CHAPTER 2

TRANSFORMING VARIABLES

In this chapter, tools for restructuring variables are introduced. IBM SPSS Statistics allows numerous ways to reconfigure, combine, and compute variables in a data set.

RECODING AND COMPUTING VARIABLES

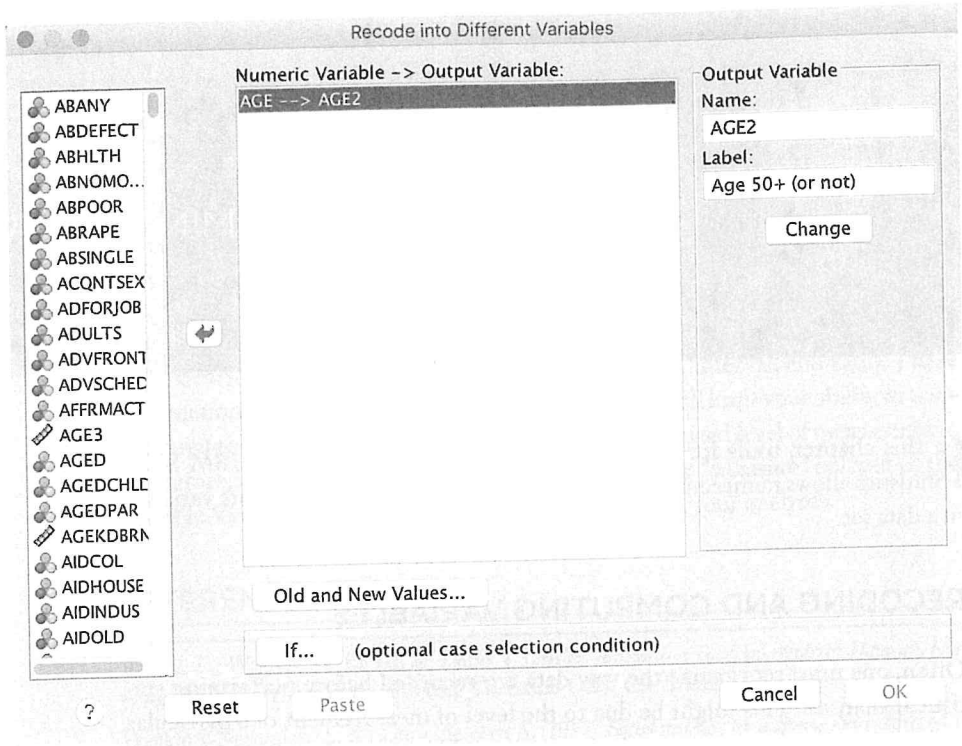
Often, one must reorganize the way data are recorded before performing statistical analyses. This might be due to the level of measurement of a particular variable a researcher wishes to change, or it could be related to the researcher's intended use of a variable. One may wish to collapse a few categories of a variable into one for appropriateness of analysis. For example, within the marital status variable, one might combine the "married" and "separated" categories to form "legally married," and combine "divorced" with "single, never married" to form "unmarried." "Recode" is the SPSS Statistics function that allows a researcher to recategorize a variable to suit the needs of the analysis.

There are many times when a researcher needs to produce a new variable from existing information in a data set but that information is not contained solely within one variable. SPSS Statistics has a "Compute" function that allows a user to both perform mathematical operations on variable data and combine data from multiple distinct variables within the file.

For this recoding example, we will use the General Social Survey (GSS) 2016 data set. We will take a straightforward case of dichotomizing age from a ratio variable, presenting the respondent's actual age at the time of interview, into just two categories with a cut point of 50 years of age. This would allow the researcher to present ages in a small frequency table, whereas before recoding age, there are so many categories (18 to 89 years of age) that presentation in a frequency table is not feasible. To recode, or change, the categories of a variable, select the "Transform" menu, and then choose the "Recode into Different Variables . . ." option:

Transform → Recode into Different Variables . . .

You will then be shown a dialog box like the one displayed in the following image:



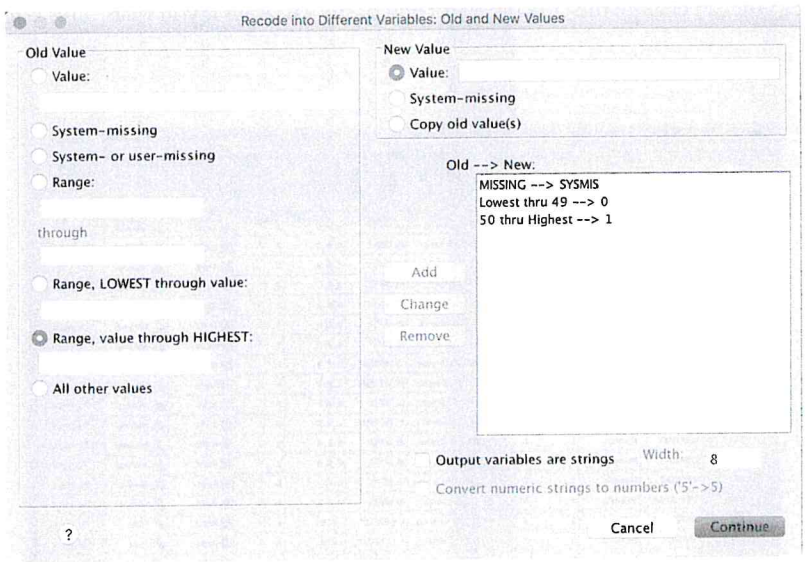
It is useful and proactive against data loss to recode into "different variables" rather than "same variables" if you are reducing information contained within the variable. For instance, if you are recoding a ratio-level variable into an ordinal or a dichotomous variable, you would want to create a different variable. The reason behind this is that the lost information resulting from the recoding would still be retained in the original variable should you want to change the way in which you recode the variable at some later point in time or should you determine that you need the more detailed ratio-level information for your analysis.

RECODING VARIABLES: DICHOTOMIES AND DUMMY VARIABLES

In this example, we recode the age variable into a dichotomy: a variable with exactly two categories, not including missing values. A dummy variable is a dichotomy usually coded with a value of "1" to indicate the existence of a particular attribute. All other attributes are coded "0" as an indication that the particular attribute represented by "1," usually the name of the variable itself, is not present. Dummy variables are particularly useful for including nominal-level variables in analyses requiring interval- or ratio-level variables to perform statistical operations, including multiple regression models (see Chapter 8, "Correlation and Regression Analysis"). It is also possible to use a series of dummy variables to represent several attributes from a given nominal variable.

To dichotomize the age variable, first select "AGE" from the list of variables n the data set on the left, and then click the arrow to drag it to the "Numeric /ariable" pane. Now, create a new name for the variable; in this example, the mun-lane "AGE2" has been conjured. You may also select a label at this time, or you :an attend to that at a later time through the Variable View tab of the Data Editor creen. For this example, the label "Age 50+ (or not)" was added. The next step is o click the "Change" button. This will enter the name "AGE2" into the "Output /ariable" location, as has been done in the screen image on page 20. Until the 'Change" button is clicked, a question mark will act as a placeholder.

