



# Recoding

Dr. Ted Brader

Dr. Abdoulaye Diop

Dr. Kien Le



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# WHY DO WE RECODE DATA?

- Answers → numbers
- Recoding = any transformation of the numerical values in the data
  - EXAMPLE: original codes: 1 = yes, 2 = no  
recoding: 0 = no, 1 = yes
- Getting the data ready for analysis and presentation.



# SOME COMMON TYPES OF RECODING

- Assigning “missing values”
- Consolidating or simplifying answers
- Folding scales
- Rescaling variables to a common range
- Reordering responses
- Centering
- Combining responses to obtain a complete answer to the full question.
- *Turning verbal transcripts of open-ended questions into numerical values.*





# Assigning Missing Values



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# WHAT ARE MISSING VALUES?

- The value of a variable is *missing* for a particular observation (respondent) when no usable value of that variable has been recorded.
- This often means the data are literally missing from the dataset.
- The computer software ignores observations with missing values when performing statistical analyses or making graphs. Only observations with valid data are included.



# WHEN DO WE CODE DATA AS MISSING?

- When the respondent does not provide an answer that is useful for a researcher's analyses or presentations.
- Common reasons for missing values in survey data:
  - *Respondent was not asked the question.* Thus, by design or interviewer error, no data exists for the respondent on this variable. (Often these responses will already be marked as missing when you get a dataset.)
  - *Respondent skips or refuses to answer a question.* Thus, by respondent choice or error, no data exists for the respondent on this question.
  - *Respondent answers "don't know" or equivalent.* In many cases, researchers do not find it useful to include these responses with their analysis of more substantive answers.



# EXAMPLE:

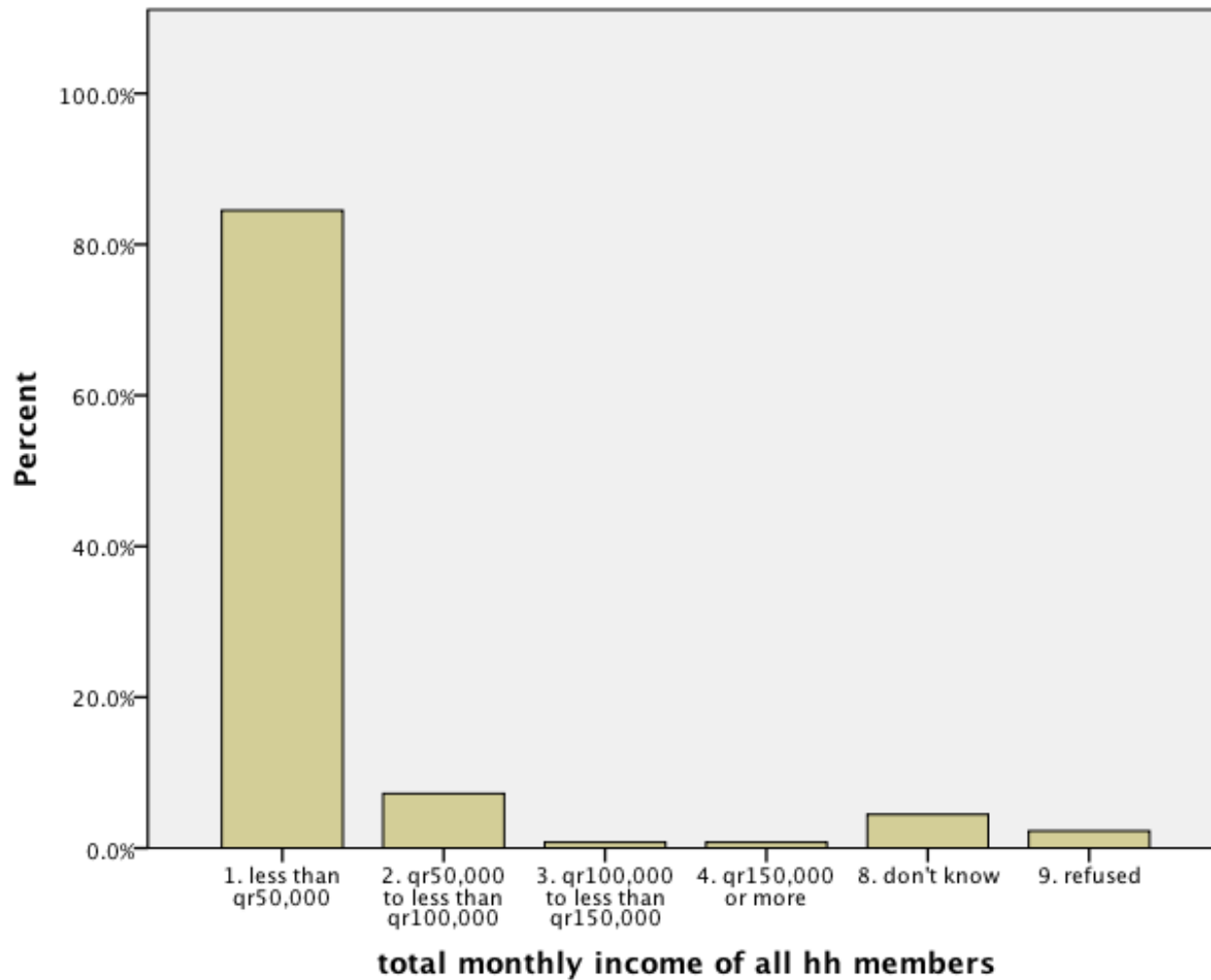
## RAW FREQUENCIES OF HOUSEHOLD INCOME

total monthly income of all hh members

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. less than qr50,000	1231	57.6	84.5	84.5
	2. qr50,000 to less than qr100,000	105	4.9	7.2	91.7
	3. qr100,000 to less than qr150,000	11	.5	.8	92.5
	4. qr150,000 or more	11	.5	.8	93.3
	8. don't know	65	3.0	4.5	97.7
	9. refused	33	1.5	2.3	100.0
	Total	1457	68.1	100.0	
Missing	System	682	31.9		
Total		2139	100.0		



