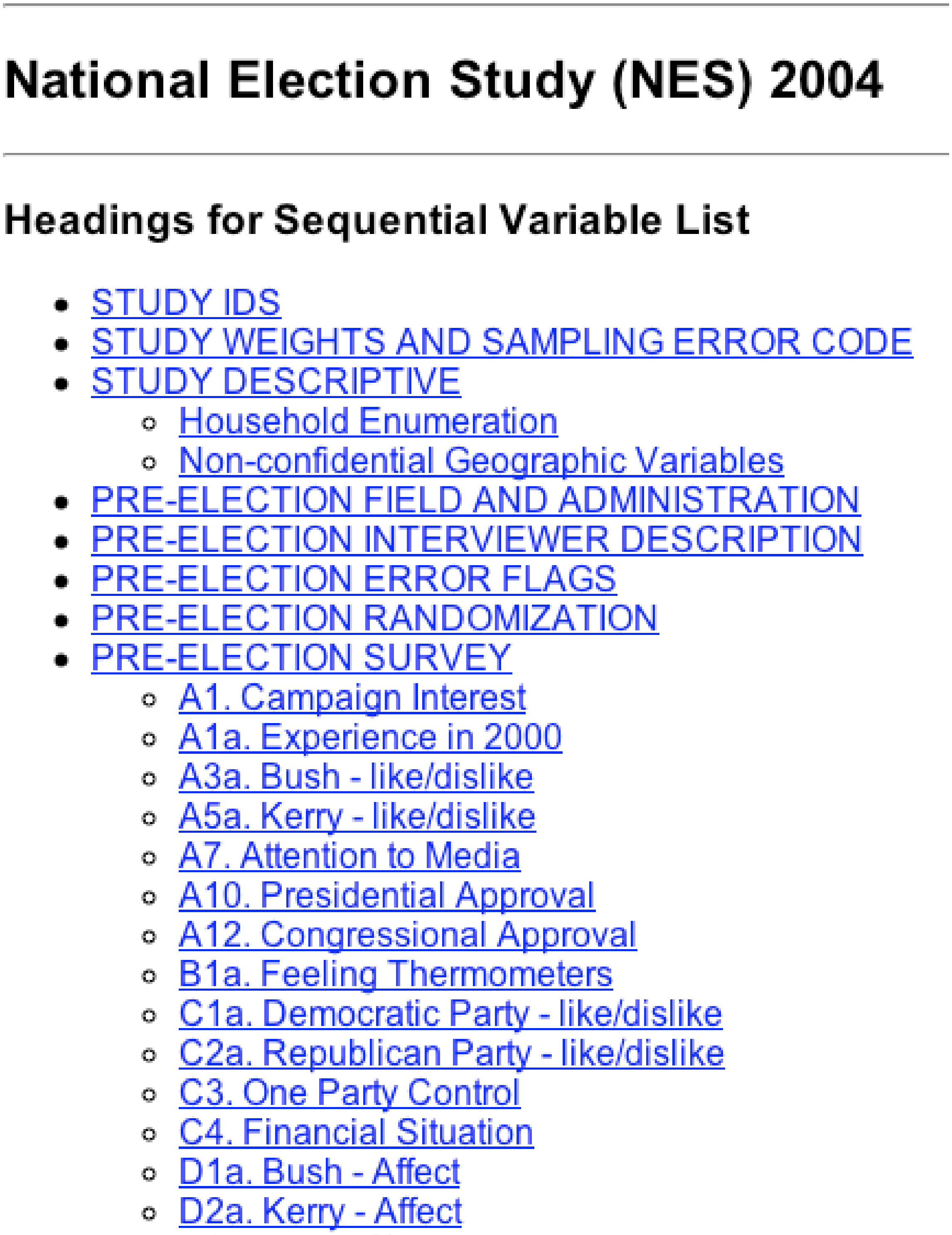
○ Download a customized subset

Once this window is open, you will see the following link on the left side of the screen:



Sequential Variable List

When you click on it, you will see the following on the right hand side:



National Election Study (NES) 2004

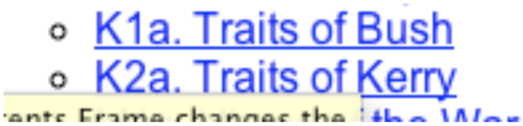
Headings for Sequential Variable List

- STUDY IDS
- STUDY WEIGHTS AND SAMPLING ERROR CODE
- STUDY DESCRIPTIVE
  - Household Enumeration
  - Non-confidential Geographic Variables
- PRE-ELECTION FIELD AND ADMINISTRATION
- PRE-ELECTION INTERVIEWER DESCRIPTION
- PRE-ELECTION ERROR FLAGS
- PRE-ELECTION RANDOMIZATION
- PRE-ELECTION SURVEY
  - A1. Campaign Interest
  - A1a. Experience in 2000
  - A3a. Bush - like/dislike
  - A5a. Kerry - like/dislike
  - A7. Attention to Media
  - A10. Presidential Approval
  - A12. Congressional Approval
  - B1a. Feeling Thermometers
  - C1a. Democratic Party - like/dislike
  - C2a. Republican Party - like/dislike
  - C3. One Party Control
  - C4. Financial Situation
  - D1a. Bush - Affect
  - D2a. Kerry - Affect

The above list shows a thematic breakdown of all the items that are included in the 2004 NES survey. These may be helpful for you to figure out which survey items you would like to analyze.

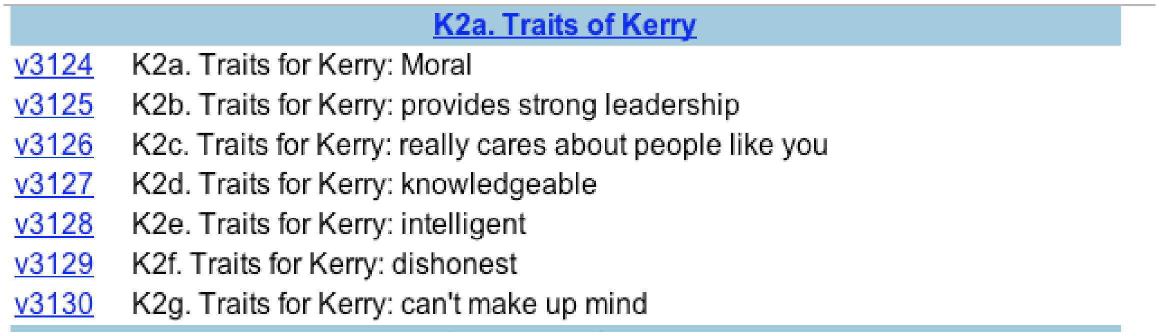
In this example, we are going to use the items from section K2a. "Traits of Kerry".