Now, it is necessary to give SPSS Statistics instructions for *how* the variable is to ɔe recoded. In this example, we want to change all ages up to and including 49 into ı category called "0," and all ages 50 and greater into another category called "1." Click the "Old and New Values . . ." button in the "Recode into Different Variables" lialog box, and another dialog box will appear on top, like the one that follows:



To implement the changes, first, under "Old Value," select the radio button that reads "Range, LOWEST through value." Then enter "49" in the box underneath. Then, under "New Value," select "Value" and enter "0." Now, click the "Add" button. This instructs SPSS Statistics to transform all ages up to and including 49 into the category "0."

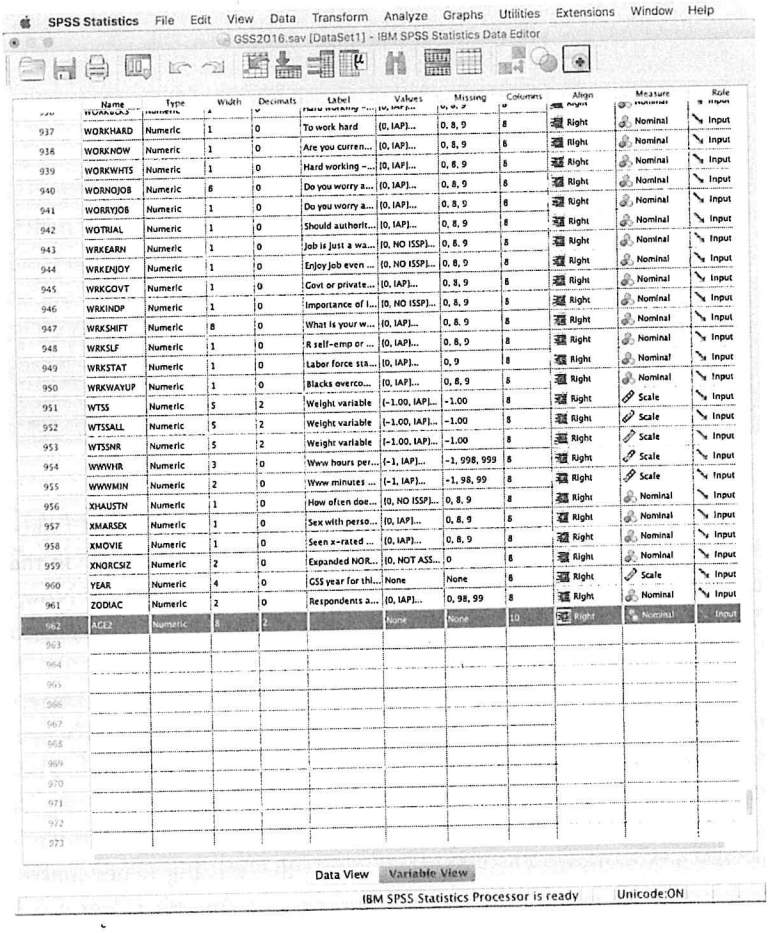
Next, under "Old Value," select the radio button associated with "Range, value through HIGHEST." Enter "50" in the box beneath that heading. Then, under "New Value," select "Value" and enter "1." Again, click the "Add" button. This now instructs SPSS Statistics to transform all ages 50 and beyond into the category "1."

Also, under "Old Value," select "System- or user-missing." Under "New Value" select "System-missing." This will ensure that missing values continue to be treated as such, even if they had been recorded as numeric values. Click the "Add" button once again to confirm that this instruction is added to the list.

Now your instructions have been entered, and you can click the "Continue" button; this will close the current dialog box and return you to the original "Recode into Different Variables" dialog box. Once there, you must click the "OK" button for SPSS Statistics to process your request to recode and then create the new variable.

If the "OK" button is dimmed and SPSS Statistics will not allow you to click it, then one of the preceding steps must not have been completed. The one most often overlooked is the requirement to click the "Change" button, which adds the new variable name for the output variable.

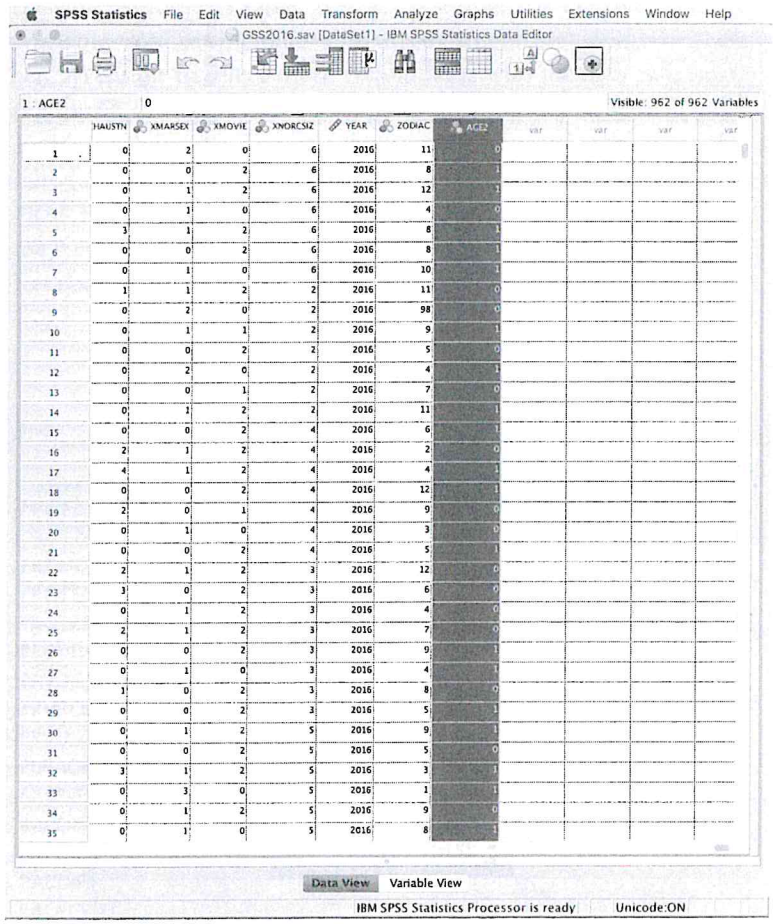
After you have completed the recoding, notice that the new variable is appended to the bottom of the variable list in the Variable View of the Data Editor window. It also happens that the variable is appended to the end (all the way to the right) of the Data View. You can move that variable to another place in the list if you wish by first selecting its entire row and then dragging the row (represented by a red line) up the screen to place it between two of the other variables above. In fact, you can change the "file order" of any of the variables if you wish. This might