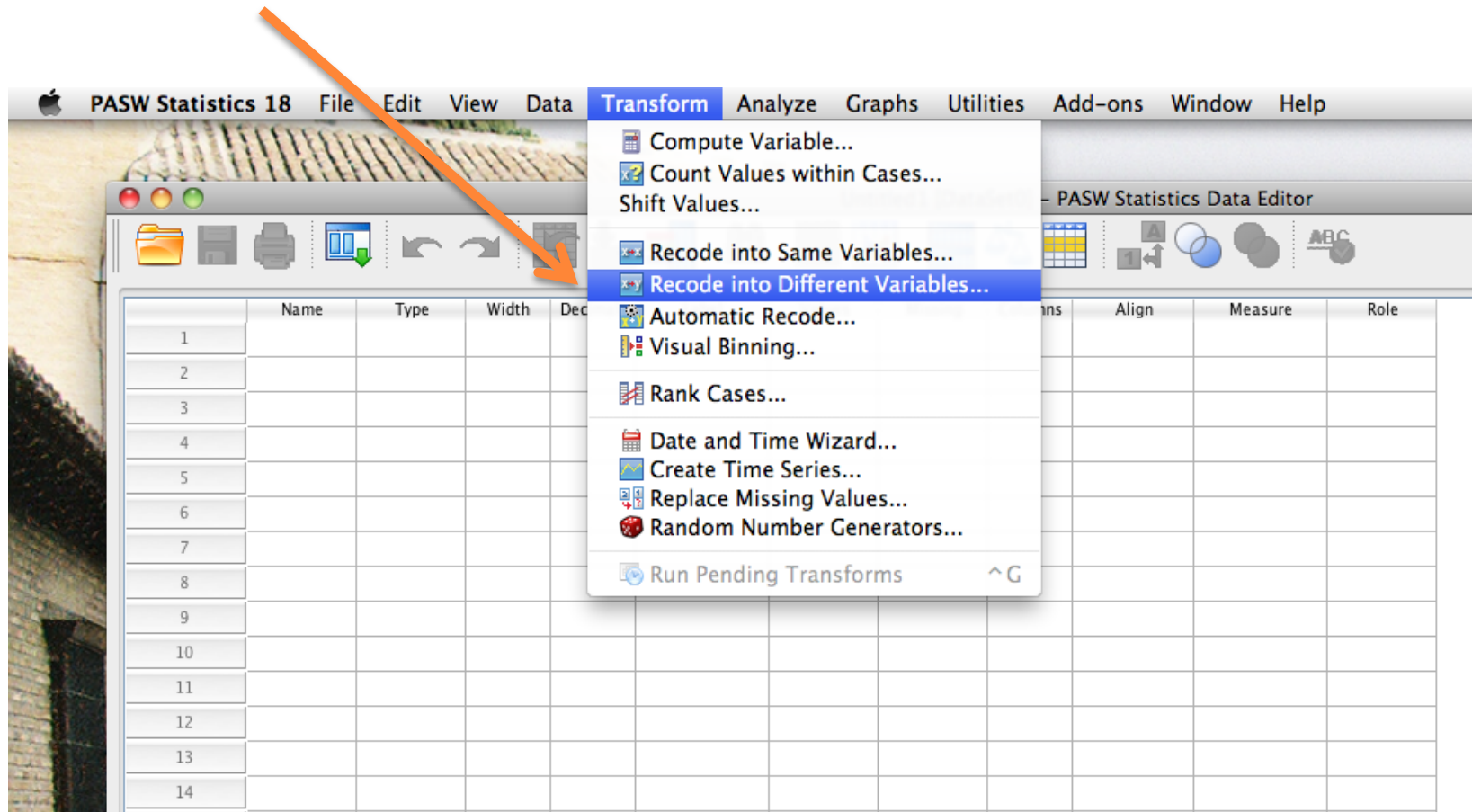
# EXAMPLE: GRAPH OF RAW INCOME DATA





# RECODING IN SPSS

Select Menu Option: “Recode into Different Variables”



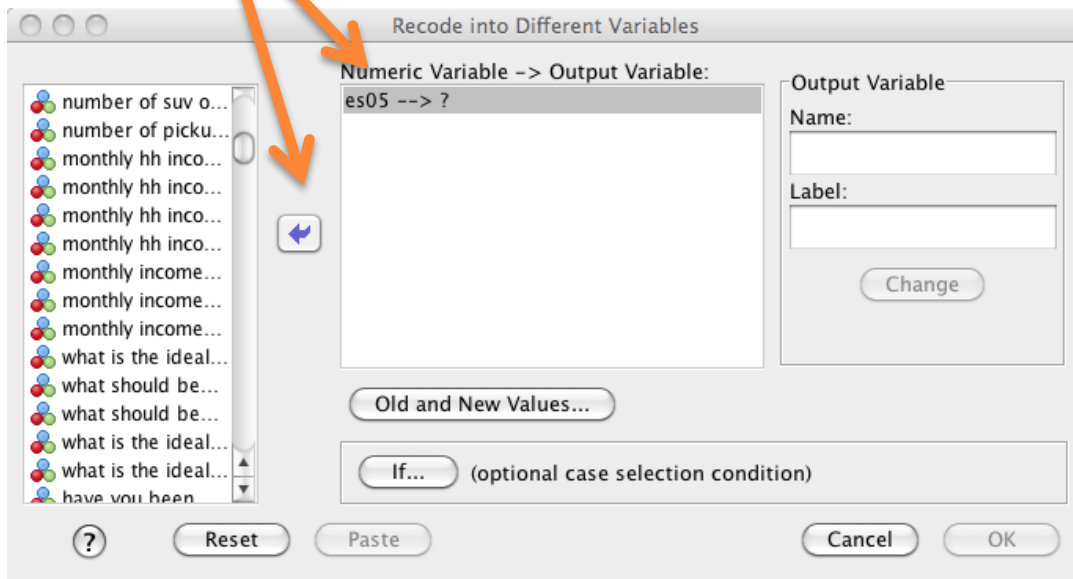
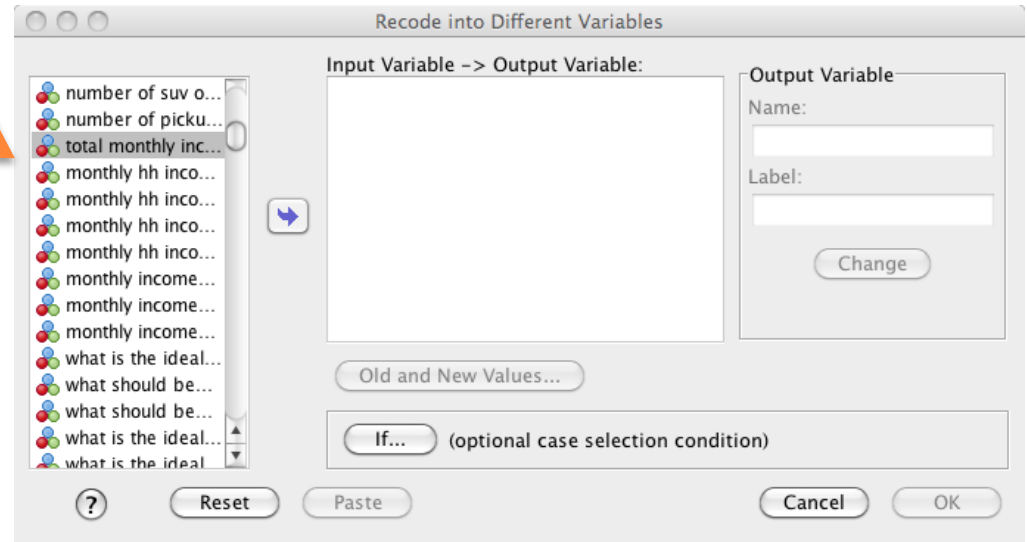
*Why Not “Recode into Same Variables”???*