CLICK ON THE TRAITS OF KERRY LINK:



- K1a. Traits of Bush
- K2a. Traits of Kerry

And then you will see the following:



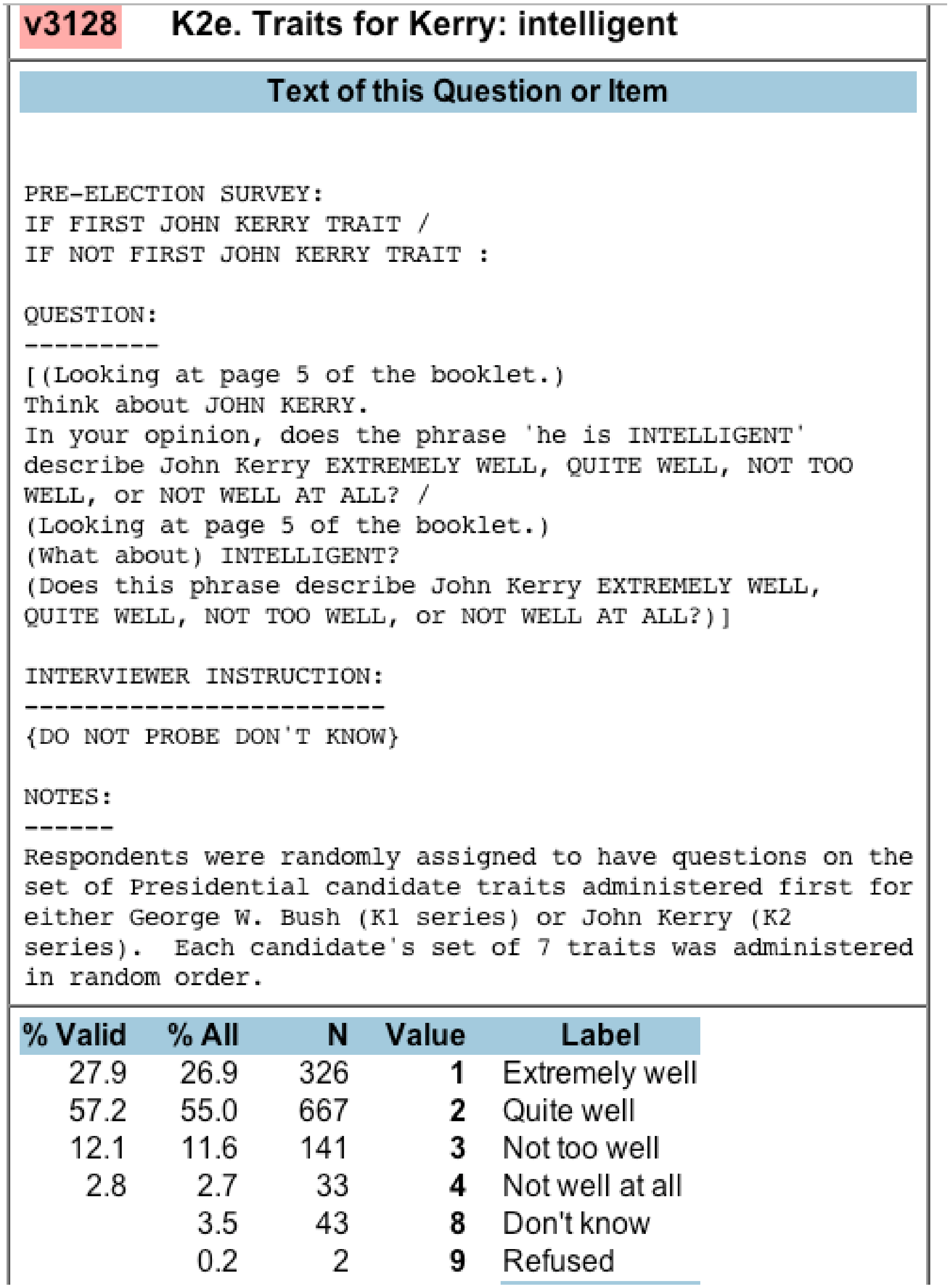
K2a. Traits of Kerry

v3124	K2a. Traits for Kerry: Moral
v3125	K2b. Traits for Kerry: provides strong leadership
v3126	K2c. Traits for Kerry: really cares about people like you
v3127	K2d. Traits for Kerry: knowledgeable
v3128	K2e. Traits for Kerry: intelligent
v3129	K2f. Traits for Kerry: dishonest
v3130	K2g. Traits for Kerry: can't make up mind

We are going to produce a tabulation of item v3128.

CLICK ON THE VARIABLE NAME, "V3128".

You will see the following:



v3128 K2e. Traits for Kerry: intelligent

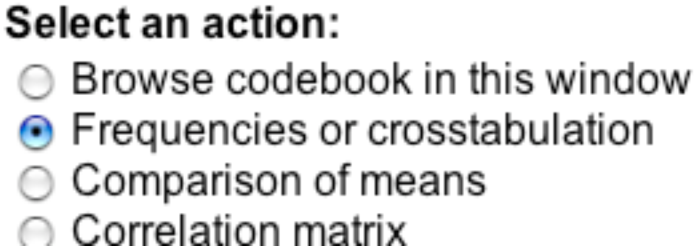
Text of this Question or Item

```
PRE-ELECTION SURVEY:
IF FIRST JOHN KERRY TRAIT /
IF NOT FIRST JOHN KERRY TRAIT :

QUESTION:
---------
[(Looking at page 5 of the booklet.)
Think about JOHN KERRY.
In your opinion, does the phrase 'he is INTELLIGENT'
describe John Kerry EXTREMELY WELL, QUITE WELL, NOT TOO
WELL, or NOT WELL AT ALL? /
(Looking at page 5 of the booklet.)
(What about) INTELLIGENT?
(Does this phrase describe John Kerry EXTREMELY WELL,
QUITE WELL, NOT TOO WELL, or NOT WELL AT ALL?)]

INTERVIEWER INSTRUCTION:
------------------------
{DO NOT PROBE DON'T KNOW}

NOTES:
------
Respondents were randomly assigned to have questions on the
set of Presidential candidate traits administered first for
either George W. Bush (K1 series) or John Kerry (K2
series).  Each candidate's set of 7 traits was administered
in random order.
```

% Valid	% All	N	Value	Label
27.9	26.9	326	**1**	Extremely well
57.2	55.0	667	**2**	Quite well
12.1	11.6	141	**3**	Not too well
2.8	2.7	33	**4**	Not well at all
	3.5	43	**8**	Don't know
	0.2	2	**9**	Refused

What we want to do is to produce a tabulation of the responses to the question showing a percentage breakdown exactly as it shows above. – But with one exception. When you produce a cross-tabulation, SDA automatically excludes the responses for "Don't Know" or "Refused". The "Value" for each of these responses are the numeric placeholders stored in the dataset. So, if you were to visually see the dataset, you would observe an "8" recorded for all those persons who said they didn't know the answer to this question (43 people total). Remember that these numbers are meaningful only when accompanied by the label that tells us what they mean.

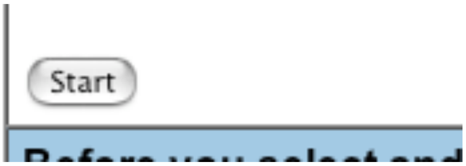
GO BACK TO THE START PAGE FOR THE NES 2004 STUDY. THIS TIME, CLICK ON "Frequencies or crosstabulation". THEN CLICK "Start".