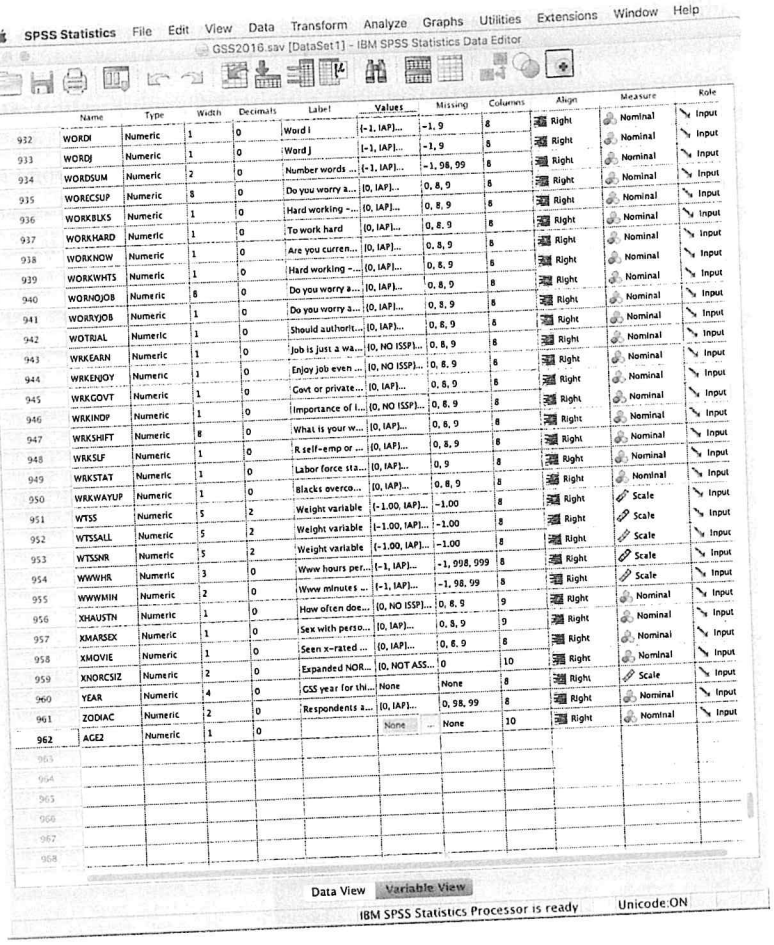
have utility for you if, for instance, you would like two or more variables to be near each other to compare or verify values in the Data View window.

For the variable information, you may wish to change the "Width" column from 8 (the default) to 1, because all possible values are only one digit, including missing case options, such as "IAP" or "Inapplicable" (7), "DK" or "Don't know" (8), and "NA" or "No answer" (9). Similarly, there are no decimals for this variable, so change the "Decimals" column from 2 to 0. Because this variable is dichotomous, select "Nominal" under the "Measure" column.

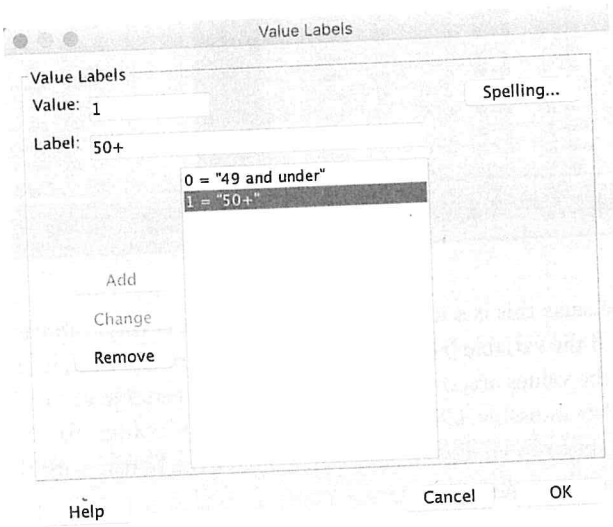
If you switch to the Data View, you can also see that the variable has been appended at the end of the file (all the way to the right), as shown in the following image:



Now, because this is a newly created variable, it is very important to insert value labels. If the variable label and/or value labels are not clear, it can be easy to forget what the values are, or in which direction the variable was coded, thereby making the data unusable. Click the cell in the "Values" column for the variable to which you want to append value labels. Then click the button with three dots, as shown in the following image:



Then, the "Value Labels" dialog box will appear:



As before with value labels, enter both the value and the label. Then click the "Add" button after each value and label you enter. When all value and label combinations have been entered, click "OK," and the value labels will be updated. You can inspect the value labels in the Variable View window by clicking the button with three dots on it in the "Values" column of the variable "AGE2."

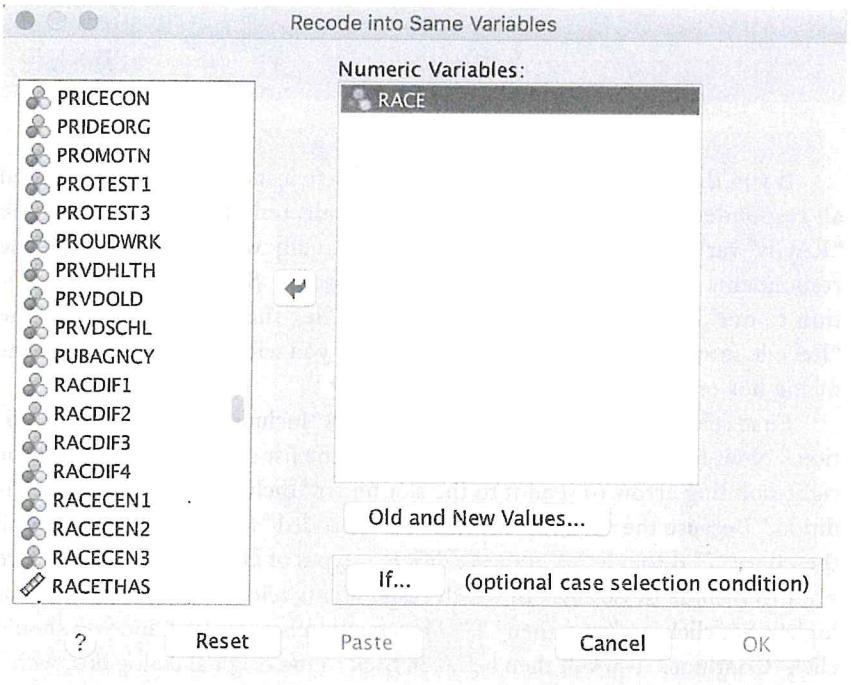
RECODING USING TWO OR MORE VARIABLES TO CREATE A NEW VARIABLE

Suppose you want to create a new variable for race/ethnicity that includes information from the variable "RACE" as well as the variable "HISPANIC." The variable "RACE" includes only three categories: "White," "Black," and "Other." We want to select those respondents who identify as Hispanic and create a fourth category labeled as such.

In this example, we will recode the race variable into itself using information from the variable "HISPANIC." If you would like to preserve the original variable "RACE3," you can do that by duplicating it first; you might call it "RACE3" because it consists of three categories.

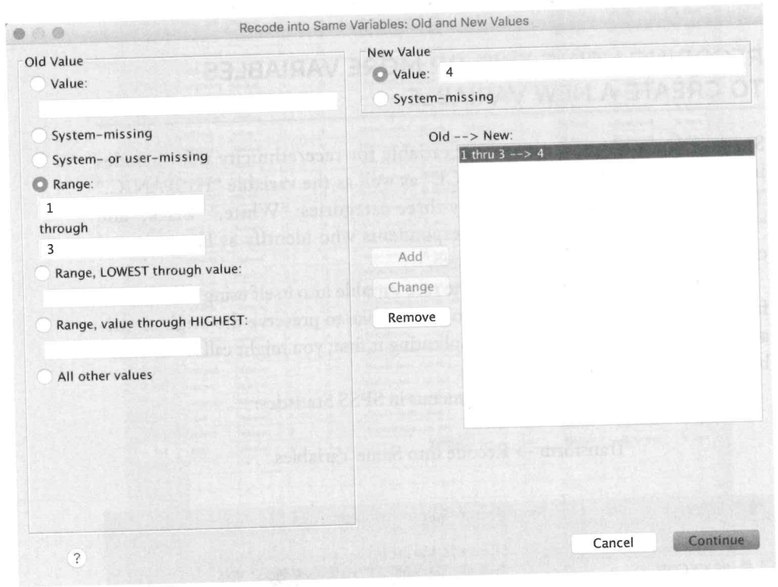
Let's begin. Select the following menus in SPSS Statistics:

Transform → Recode into Same Variables . . .