# RECODING IN SPSS

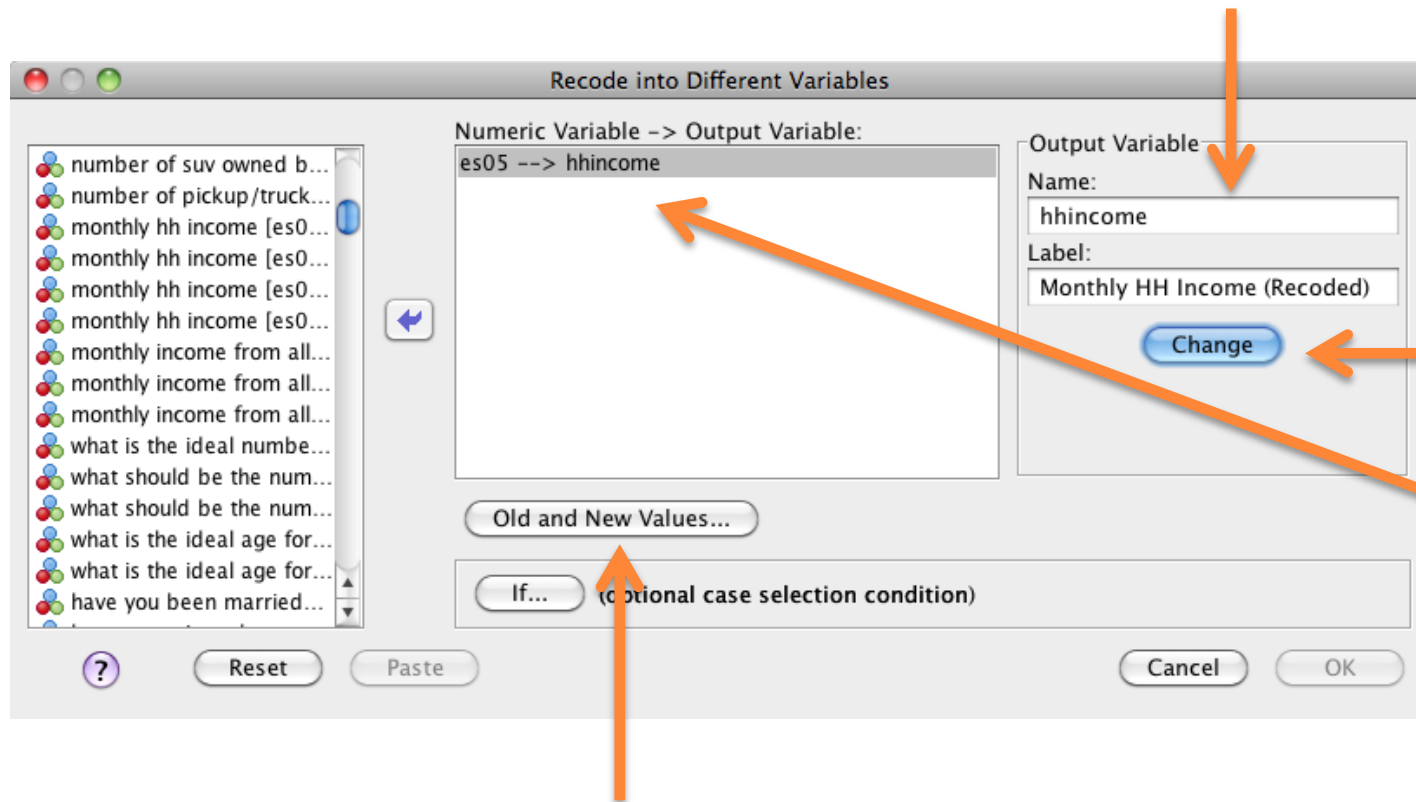
Find the variable you wish to recode in the list of variables on the left. And select it, so it is highlighted.

Then click on the arrow button to indicate that this is the input variable.



# RECODING IN SPSS

Enter the name you want to give to the new (recoded) variable in the box. This is usually a short continuous set of letters and numbers. You can also give the new variable a longer label that provides more explanation.



Then click on the Change button to indicate that this is the output variable.

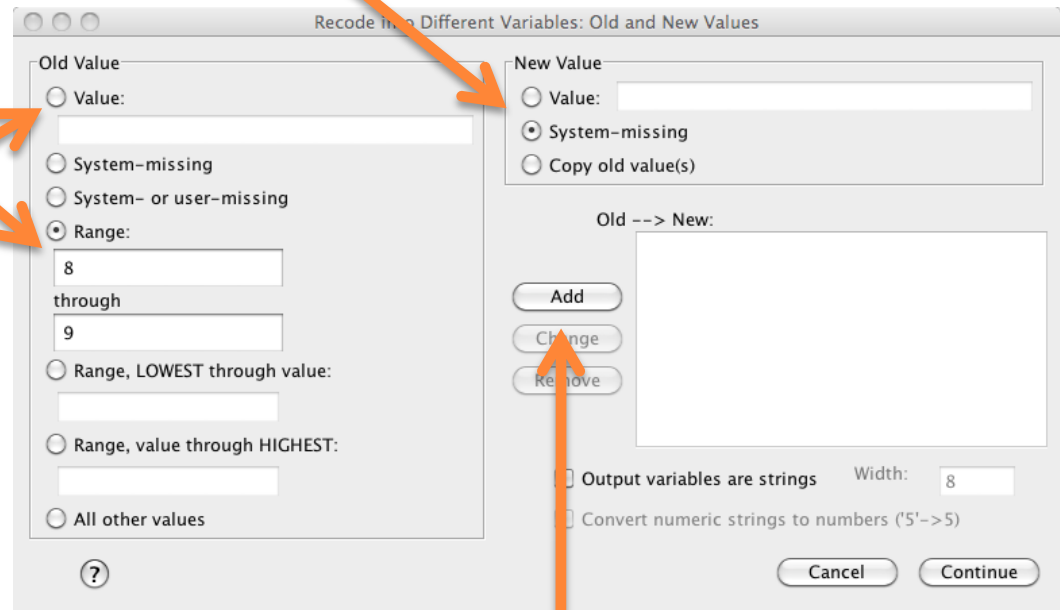
You are now ready to specify how the values of the old variable should be recoded into a new variable. Click on the Old and New Values button.



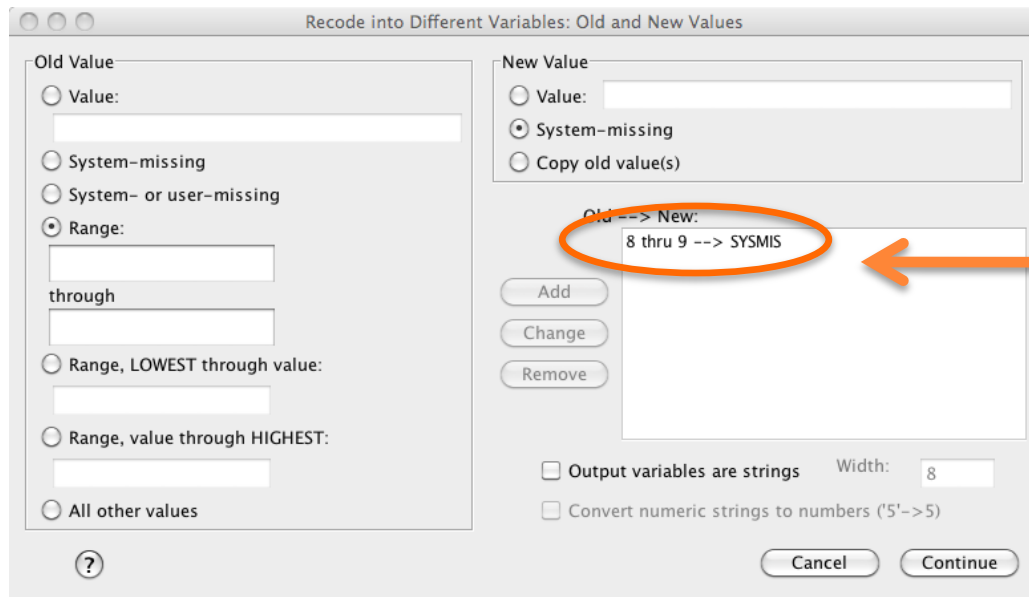
# RECODING IN SPSS

Identify the values of the old variable that you wish to recode as missing values. Enter these on the left side. In this case, we want to change 8 (don't know) and 9 (refuse). We can enter them one at time under Value, or together under Range (since they form a continuous range of numbers).