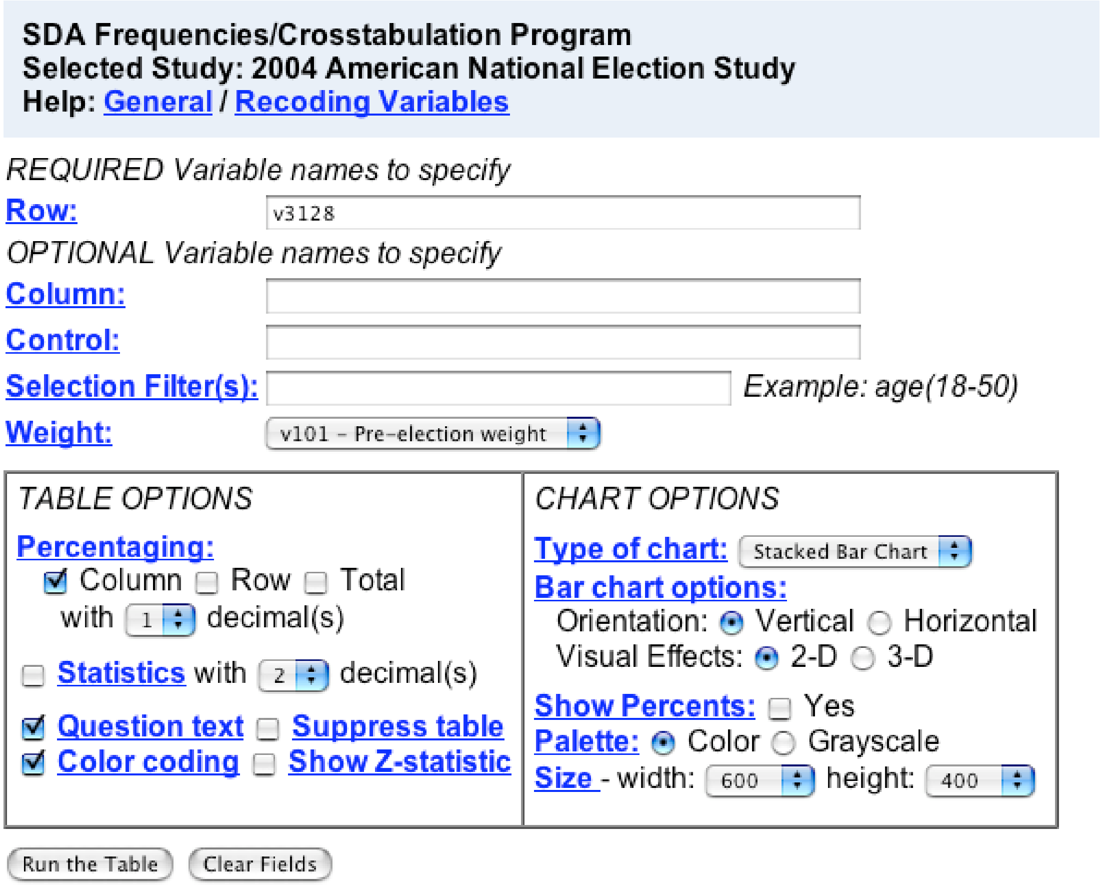
Select an action:

○ Browse codebook in this window
● Frequencies or crosstabulation
○ Comparison of means
○ Correlation matrix

…



After you click start, you will see the following screen:



SDA Frequencies/Crosstabulation Program
Selected Study: 2004 American National Election Study
Help: General / Recoding Variables

REQUIRED Variable names to specify

Row: v3128

OPTIONAL Variable names to specify

Column:

Control:

Selection Filter(s): *Example: age(18-50)*

Weight: v101 - Pre-election weight

TABLE OPTIONS

Percentaging:
☑ Column □ Row □ Total
with 1 decimal(s)

□ Statistics with 2 decimal(s)

☑ Question text □ Suppress table
☑ Color coding □ Show Z-statistic

CHART OPTIONS

Type of chart: Stacked Bar Chart

Bar chart options:
Orientation: ● Vertical ○ Horizontal
Visual Effects: ● 2-D ○ 3-D

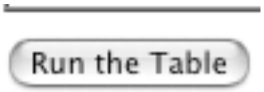
Show Percents: □ Yes

Palette: ● Color ○ Grayscale

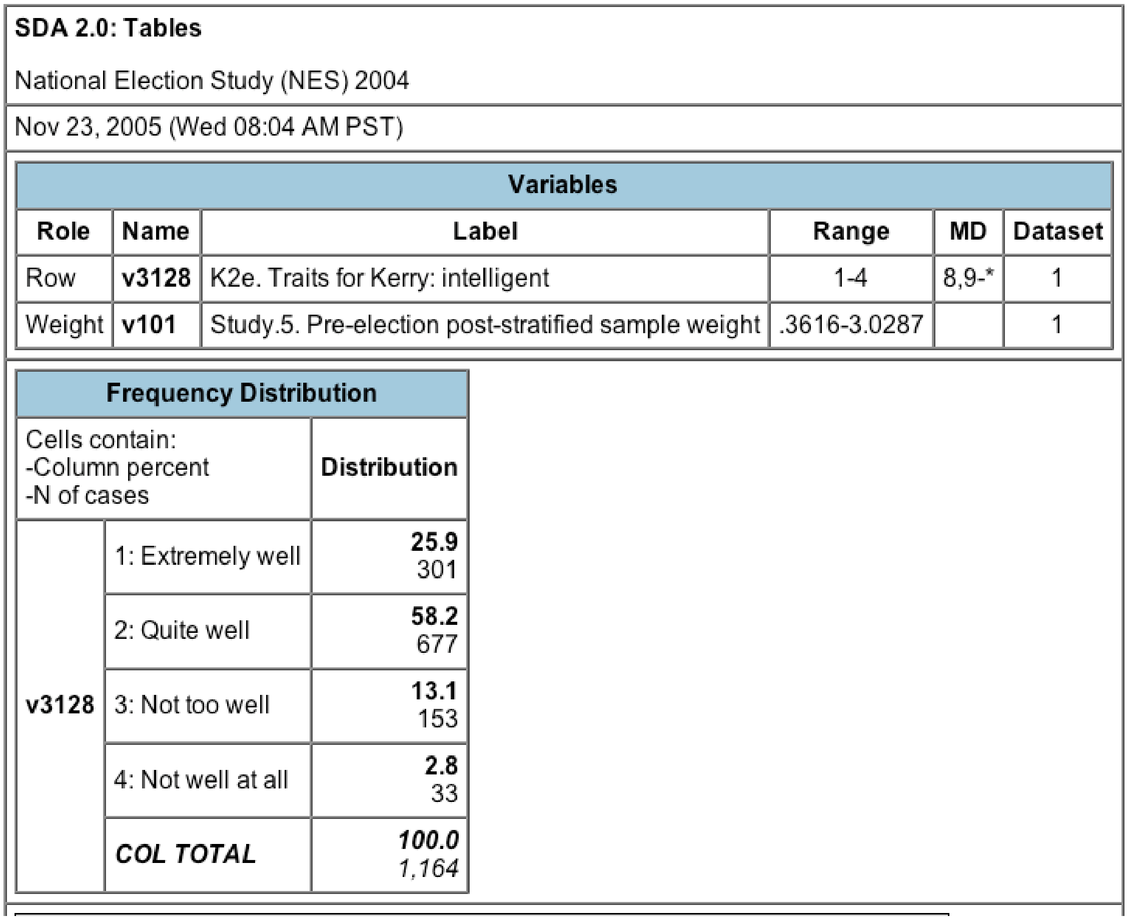
Size - width: 600 height: 400

ENTER "v3128" AS THE ROW VARIABLE. CLICK 'Question Text" so you can see the question wording underneath the tabulation.