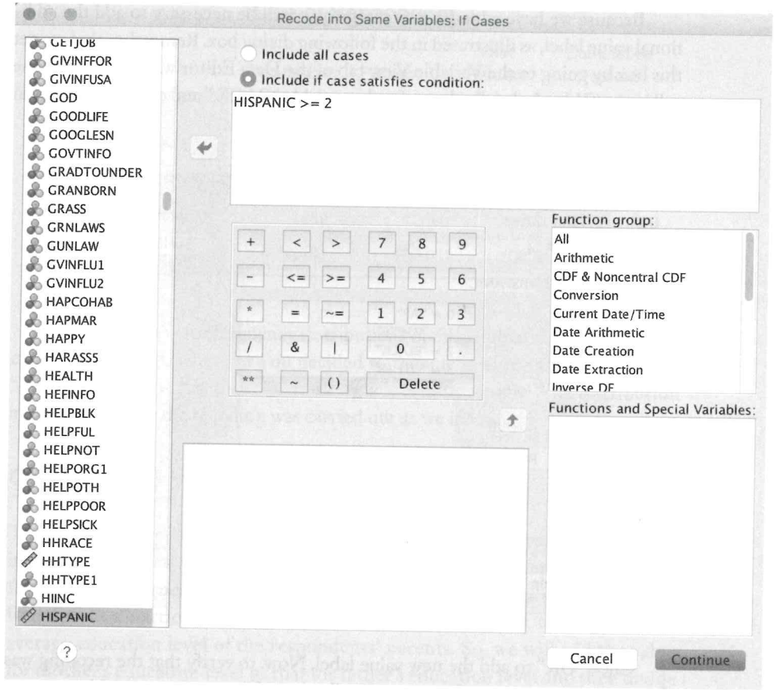
When the "Recode into Same Variables" dialog box appears, select and move the "RACE" variable to the "Numeric Variables" slot.

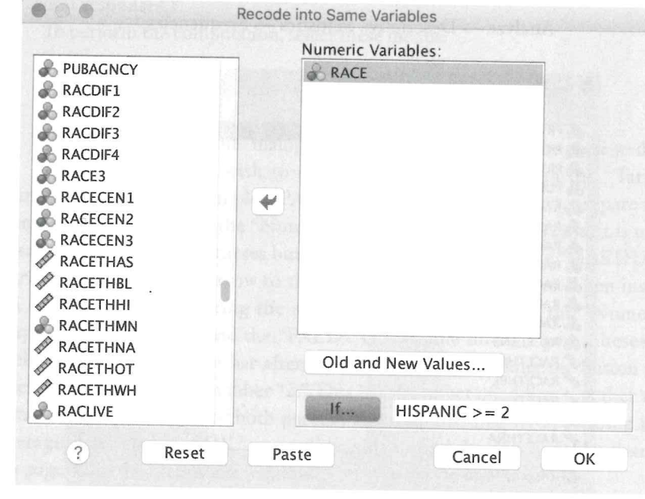
Now, click "Old and New Values . . ." Here, you will make the following entry: Old values from "1" through "3" will be routed to the new value of "4." See the following image to verify your entry. Then click "Continue."



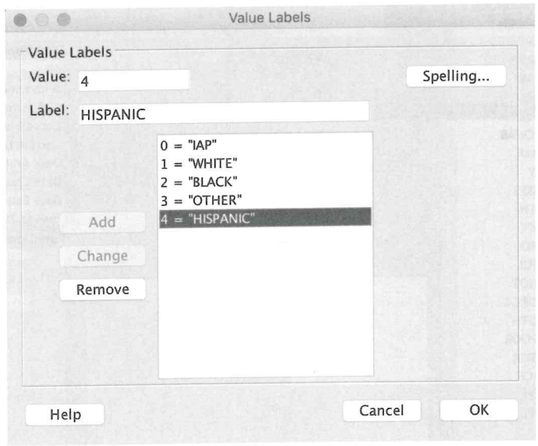
If you think about the logic of what we have just done, we have recoded all respondents (regardless of whether they selected "1," "2," or "3" on the "RACE" variable) to "HISPANIC." But we actually want to recode only the respondents who reported that they were Hispanic. So, let's add that condition to our SPSS Statistics command now. Click the "If . . ." button in the "Recode into Same Variables" dialog box, and you will be presented with the dialog box on page 27.

First, select the radio button associated with "Include if case satisfies condition." Now, find the variable "HISPANIC" in the list on the left, and click the right-pointing arrow to send it to the slot under "Include if case satisfies condition." Because the variable "HISPANIC" is coded "1" for not Hispanic, and the values "2" through "50" indicate different types of Hispanic recognition, we want to include those cases in which respondents selected the number "2" or higher. So, click ">=" and then "2." The equation is complete, and you should click "Continue." You will then be taken back to the original dialog box, which will be updated to reflect the new condition that was just entered. Make sure you see that, as pictured at the bottom of page 27, and then click "OK."



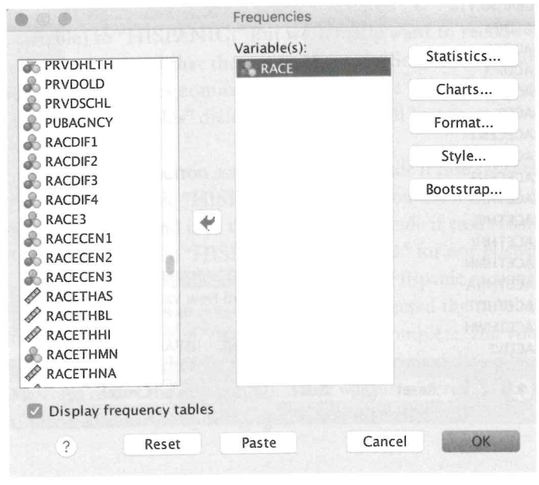


Because we have added a new category, it will be necessary to add the additional value label, as illustrated in the following dialog box. Remember that we get this box by going to the Variable View tab of the Data Editor window, finding the cell in the "Value Labels" column for the variable "RACE," and clicking the button with the three dots.



Click "OK" to add the new value label. Now, to verify that the recoding was done correctly, request a frequency distribution of the new race variable:

Analyze → Descriptive Statistics → Frequencies . . .



Race of Respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	WHITE	1893	66.0	66.0	66.0
	BLACK	468	16.3	16.3	82.4
	OTHER	137	4.8	4.8	87.1
	HISPANIC	369	12.9	12.9	100.0
	Total	2867	100.0	100.0	

You can also ask for frequency distributions of the original race variable, now called "RACE3" or whatever you decided to rename it, along with the variable "HISPANIC" to verify that the recoding was done properly. The distribution above reveals that the recoding was carried out as we intended.

COMPUTING VARIABLES

There are numerous reasons why a user of SPSS Statistics would be interested in computing a new variable. For example, one may want to construct an index from individual questions, or one may wish to compute the logarithmic (log) function of a particular variable. In this example, we want to compute the average education level of the respondents' parents. So, we will add the value for mother's education level to that for father's education level and then divide by 2. (In more sophisticated approaches, we might divide by the number of parental responses.)