Select "System-missing" as the new value on the right side.



The dialog box "Recode into Different Variables: Old and New Values" is shown. On the left, under "Old Value", the "Range:" option is selected with "8" in the first box and "9" in the second box. On the right, under "New Value", the "System-missing" option is selected. Below this, the "Old --> New:" list is empty. Buttons "Add", "Change", and "Remove" are visible. At the bottom right are "Cancel" and "Continue" buttons.



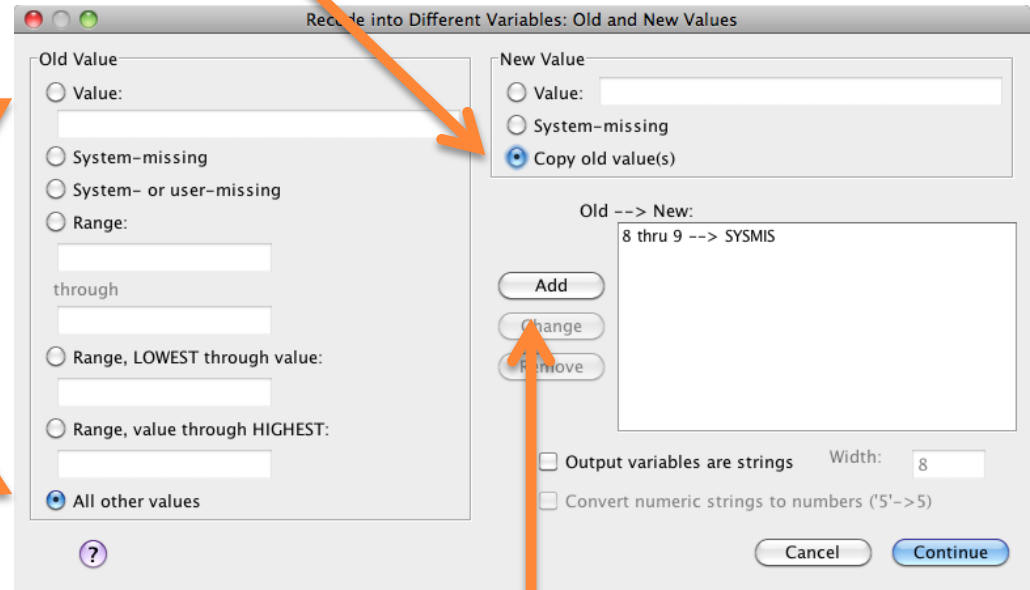
The same dialog box is shown, but now the "Old --> New:" list contains the instruction "8 thru 9 --> SYSMIS", which is circled in orange. The "Add" button is located below the list. The "Old Value" section remains the same. The "New Value" section still has "System-missing" selected. The "Cancel" and "Continue" buttons are at the bottom.

Then click the Add button to enter this recoding instruction into the list of recoding instructions, which will appear in the box on the right.

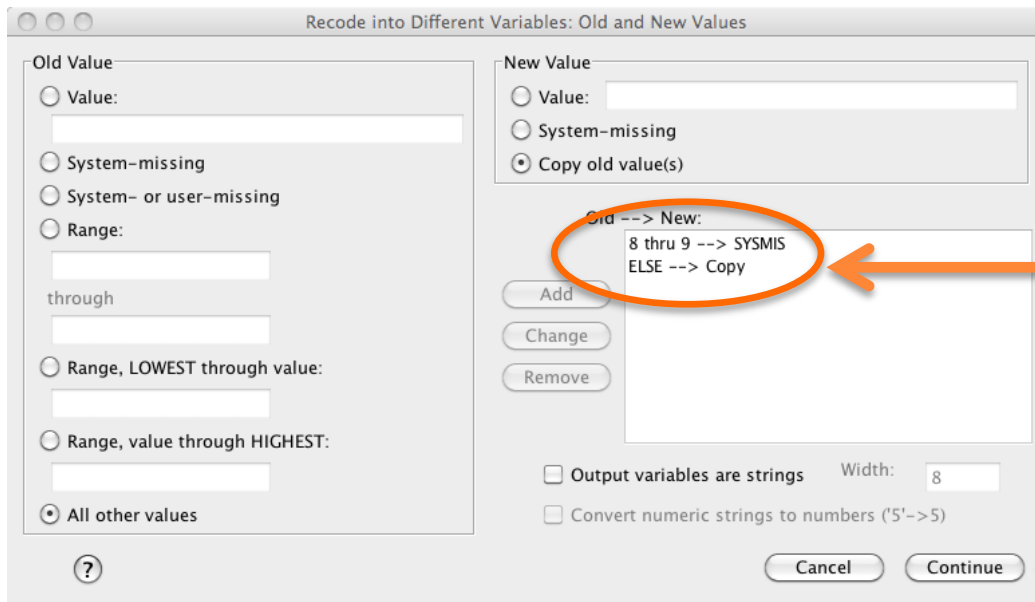
# RECODING IN SPSS

For now, those are the only values we wanted to change in the old variable. We want all other values of the new variable to be the same as the old variable. So we select “All other values” on the left side.

And select “Copy old values” on the right side.



The dialog box is titled "Recode into Different Variables: Old and New Values". On the left, under "Old Value", the "All other values" radio button is selected. On the right, under "New Value", the "Copy old value(s)" radio button is selected. Below these, the "Old --> New:" list contains the instruction "8 thru 9 --> SYSMIS". The "Add" button is highlighted with an orange arrow. At the bottom right, there are "Cancel" and "Continue" buttons.