CLICK



And then at the next screen you will see the following:



SDA 2.0: Tables

National Election Study (NES) 2004

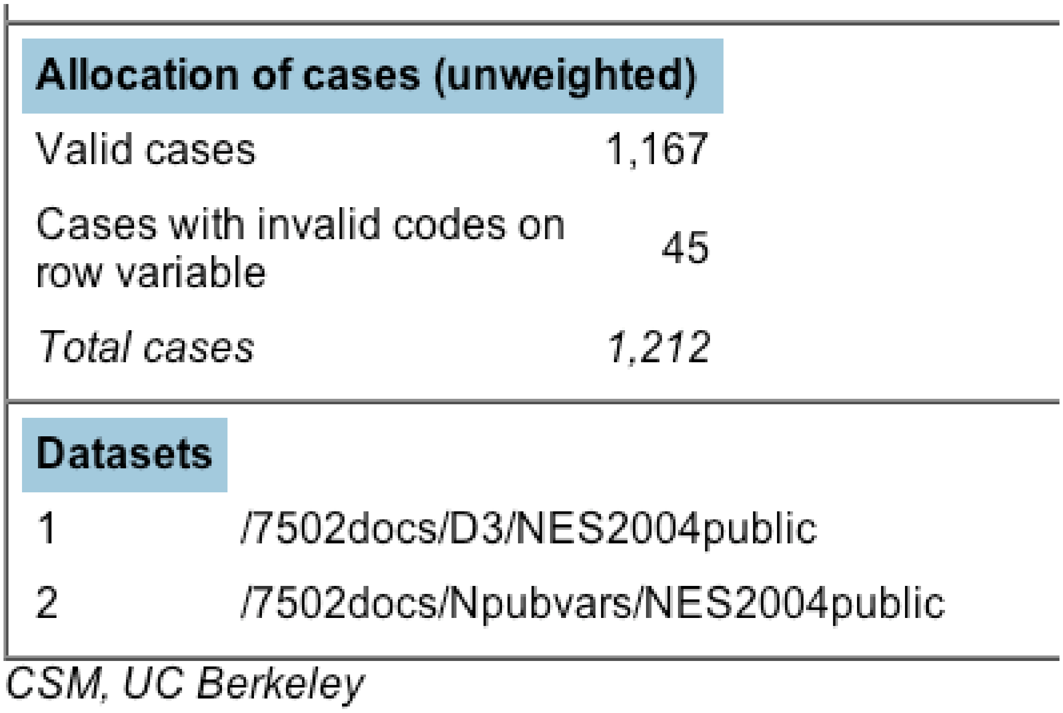
Nov 23, 2005 (Wed 08:04 AM PST)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**v3128**	K2e. Traits for Kerry: intelligent	1-4	8,9-*	1
Weight	**v101**	Study.5. Pre-election post-stratified sample weight	.3616-3.0287		1

Frequency Distribution		
Cells contain: -Column percent -N of cases		**Distribution**
v3128	1: Extremely well	**25.9** 301
	2: Quite well	**58.2** 677
	3: Not too well	**13.1** 153
	4: Not well at all	**2.8** 33
	COL TOTAL	***100.0*** *1,164*

These results should look familiar. The percentages are in boldface, while the frequencies are in regular typeface. Play around with the options on your own. Try reducing the number of decimals in the percentages.

Note the section at the bottom of the screen:



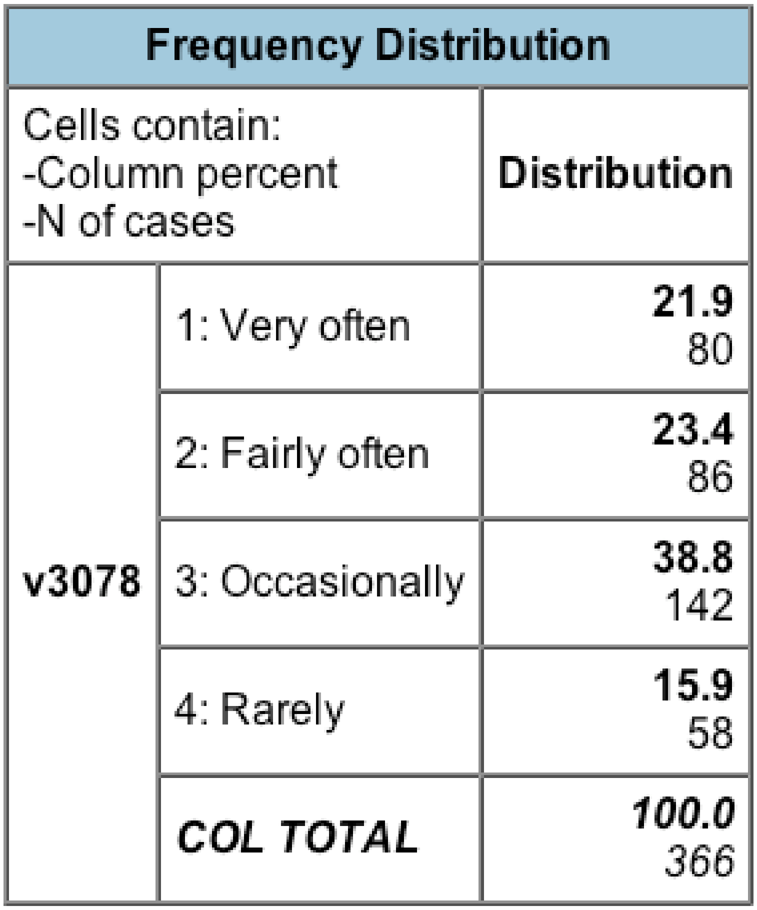
Allocation of cases (unweighted)	
Valid cases	1,167
Cases with invalid codes on row variable	45
Total cases	*1,212*

Datasets	
1	/7502docs/D3/NES2004public
2	/7502docs/Npubvars/NES2004public

CSM, UC Berkeley

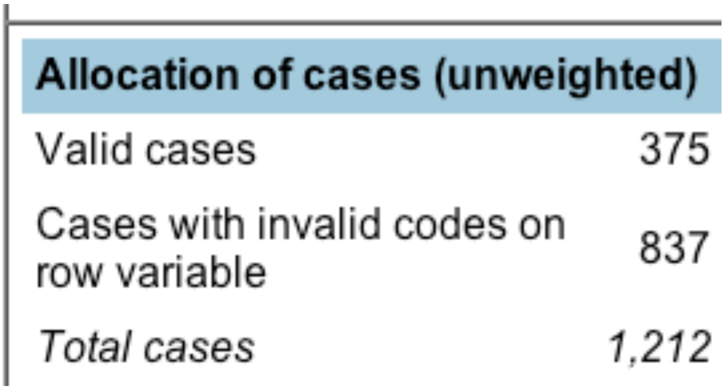
The "Valid cases" are the number of survey respondents, out of a total of 1,212 surveyed, who answered the question. 45 people, however, refused to answer it.

Now on your own, find the variable asking respondents how angry they felt about John Kerry And produce a tabulation of the reponses, which should look like the following:



Frequency Distribution		
Cells contain: -Column percent -N of cases		**Distribution**
v3078	1: Very often	**21.9** 80
	2: Fairly often	**23.4** 86
	3: Occasionally	**38.8** 142
	4: Rarely	**15.9** 58
	COL TOTAL	***100.0*** *366*

Notice that there are only 366 responses included in this item. The others, which were assigned missing data codes for this survey item, were not included in the tabulation. As the lower part of the table tells you, 837 people in the survey either said they did not feel "angry" at Kerry or they did not answer the question in the first place.



Allocation of cases (unweighted)	
Valid cases	375
Cases with invalid codes on row variable	837
Total cases	*1,212*

Notice how in the above table, I changed the number of decimals for the percentages to "0". You should do this as well, since decimal differences aren't really all that meaningful for our analyses.

Also, you should uncheck the "color coding" box. It gives you a more readable output. The color coding is intended to give you a sense of whether, according to inferential statistical theory, the differences you observe across the table are probably not just due to sampling error. Such theory is way beyond the scope of this course, so feel free to ignore it.