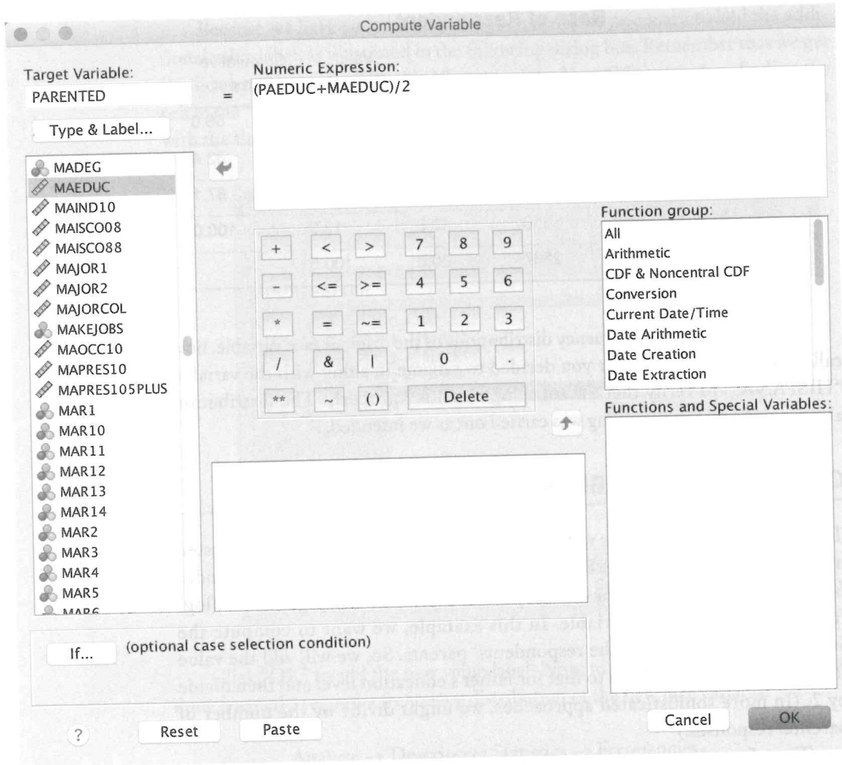
To perform the computation, select these menus:

Transform → Compute Variable . . .

The "Compute Variable" dialog box, shown on page 30, will be presented.

Type the name you wish to assign to the new variable in the "Target Variable" box; in this example, "PARENTED" was chosen. Next, prepare the computation equation in the "Numeric Expression" box. In this case, it is necessary to select the parentheses button "( )" first. Then insert the "MAEDUC" variable by clicking the arrow to the right of the list of variables. Then insert an addition sign by clicking the appropriate button beneath the "Numeric Expression" box. Now, add the "PAEDUC" variable into the parentheses as well. Next, insert a divisor bar after the parentheses by clicking the button for such, and then click the number "2." This has the effect of adding together the total years of education of both parents and then dividing by 2, yielding the average. After clicking "OK," you will be able to find the new variable, as seen on page 31 in the Data View window.

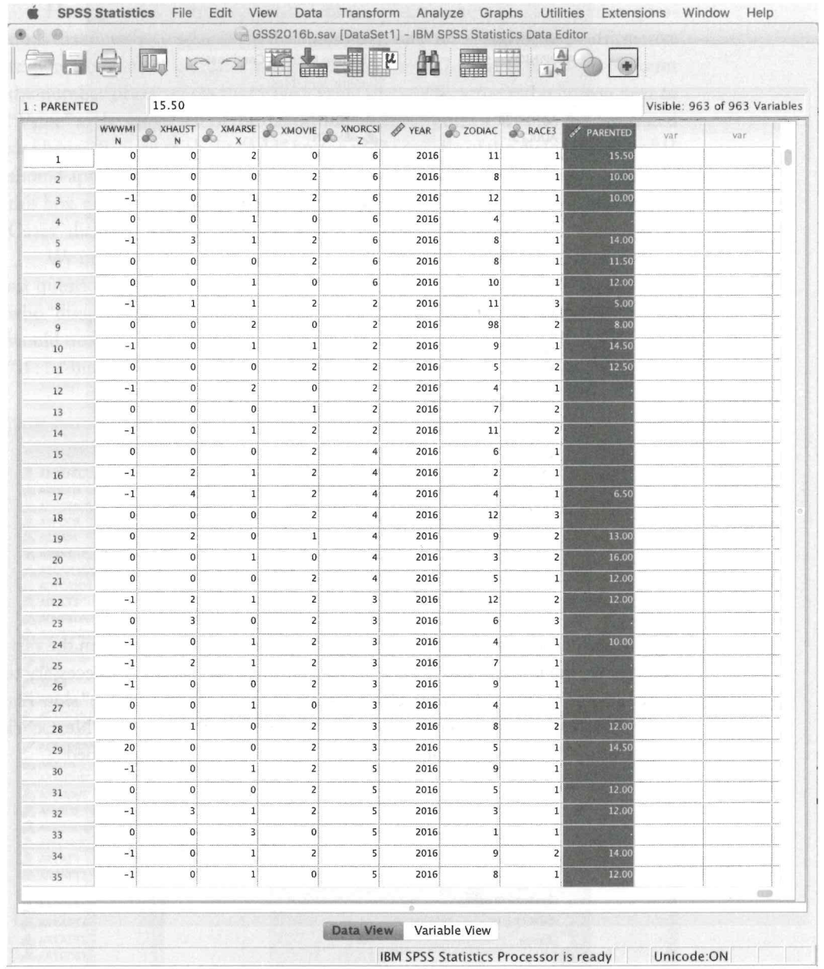
For users who are more comfortable with this process, note that you can type the expression you wish SPSS Statistics to calculate directly into the



"Numeric Expression" box. In this case, you could have typed the expression "(MAEDUC + PAEDUC) / 2" using your computer keyboard. If you use this method, be very careful to get the spelling of the variable names exactly correct, or SPSS Statistics will not be able to execute your command.

The "Type & Label . . ." button opens a small dialog box that offers an immediate opportunity to enter a variable label and also to define the type of variable (e.g., numeric).

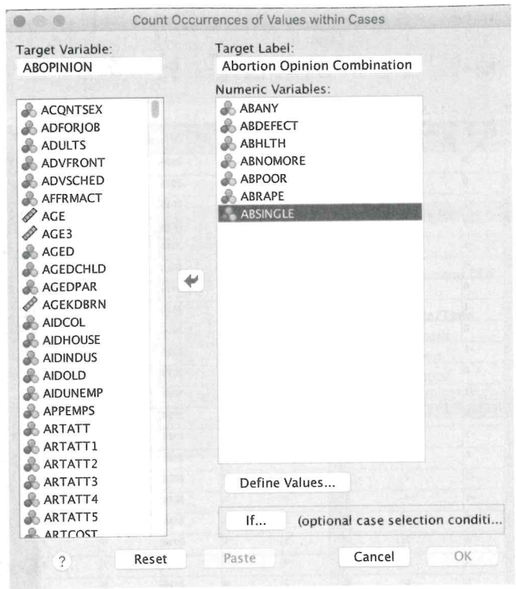
Note the functions that are available: Each selection in the "Function group" pane will bring up a separate list of functions in the "Functions and Special Variables" pane in the lower right corner of the "Compute Variable" box. There are statistical, trigonometric, date, time, string, and other functions that can be used to compute just about anything. Also, if you wish to set up a conditional computation—such that a computation is made in only one case, or there are to be different computations for different cases on the basis of some predetermined condition—then select the "If . . ." button in the lower left corner and enter one or more conditions. The same functions and keypads are provided to instruct SPSS Statistics how to determine the criteria for the conditional computation.