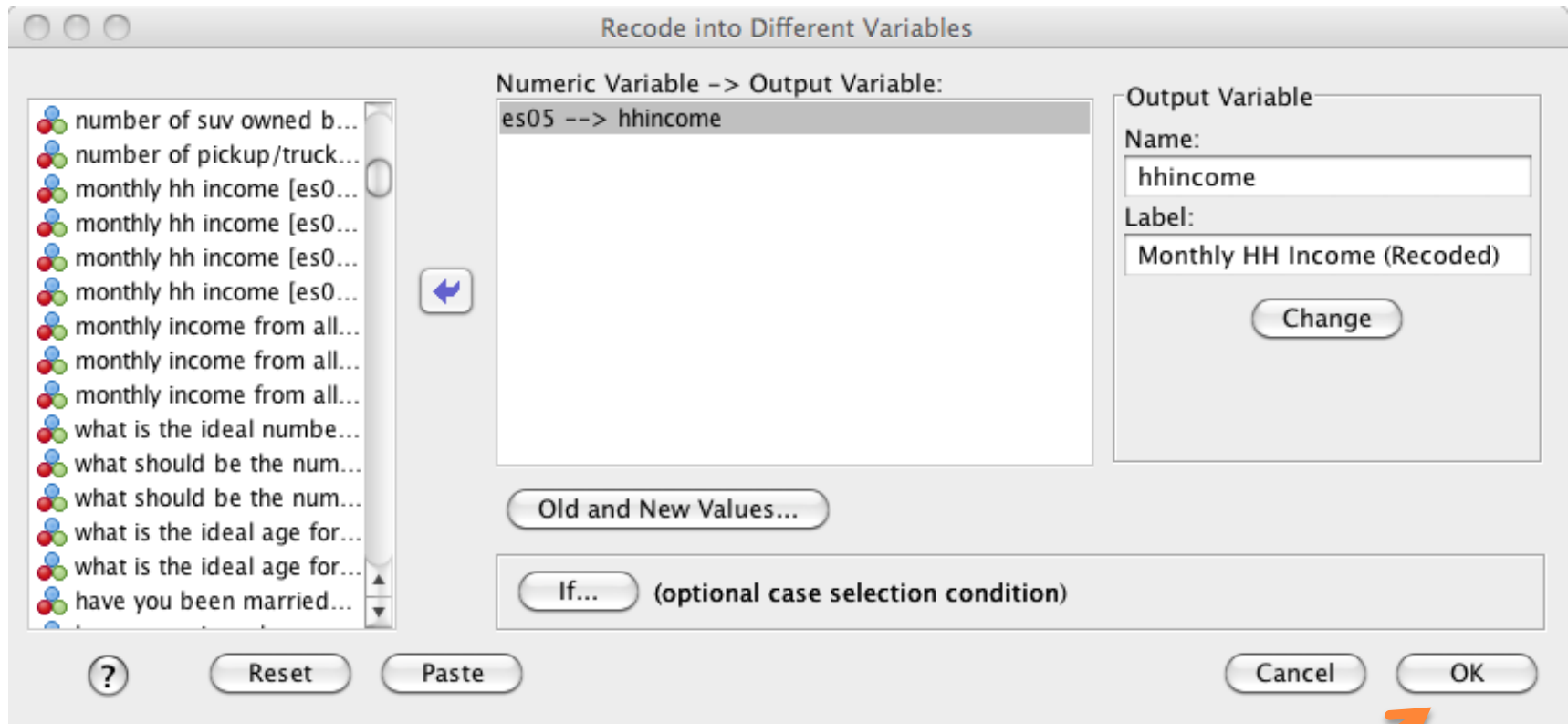


The dialog box is the same as the previous one, but now the "Old --> New:" list contains two instructions: "8 thru 9 --> SYSMIS" and "ELSE --> Copy". The entire list area is circled in orange. The "Add" button is no longer highlighted. The "Continue" button is highlighted with an orange arrow.

Then click the Add button to enter this recoding instruction into the list of recoding instructions.

Now that the list of instructions is complete, we can click Continue.

# RECODING IN SPSS



We are now ready to tell SPSS to perform the recoding. Do this by clicking OK.

# EXAMPLE:

## FREQUENCIES OF RECODED INCOME DATA

total monthly income of all hh members

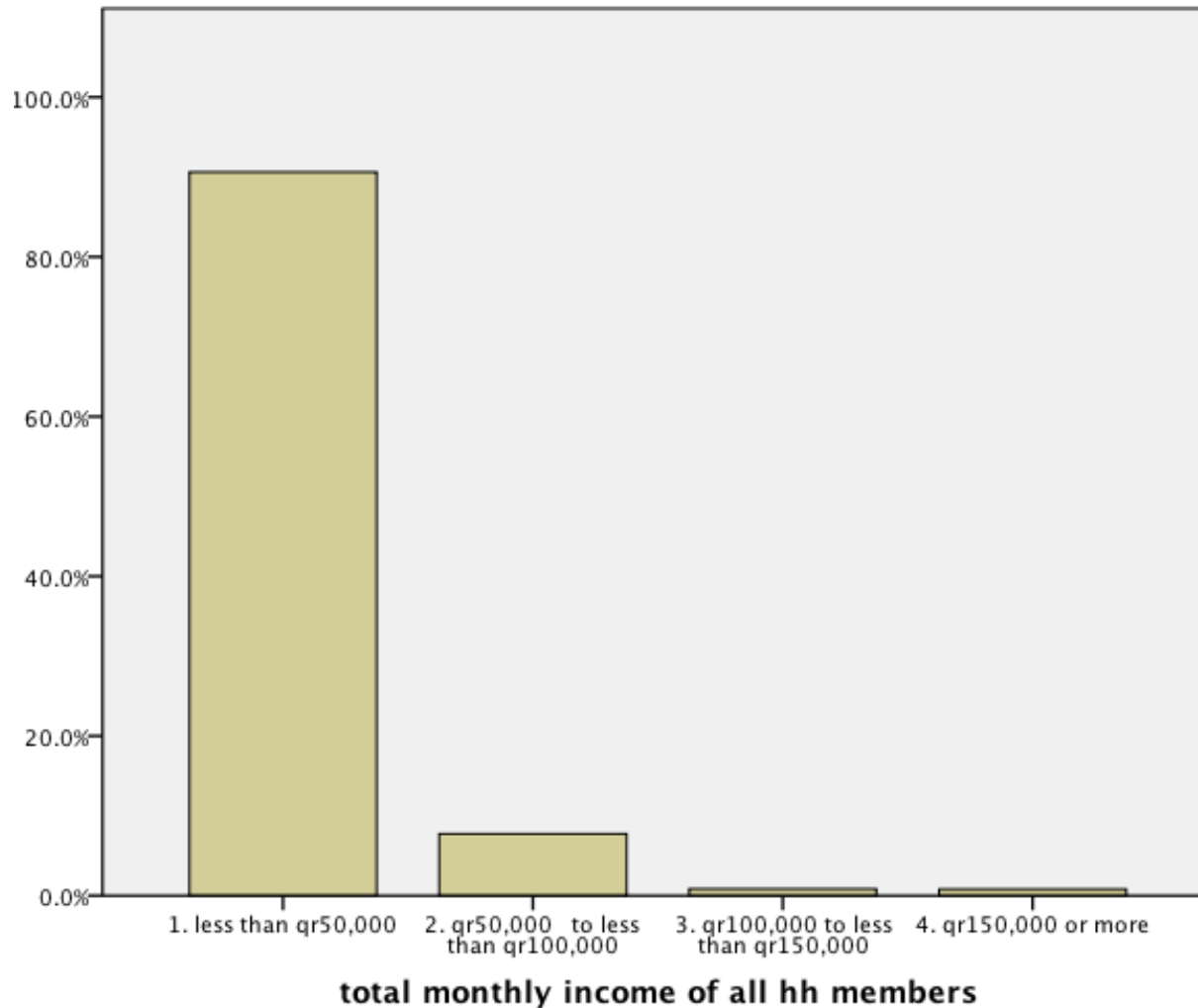
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1. less than qr50,000	1231	57.6	90.6	90.6
	2. qr50,000 to less than qr100,000	105	4.9	7.7	98.3
	3. qr100,000 to less than qr150,000	11	.5	.8	99.2
	4. qr150,000 or more	11	.5	.8	100.0
	Total	1359	63.5	100.0	
Missing	System	780	36.5		
Total		2139	100.0		