III. Producing a Cross-Tabulation

In this example, we are going to produce a cross-tabulation, comparing survey respondent's perception of the extent to which the phrase "intelligent" describes Kerry as a function of party identification. Or in other words, we want to test the intuition that perceptions of candidate character are so highly partisan, that there is little influence – if any – from the actual candidate's own personality. So, our working hypothesis would go something like this "Democratic party identifiers, rather than Republicans or independents, are much more likely to perceive Kerry as at least somewhat "intelligent", rather than "not at all" intelligent…"

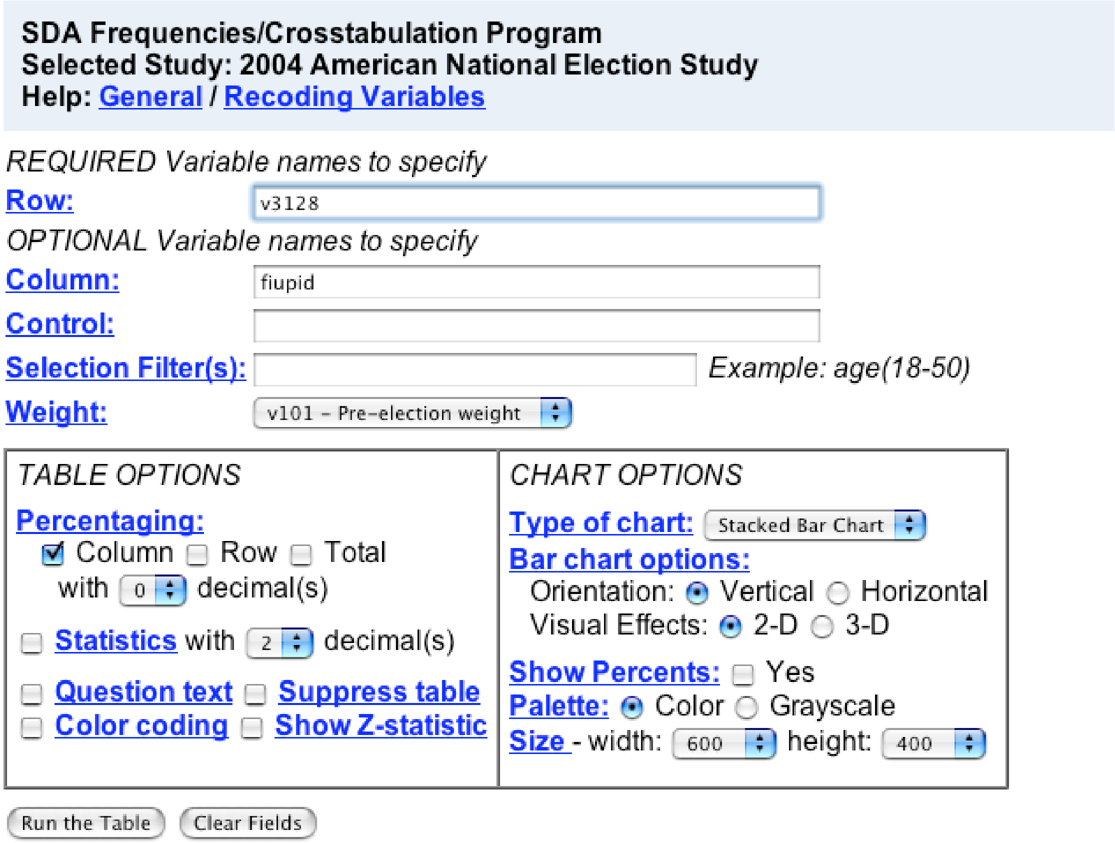
So the hypothesis implies that party identification is our independent variable, while Kerry trait perception is the dependent variable.

Imagine that we have created a new party identification variable called “fiupid”. In our class, we did not create this variable. To follow along with the example, use another variable for the actual exercise, if desired. (**All of the derived, or recoded, variables are stored in a public workspace. Anyone can use any of these variables.**

So, for the cross-tabulation, we will produce a cross-tabulated result for this Kerry trait by Party identification. Note that we are not producing a set of control tables here. All we are trying to do right now is go over a simple example to show the steps to produce a cross-tabulation.

GO BACK TO THE START SCREEN AND CLICK AGAIN ON THE FREQUENCIES/CROSS TABULATION LINK.

THIS TIME, ENTER "v3128" AS THE ROW VARIABLE AND "fiupid" as the column variable. **YES, EVERYTHING ON SDA IS CASE-SENSITIVE.**



SDA Frequencies/Crosstabulation Program
Selected Study: 2004 American National Election Study
Help: General / Recoding Variables

REQUIRED Variable names to specify

Row: v3128

OPTIONAL Variable names to specify

Column: fiupid

Control:

Selection Filter(s): *Example: age(18-50)*

Weight: v101 - Pre-election weight

TABLE OPTIONS

Percentaging:
☑ Column □ Row □ Total
with 0 decimal(s)

□ Statistics with 2 decimal(s)

□ Question text □ Suppress table
□ Color coding □ Show Z-statistic

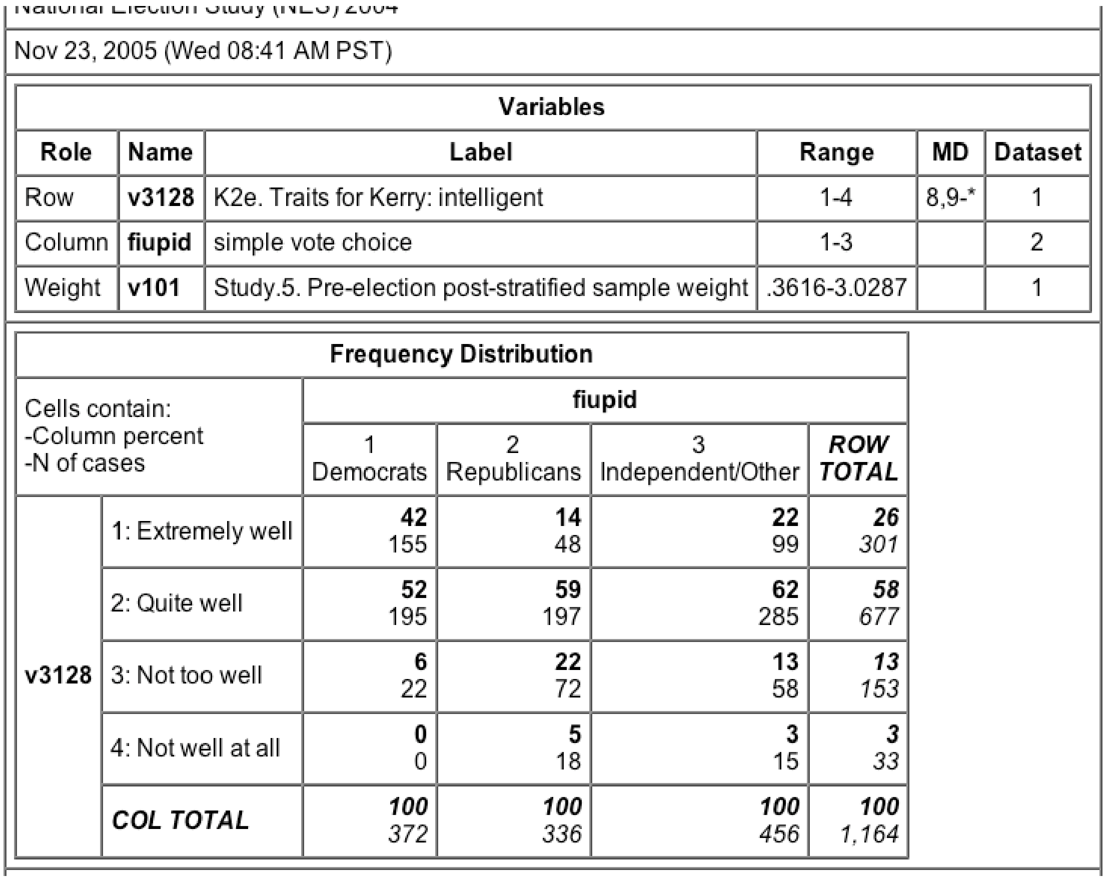
CHART OPTIONS

Type of chart: Stacked Bar Chart

Bar chart options:
Orientation: ● Vertical ○ Horizontal
Visual Effects: ● 2-D ○ 3-D