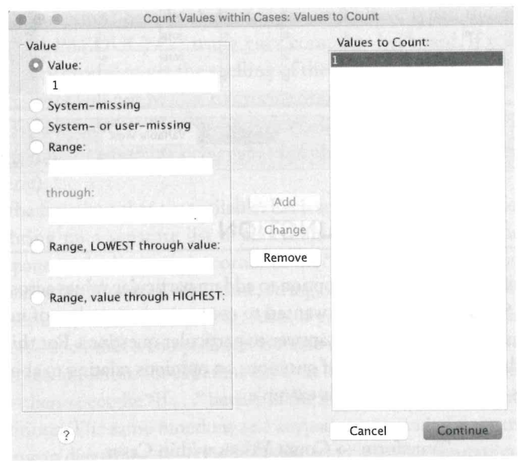
USING THE COUNT FUNCTION

SPSS Statistics allows users the option to add up particular values across different variables. Suppose a researcher wanted to count up the number of instances in which a respondent gave a "yes" answer to particular questions. For this example, consider the GSS (2014) series of questions on opinions relating to abortion. Use the menus below to carry out this example:

Transform → Count Values within Cases . . .

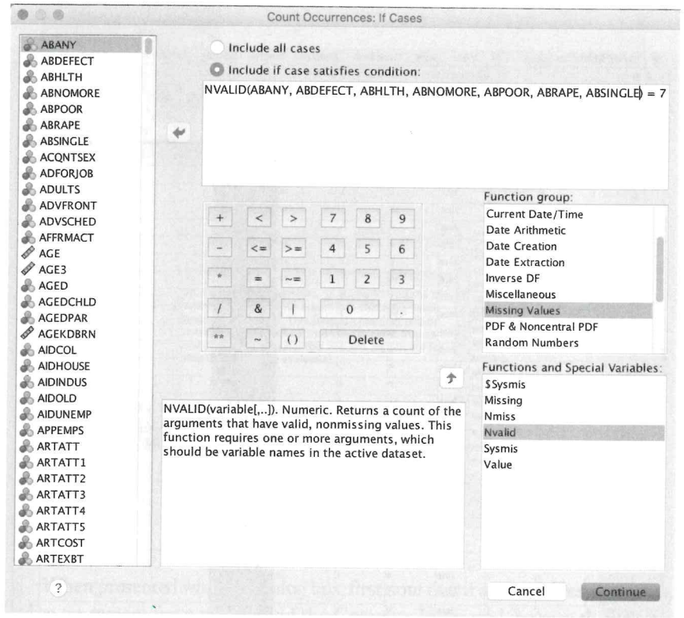


In the dialog box that appears, move the appropriate variables from the variable list on the left into the "Numeric Variables" box. It will also be necessary to enter a name for the new variable to be created in the "Target Variable" slot. The target label can be conveniently entered in the appropriate slot as well. Next, click the "Define Values . . ." button. The following dialog box will be provided:



Here, you will want to select the values to be counted. For the opinion questions that have been selected in this example, a code of "1" indicates an affirmative response (and a "2" indicates a negative response). Therefore, you want to count the number of responses that have a value of "1." Click the radio button next to "Value" at the upper left of the dialog box, and then enter a "1" into the associated slot beneath. Now, click the "Add" button in the middle of the dialog box. The "1" should appear in the "Values to Count" area. Now, click the "Continue" button in this box, and you will be taken back to the "Count Occurrences of Values within Cases" dialog box.

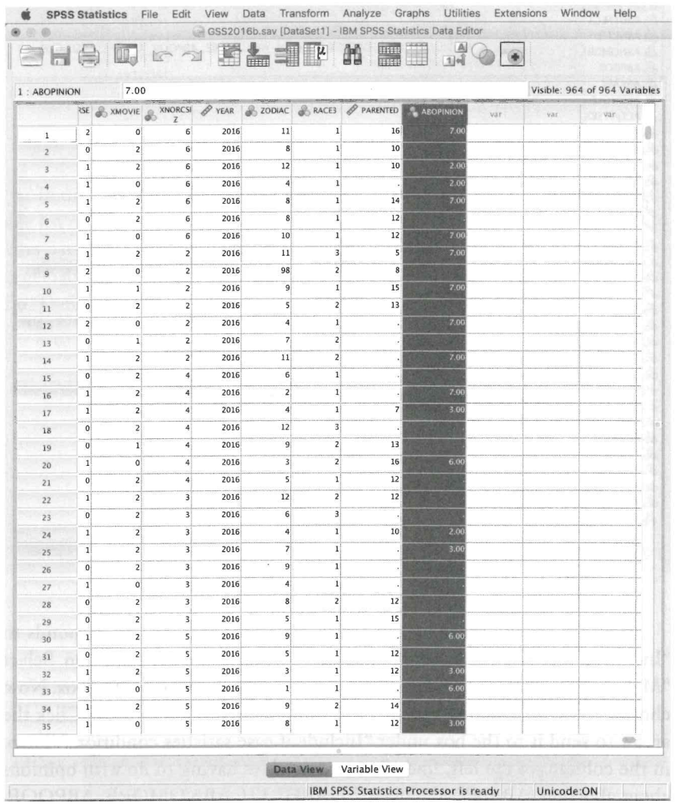
We need to make sure that someone who didn't answer one or more of these six questions—or who didn't answer any of them—is not counted as someone who "disapproves." That would corrupt our data, and the results from this analysis would not be valid. We can prevent this by using a conditional filter: Click the "If . . ." button, and the following dialog box will appear:



First, select the radio button at the top of the box that corresponds to "Include if case satisfies condition." Next we will prepare the condition. Select "Missing Values" from the "Function group" on the right side of the box. Now, choose "Nvalid" from the list of "Functions and Special Variables." Click the arrow to send it to the box under "Include if case satisfies condition." Then, in the column on the left, find all seven variables having to do with opinions about abortion (ABANY, ABDEFECT, ABHLTH, ABNOMORE, ABPOOR,

ABRAPE, and ABSINGLE) and drag them between the parentheses next to "NVALID," each separated by a comma. This will return a count for each variable that has a valid response. Therefore, if the value returned is a "7," that means that all seven variables contained valid responses for that case. We want all seven items to have valid responses; otherwise, our new variable might include truncated values (e.g., if someone answered only three questions and approved on all three, the new variable would make it look as if he or she approved of only three out of seven, or fewer than half, when the respondent actually approved in all cases for which he or she replied). So, set this expression equal to "7," as displayed in the previous illustration.