*Note: If you wish to use the old labels for the values of the variable or create new labels of your own for these values, you will need to do that using the menu option “Define Variable Properties” under the Data menu. Otherwise, the recoded variable’s values will appear simply as the actual numerical values.*



# EXAMPLE 1:

## GRAPH OF RECODED INCOME DATA





# WHY WORRY ABOUT THROWING OUT MISSING DATA?

- May Lose Information
- Lose Statistical Power *(leap ahead to hypothesis testing)*
  - Power of a statistical test = the probability of drawing a “false negative” inference, i.e., concluding that a relationship does not exist when it truly does exist.
  - When our tests have more statistical power, we can more easily detect relationships, even when those relationships are small.
  - Best way to increase power is to collect and use more observations (a larger sample).
- Risk a Loss of Representativeness
- Strategies for Minimizing the Loss of Missing Data  
*(more on Day 2)*





# Consolidating or Simplifying Answers



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# CONSOLIDATING RESPONSES: SOME EXAMPLES

- Create a simpler scale.
- Reduce the number of categories to focus on a comparison of interest.
- Create a count of items or options selected in answer to question or set of questions.



# CREATE A SIMPLER SCALE

## Original Scale

- 1 = much better
- 2 = somewhat better
- 3 = about the same
- 4 = somewhat worse
- 5 = much worse

## New (Recoded) Scale

- 1 = better
- 2 = about the same
- 3 = worse

QOL06. Would you say your own personal financial situation is now much better, somewhat better, about the same, somewhat worse or much worse now compared to two years ago?



# REDUCE THE NUMBER OF CATEGORIES TO FOCUS ON A COMPARISON OF INTEREST

## Original Categories

- 1 = less than one year
- 2 = one to two years
- 3 = three to five years
- 4 = six to ten years
- 5 = eleven to nineteen years
- 6 = twenty years or more, but not all of my life
- 7 = all my life

## New (Recoded) Categories

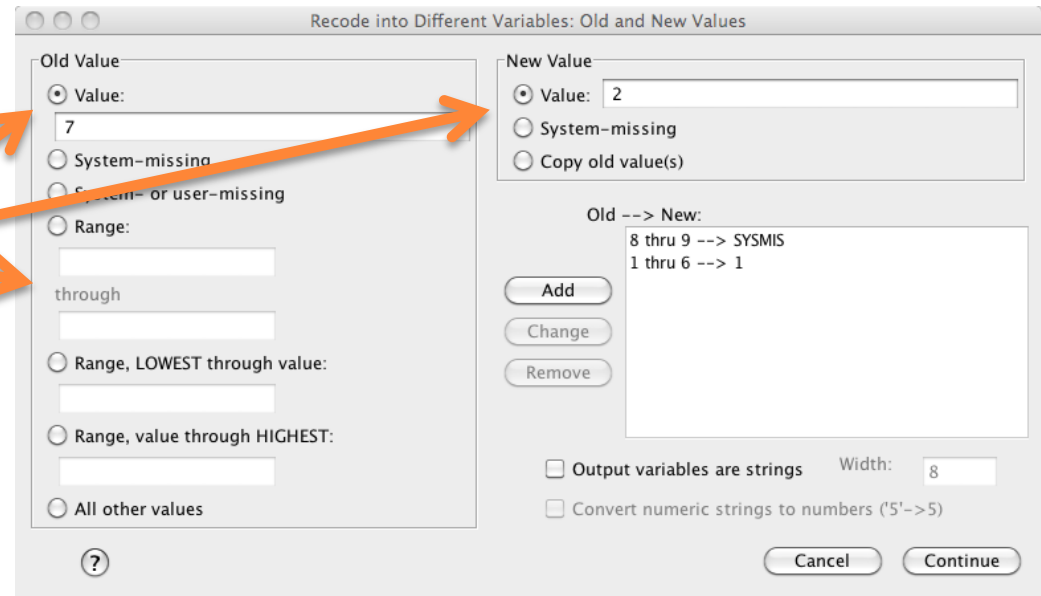
- 1 = less than my entire life
- 2 = all my life

DEM01. How long have you lived in Qatar?



# RECODING IN SPSS

Reminder: Enter range of old values, 1 through 6, to take new value of 1. Enter old value of 7 to take new value of 2. And assign 8 through 9 to missing values.



The dialog box is titled "Recode into Different Variables: Old and New Values". It has two main sections: "Old Value" and "New Value".

**Old Value:**

- ☒ Value: 7
- ☐ System-missing
- ☐ System- or user-missing
- ☐ Range: [ ] through [ ]
- ☐ Range, LOWEST through value: [ ]
- ☐ Range, value through HIGHEST: [ ]
- ☐ All other values

**New Value:**

- ☒ Value: 2
- ☐ System-missing
- ☐ Copy old value(s)

**Old --> New:**

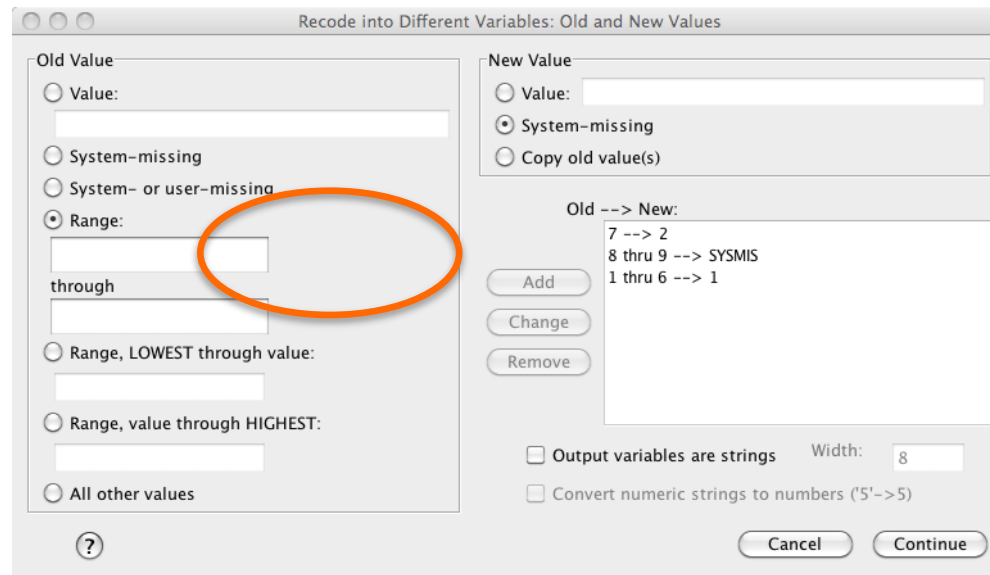
- 8 thru 9 --> SYSMIS
- 1 thru 6 --> 1

Buttons: Add, Change, Remove

☐ Output variables are strings Width: 8

☐ Convert numeric strings to numbers ('5'-->5)

Buttons: Cancel, Continue



The dialog box is titled "Recode into Different Variables: Old and New Values". It has two main sections: "Old Value" and "New Value".