Show Percents: □ Yes
Palette: ● Color ○ Grayscale
Size - width: 600 height: 400

After you click "Run the Table", you should see the following results:



National Election Study (NES) 2004

Nov 23, 2005 (Wed 08:41 AM PST)

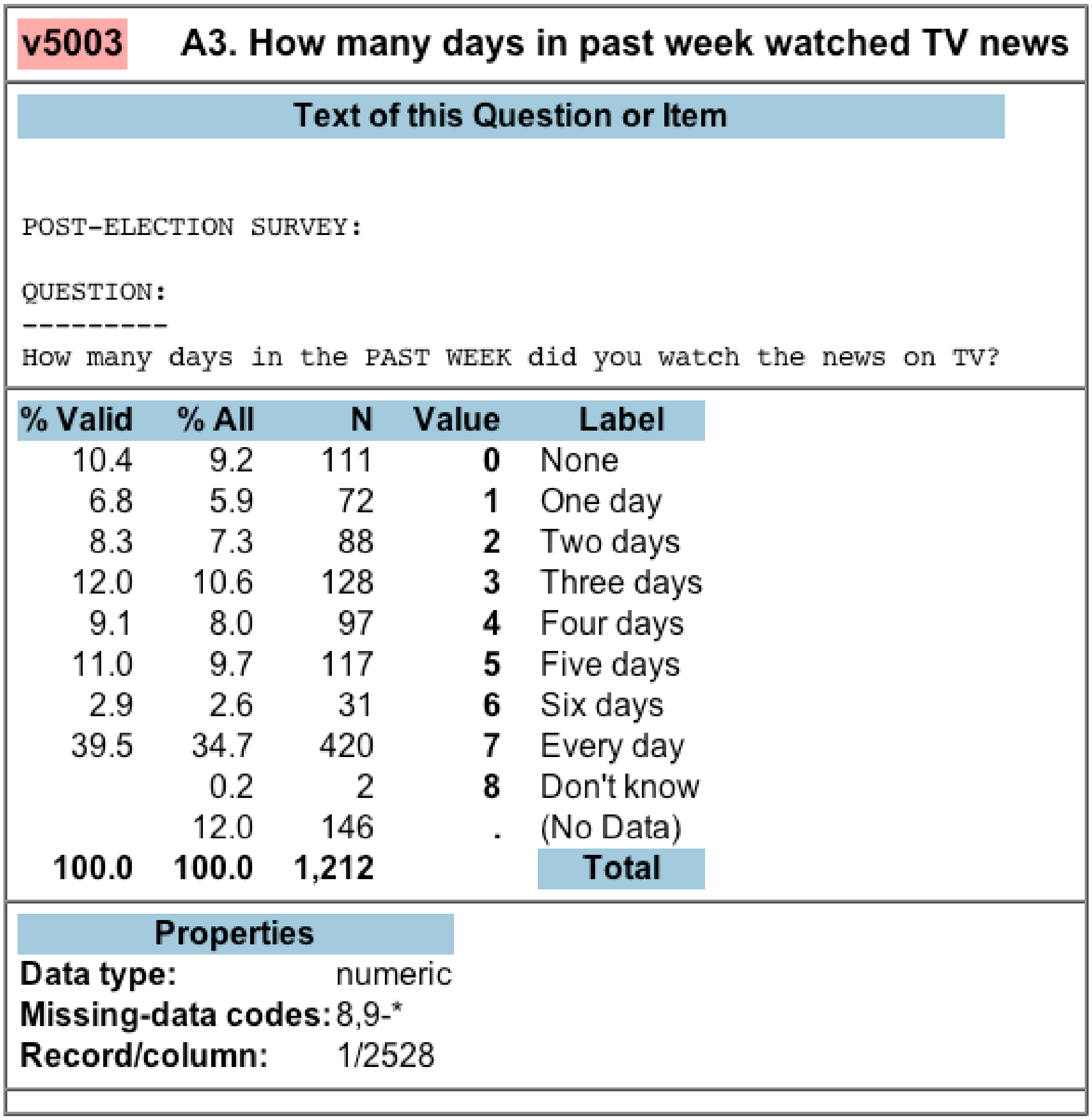
Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**v3128**	K2e. Traits for Kerry: intelligent	1-4	8,9-*	1
Column	**fiupid**	simple vote choice	1-3		2
Weight	**v101**	Study.5. Pre-election post-stratified sample weight	.3616-3.0287		1

Frequency Distribution

Cells contain: -Column percent -N of cases		**fiupid**			
		1 Democrats	2 Republicans	3 Independent/Other	***ROW TOTAL***
v3128	1: Extremely well	**42** 155	**14** 48	**22** 99	***26*** *301*
	2: Quite well	**52** 195	**59** 197	**62** 285	***58*** *677*
	3: Not too well	**6** 22	**22** 72	**13** 58	***13*** *153*
	4: Not well at all	**0** 0	**5** 18	**3** 15	***3*** *33*
	COL TOTAL	***100*** *372*	***100*** *336*	***100*** *456*	***100*** *1,164*

IV. Recoding Variables

onsider the question below about television news viewing. The question asks respondents about the number of days they watched television news in the past week. If we wanted to use this question in a cross-tabulation, it is coded into too many categories.



v5003 **A3. How many days in past week watched TV news**

Text of this Question or Item

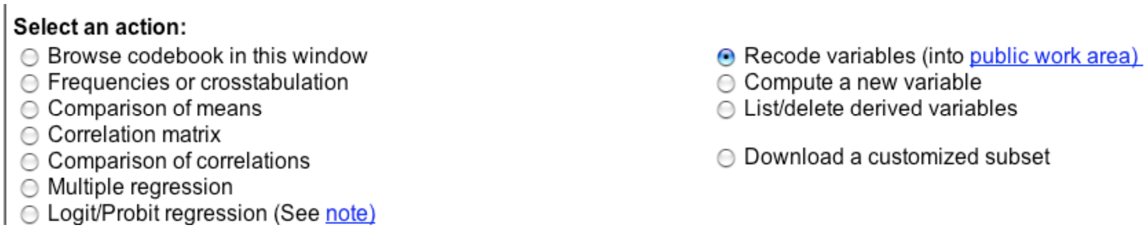
```
POST-ELECTION SURVEY:

QUESTION:
---------
How many days in the PAST WEEK did you watch the news on TV?
```

% Valid	% All	N	Value	Label
10.4	9.2	111	**0**	None
6.8	5.9	72	**1**	One day
8.3	7.3	88	**2**	Two days
12.0	10.6	128	**3**	Three days
9.1	8.0	97	**4**	Four days
11.0	9.7	117	**5**	Five days
2.9	2.6	31	**6**	Six days
39.5	34.7	420	**7**	Every day
	0.2	2	**8**	Don't know
	12.0	146	**.**	(No Data)
100.0	**100.0**	**1,212**		**Total**

Properties	
Data type:	numeric
Missing-data codes:	8,9-*
Record/column:	1/2528

We could use it after re-coding the number of days into a set of manageable categories, such as "Low", "medium", and "high". While arbitrary, perhaps we could agree that 0-2 days is "low", 3-4 is "medium", while 5-7 is "high". So, we need to recode it.