Now click "Continue" and then the "OK" button in the prior dialog box, to which you will be returned after this one closes. Other than a record of the SPSS commands executed, no output will be generated. A new variable, however, will be created. See the following screen image for the data contained within the new variable that has been created:

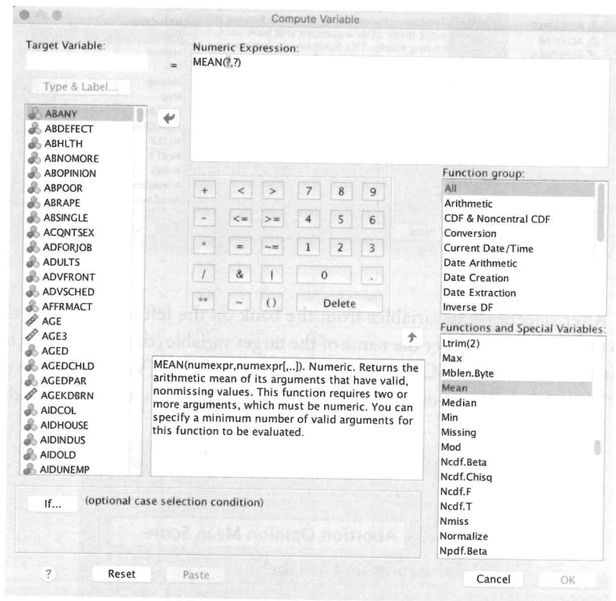


Note that the values, except of course for missing cases, are all appended with ".00." You will want to go back to the Variable View tab of the Data Editor screen to clean up the details, such as width, decimals (change from 2 to 0), and measure.

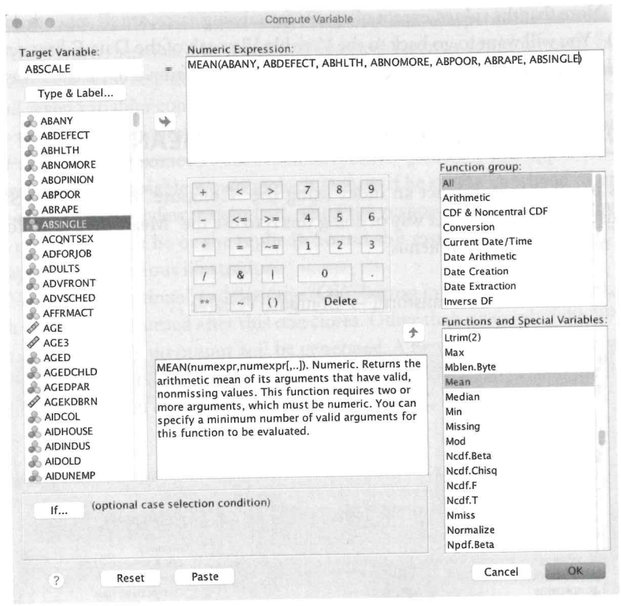
COMPUTING AN INDEX USING THE MEAN

It is possible to construct an index using the "Compute" command in SPSS Statistics. The most direct way of doing this is to use the "Mean" function. To use this method, click these menus:

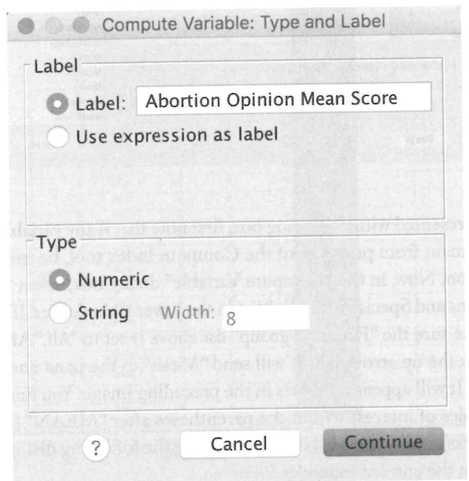
Transform → Compute Variable . . .



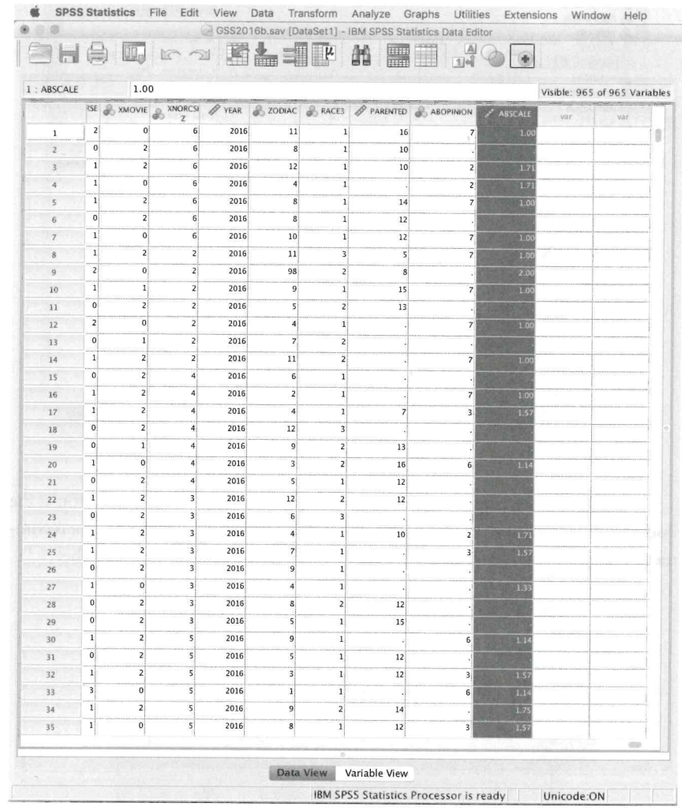
When presented with this dialog box, first note that if any variables, functions, and so on remain from prior use of the Compute Index tool, be sure to click the "Reset" button. Now, in the "Compute Variable" dialog box, select "Mean" from the "Functions and Special Variables" list in the lower right corner. If you don't see "Mean," make sure the "Function group" list above is set to "All." After you select "Mean," click the up arrow, which will send "Mean" to the pane under "Numeric Expression." It will appear as it does in the preceding image. You must then insert all the variables of interest within the parentheses after "MEAN" (replacing the question marks), each separated by a comma. See the following dialog box for how this is done in the current example:



After selecting the variables from the bank on the left and dragging them into the parentheses, enter the name of the target variable (the new variable to be created, in this case "ABSCALE") in the pane under "Target Variable." Then click the "Type & Label . . ." button. The following short dialog box will be provided:

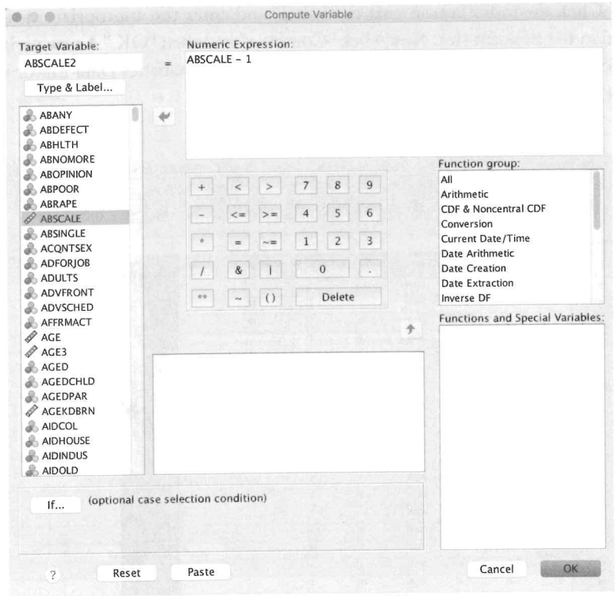


Click the radio button next to "Label" and enter the appropriate variable label in the adjacent slot. Now, click "Continue" and then "OK." A new variable, "ABSCALE," will be created and appear in the SPSS Statistics Data Editor window as shown below:

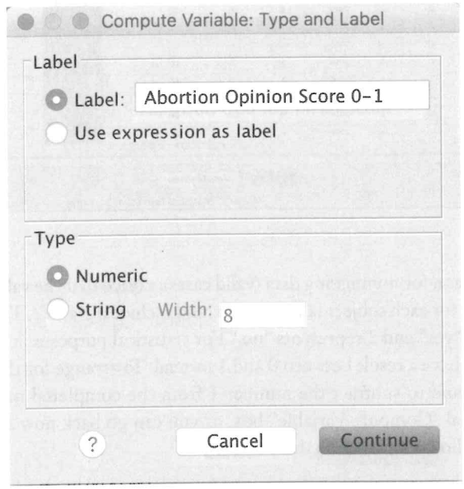


In this case, for nonmissing data (valid cases), notice that the value computed for the mean for each subject falls between (and includes) 1 and 2. This is because 1 represents "yes" and 2 represents "no." For statistical purposes, it is sometimes beneficial to have a result between 0 and 1 instead. To arrange for this, you could add a command to subtract the number 1 from the completed mean function in the original "Compute Variable" box, or you can go back now and make the change, as follows. Again, click these menus:

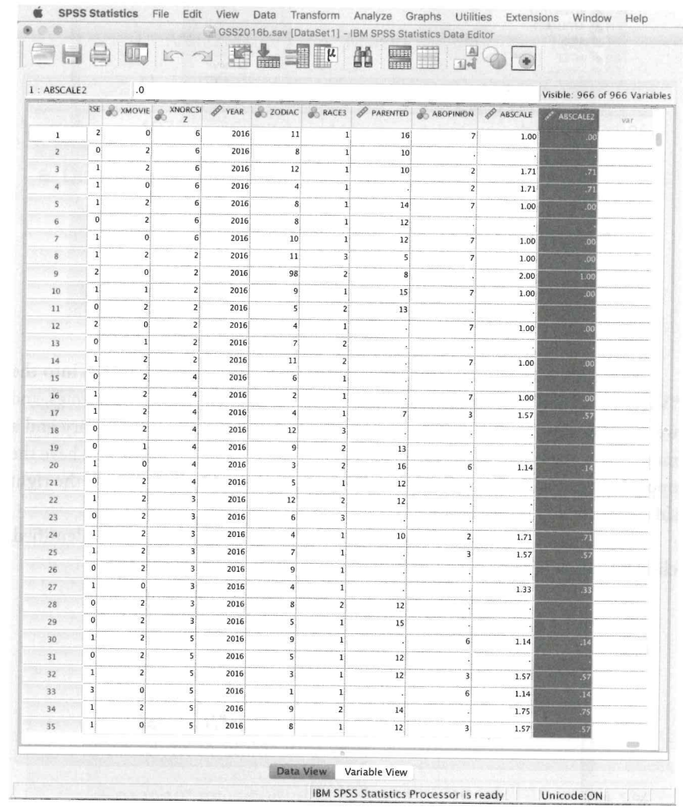
Transform → Compute Variable . . .



If the dialog box presented is populated with data, click the "Reset" button at the bottom of the box. Now, enter a name for the new variable. (Although it is not recommended, you could overwrite the original variable name.) Click "Type & Label . . .":



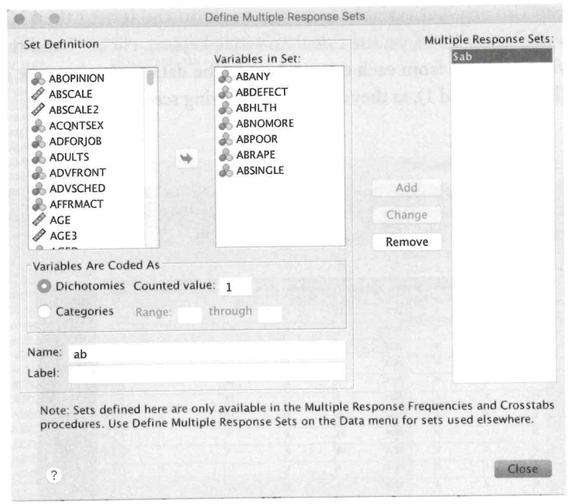
Drag the original variable from the bank on the left to the "Numeric Expression" area. Then, on the calculator-style keypad, click "-" and then "1." This will subtract 1 from each case and move the data means into the desired range (between 0 and 1), as they are in the following screen image:



MULTIPLE RESPONSE

To produce multiple response values (e.g., frequency values combined across multiple variables), choose the following menus:

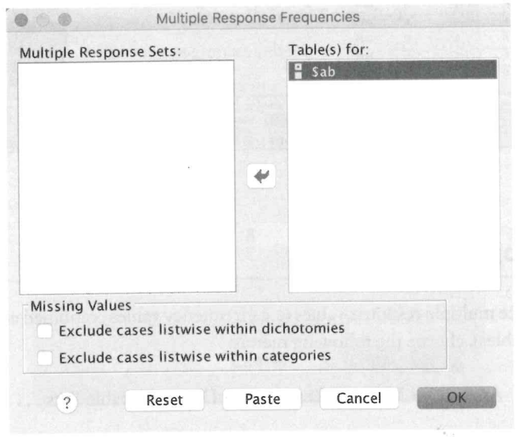
Analyze → Multiple Response → Define Variable Sets . . .



From the variable bank on the left, move all the desired variables into the "Variables in Set" box. In this example, the variables chosen are dichotomies, and we are counting the "yes" value, 1. It is also possible to select a category and a range. Now, name the set and type the name in the "Name" slot. You also have the opportunity to add a label at this time. Finally, click the "Add" button on the right side of the dialog box. You can now click "Close."

To produce a frequency table for the response set that has just been identified, click the following menus:

Analyze → Multiple Response → Frequencies . . .



Select the response set from the list on the left. (In this case, it was the only item in the list.) Move the set to the "Table(s) for" box. Then, click "OK." SPSS Statistics will produce output like the following:

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
$ab[a]	1675	58.4%	1192	41.6%	2867	100.0%

a. Dichotomy group tabulated at value 1.

$ab Frequencies

		Responses		Percent of Cases
		N	Percent	
$ab[a]	Abortion if woman wants for any reason	837	10.9%	50.0%
	Strong chance of serious defect	1353	17.6%	80.8%
	Woman's health seriously endangered	1610	21.0%	96.1%
	Married—wants no more children	859	11.2%	51.3%
	Low income—can't afford more children	816	10.6%	48.7%
	Pregnant as result of rape	1411	18.4%	84.2%
	Not married	782	10.2%	46.7%
Total		7668	100.0%	457.8%

a. Dichotomy group tabulated at value 1.

Note that the multiple response command allows the easy production of a table combining similar-style variables counting a particular category or range.

In the preceding, it is easy to see the similarities and differences in percentages of those who support abortion in the listed circumstances. This frequency table yields the percentage of respondents (here called "Percent of Cases") who approve of abortion in the particular case described. Note, for instance, that 96.1% approve when a woman's health is seriously endangered, whereas only 46.7% approve when the reason is that a woman is not married.