**Old Value:**

- ☐ Value: [ ]
- ☐ System-missing
- ☐ System- or user-missing
- ☒ Range: [ ] through [ ]
- ☐ Range, LOWEST through value: [ ]
- ☐ Range, value through HIGHEST: [ ]
- ☐ All other values

**New Value:**

- ☐ Value: [ ]
- ☒ System-missing
- ☐ Copy old value(s)

**Old --> New:**

- 7 --> 2
- 8 thru 9 --> SYSMIS
- 1 thru 6 --> 1

Buttons: Add, Change, Remove

☐ Output variables are strings Width: 8

☐ Convert numeric strings to numbers ('5'-->5)

Buttons: Cancel, Continue

# THOUGHT EXERCISE: RECODE FREQUENCY OF WORSHIP ATTENDANCE INTO TWO CATEGORIES

Valid Values	1.00	1. daily	709	33.1%
	2.00	2. more than once a week	172	8.0%
	3.00	3. once a week	241	11.3%
	4.00	4. once or twice a month	77	3.6%
	5.00	5. a few times a year	297	13.9%
	6.00	6. seldom	226	10.6%
	7.00	7. never	360	16.8%
	8.00	8. don't know	6	.3%
	9.00	9. refused	12	.6%
Missing Values	System		39	1.8%

Let's say we wish to divide people into two groups based on how often they attend places of worship.

What do we need to do to recode the original variable?

What are the options for dividing people into two groups based on their original answers?

Think about how these different options alter the comparison we are making (what it means).

Original Question (DEM08): People differ in how often they go to a place of worship. Sometimes people don't go as often as they might like because they are sick, working, or have other obligations. What about you? How often do you go to a place of worship – would you say daily, more than once a week, once a week, once or twice a month, a few times a year, seldom, or never?



# CREATE A COUNT

## Original Answers

Type of benefit	R1	R2	R3
Housing allowance	1	1	0
Food allowance	0	0	0
Health insurance	1	0	0
Accident insurance	1	0	0
Paid holidays	1	1	0
Paid sick leave	1	1	0
Free water/electricity	0	0	1
Transport to/from work	0	0	1
Access to training	1	0	0
Paid visits to home	0	0	0

## New Count Variable

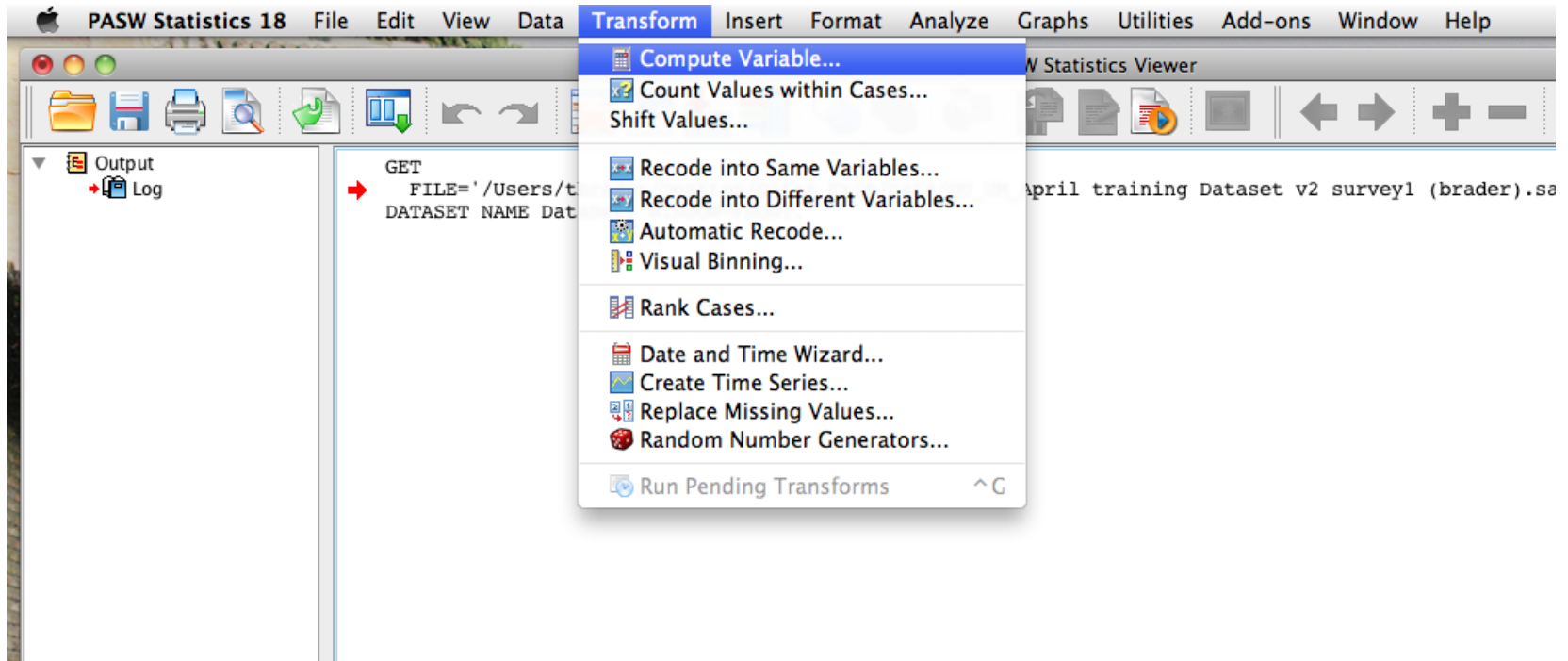
	R1	R2	R3
Count of Benefits	6	3	2

EMP14. Besides money paid for work, some people earn other benefits. Which of the following non-wage benefits do you receive from your current (primary) job?