Go back to the start screen for the 2004 study, and click "recode variables"



Select an action:

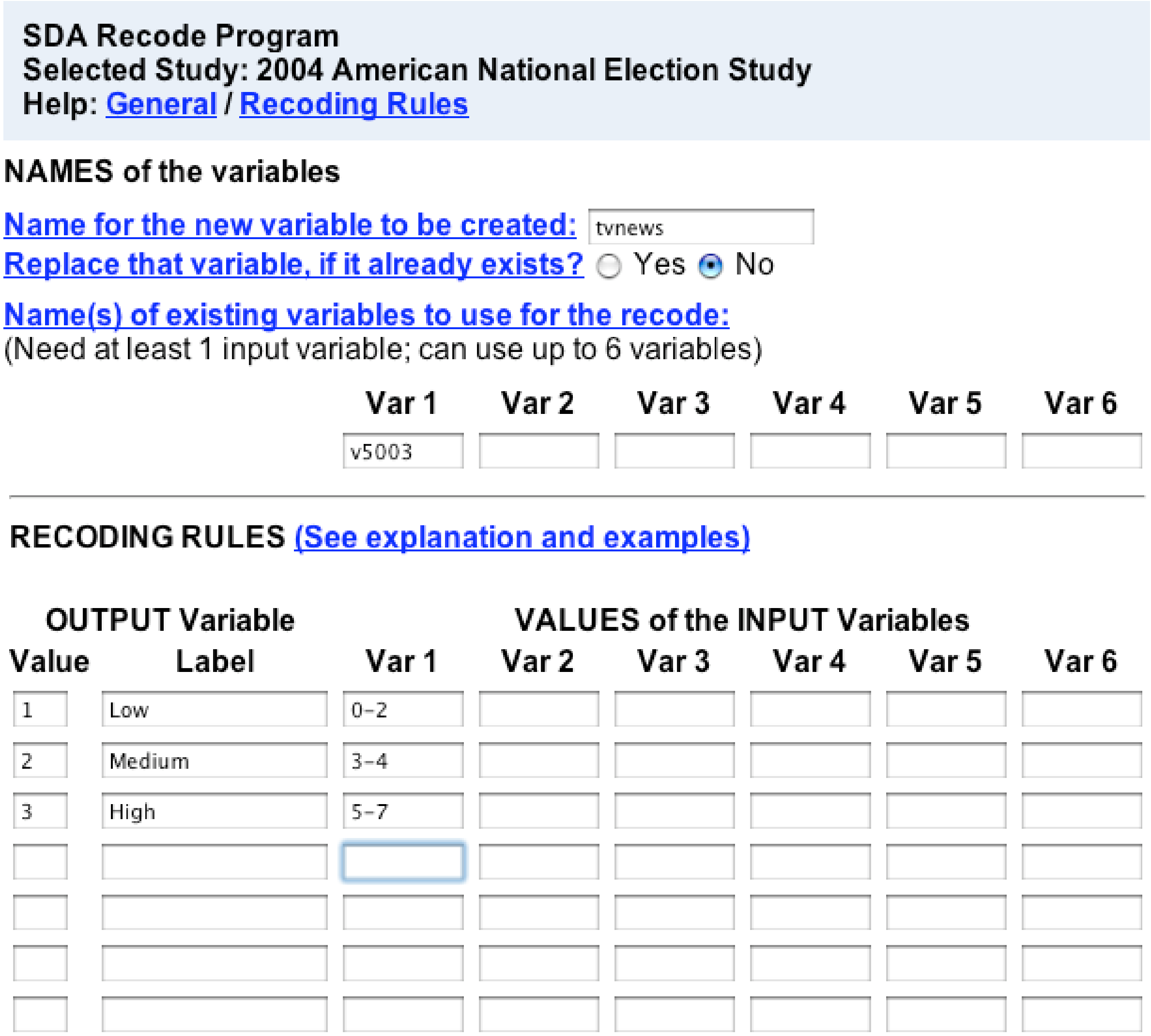
- ○ Browse codebook in this window
- ○ Frequencies or crosstabulation
- ○ Comparison of means
- ○ Correlation matrix
- ○ Comparison of correlations
- ○ Multiple regression
- ○ Logit/Probit regression (See note)

- ● Recode variables (into public work area)
- ○ Compute a new variable
- ○ List/delete derived variables

- ○ Download a customized subset

And then press "Start". You will then see the following screen. At the very top, you write a new name for your variable, whatever you want. Here, I chose "tvnews". Next, you need to tell SDA what variable you want to recode. So we enter "v5003" in the space for Var 1.

In the next section, under "Recoding Rules", we have to write out the rules. In the OUTPUT section, we are telling SDA what the numeric placeholders should be for our new variable, where the placeholder for 1 should be labeled "Low", for instance. The part to the right is more important. Here we have told SDA that the "Low" category should be compiled from responses "0-2" days per week in v5003. (THE NUMBERS UNDER THE "Values" LABEL IN THE DESCRIPTION OF THE ITEM ABOVE). You do this for each of the values of Medium and High, and you will then see the results after you press "Start Recode".



SDA Recode Program
Selected Study: 2004 American National Election Study
Help: General / Recoding Rules

NAMES of the variables

Name for the new variable to be created: tvnews
Replace that variable, if it already exists? ○ Yes ● No

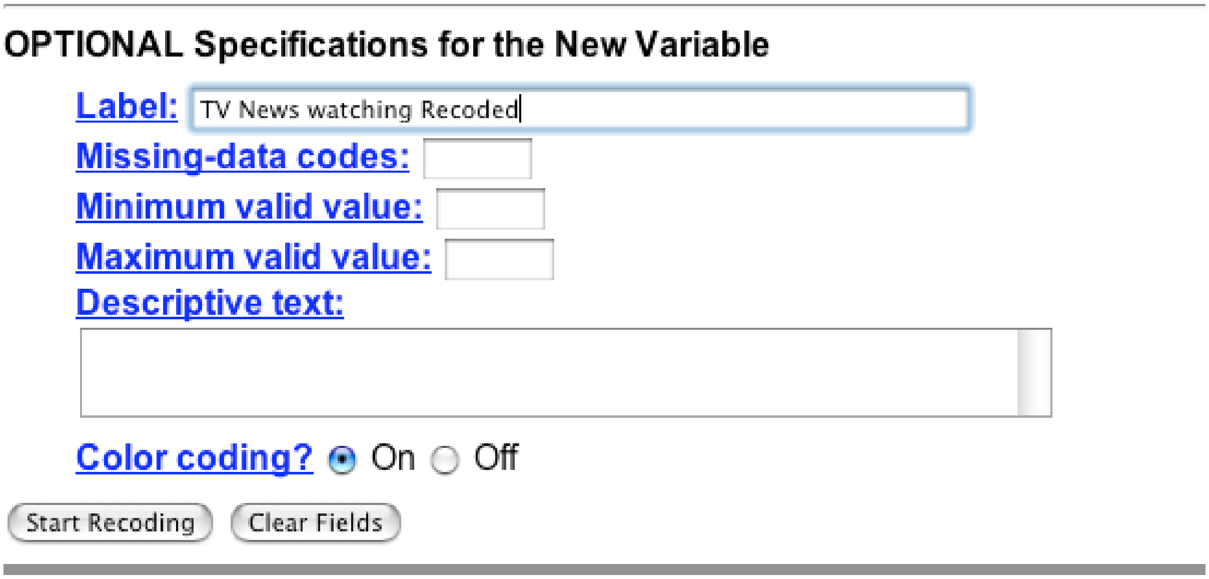
Name(s) of existing variables to use for the recode:
(Need at least 1 input variable; can use up to 6 variables)

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
v5003					

RECODING RULES (See explanation and examples)

OUTPUT Variable		VALUES of the INPUT Variables					
Value	Label	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
1	Low	0-2					
2	Medium	3-4					
3	High	5-7					

While optional, you can label your new variable:



OPTIONAL Specifications for the New Variable

Label: TV News watching Recoded

Missing-data codes:

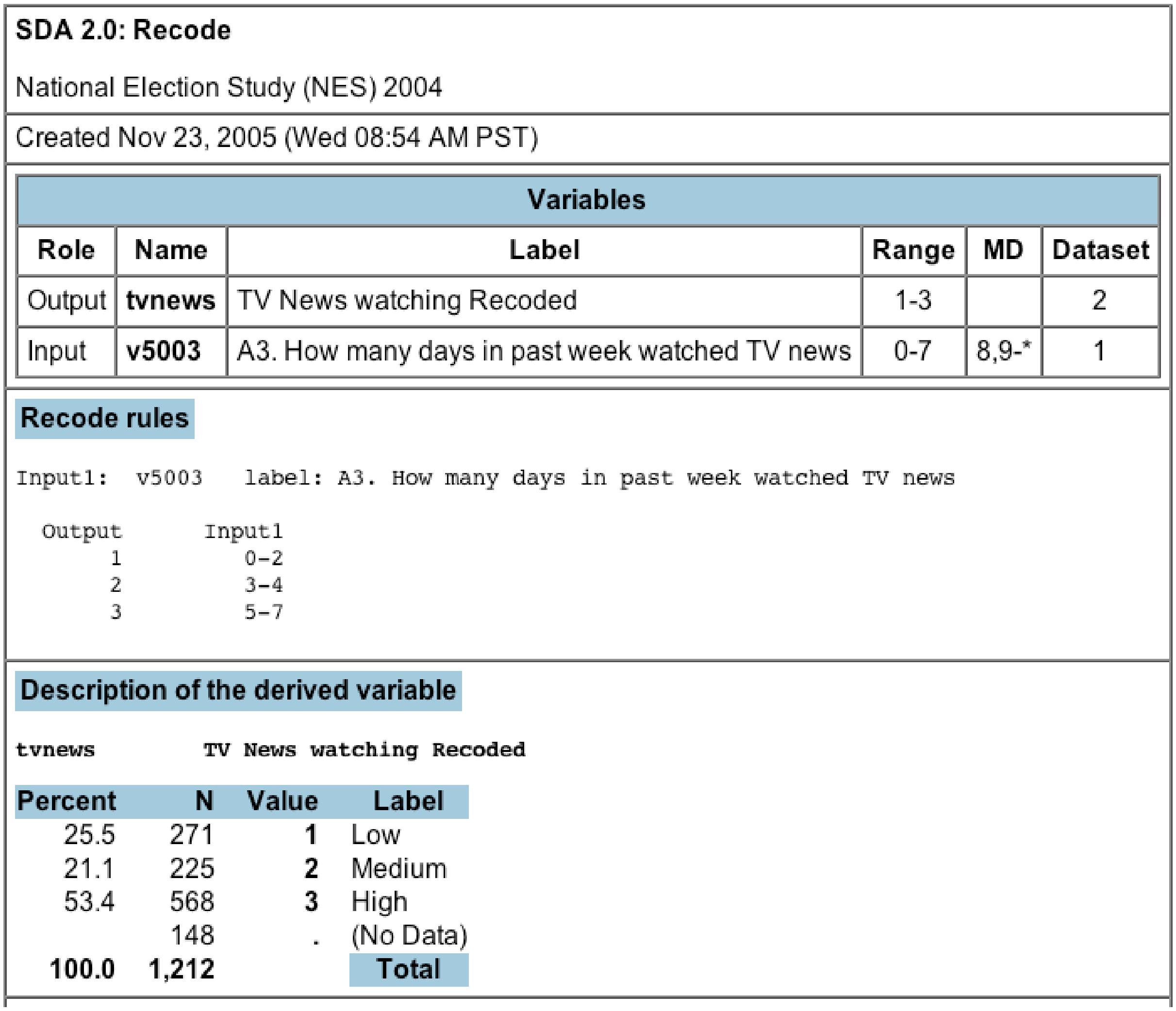
Minimum valid value:

Maximum valid value:

Descriptive text:

Color coding? ● On ○ Off

Here are the results:
I



SDA 2.0: Recode

National Election Study (NES) 2004

Created Nov 23, 2005 (Wed 08:54 AM PST)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Output	**tvnews**	TV News watching Recoded	1-3		2
Input	**v5003**	A3. How many days in past week watched TV news	0-7	8,9-*	1

Recode rules

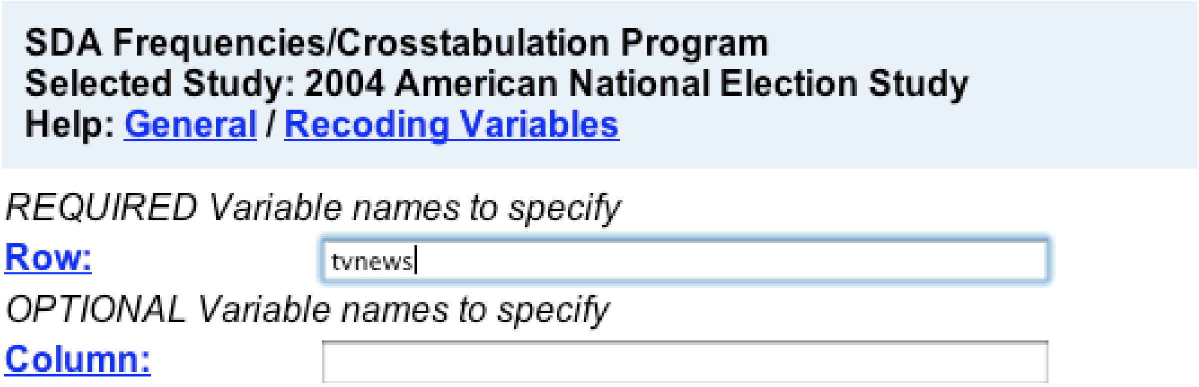
```
Input1:  v5003   label: A3. How many days in past week watched TV news

  Output      Input1
       1         0-2
       2         3-4
       3         5-7
```

Description of the derived variable

```
tvnews        TV News watching Recoded
```

Percent	N	Value	Label
25.5	271	**1**	Low
21.1	225	**2**	Medium
53.4	568	**3**	High
	148	.	(No Data)
100.0	**1,212**		**Total**

If you look at the section "Description of the derived variable" above, it looks like we recoded it correctly. So, we can check this by producing a cross-tabulation of the variable. We just use the name of the variable, "tvnews", just like any other variable name.



SDA Frequencies/Crosstabulation Program
Selected Study: 2004 American National Election Study
Help: General / Recoding Variables

REQUIRED Variable names to specify

Row: tvnews

OPTIONAL Variable names to specify

Column:

In the next posting, we will go over an example of creating an entirely new variable derived from one or more sources.