# RECODING IN SPSS



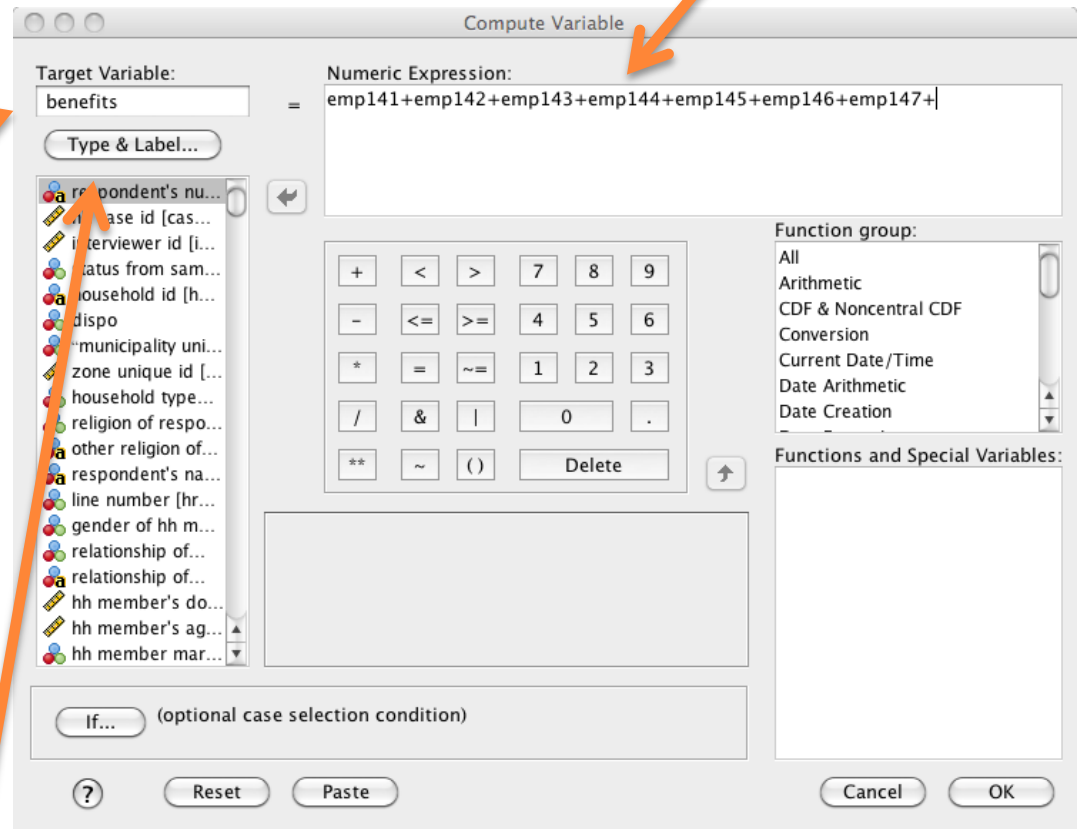
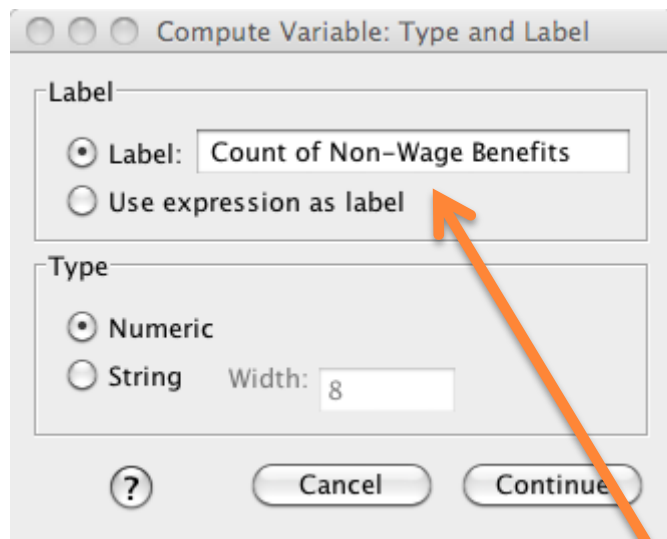
Select Menu Option: “Compute Variable”



# RECODING IN SPSS

Enter mathematical formula into Numeric Expression box. In this case, summing the variables emp141 through emp1410.

Enter name of new variable into Target Variable box.



If you want to give the new variable an explanatory label, click Type and Label button. In the new dialog box, enter the name of the label and then click the Continue button.





# Reordering the Values



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# CHANGE THE VALUES TO AN ORDER THAT IS MORE INTUITIVE OR FITS THE HYPOTHESIS

## Original Variable

- 1 = Excellent
- 2 = Good
- 3 = Fair
- 4 = Poor

## New (Recoded) Variable

- 1 = Poor
- 2 = Fair
- 3 = Good
- 4 = Excellent

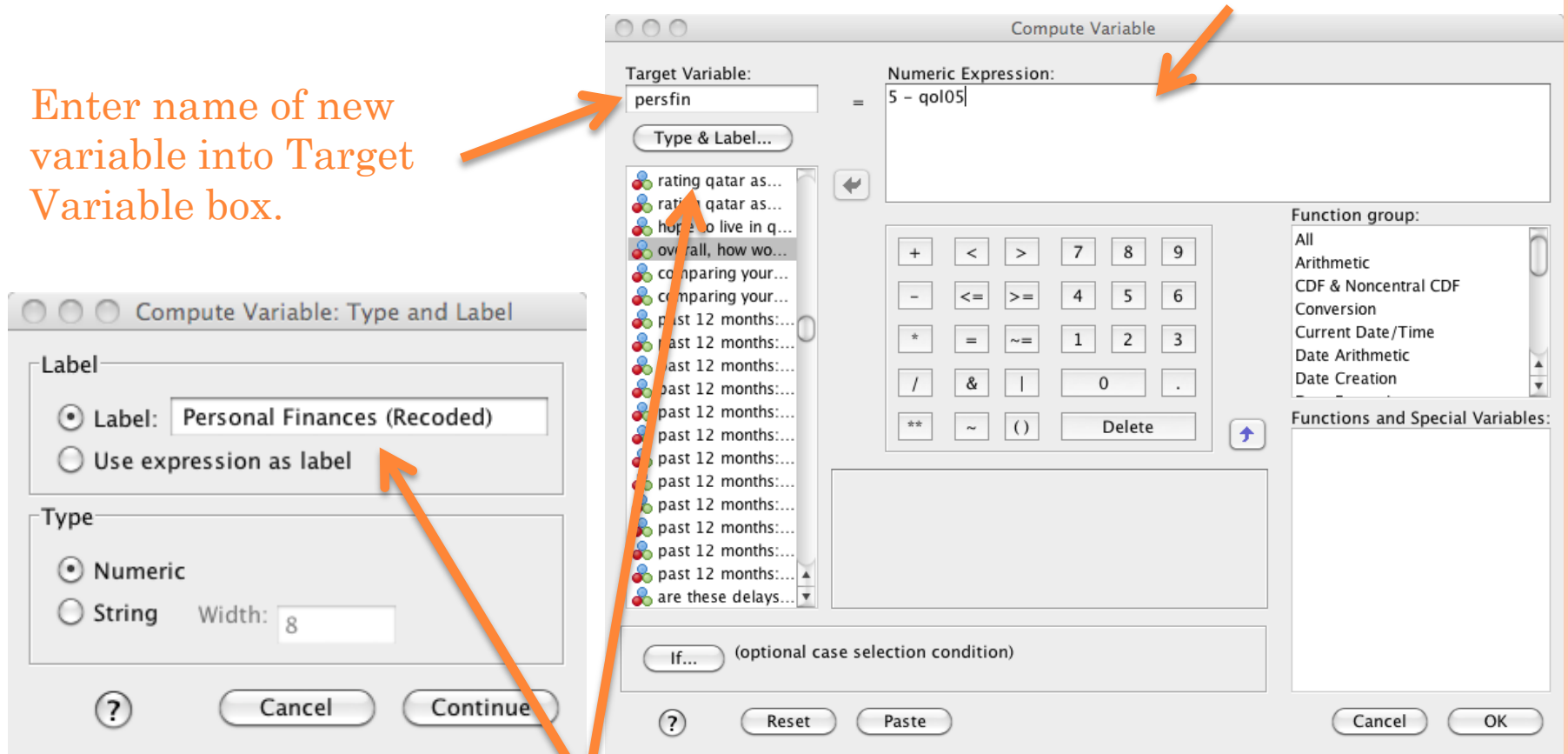
QOL05. Overall, how would you rate your own personal financial situation these days? Would you say you are in excellent shape, good shape, fair shape, or poor shape financially?



# RECODING IN SPSS

Enter mathematical formula into Numeric Expression box. In this case, subtract the original variable (qol05) from the number 5.

Enter name of new variable into Target Variable box.



If you want to give the new variable an explanatory label, click Type and Label button. In the new dialog box, enter the name of the label and then click the Continue button.



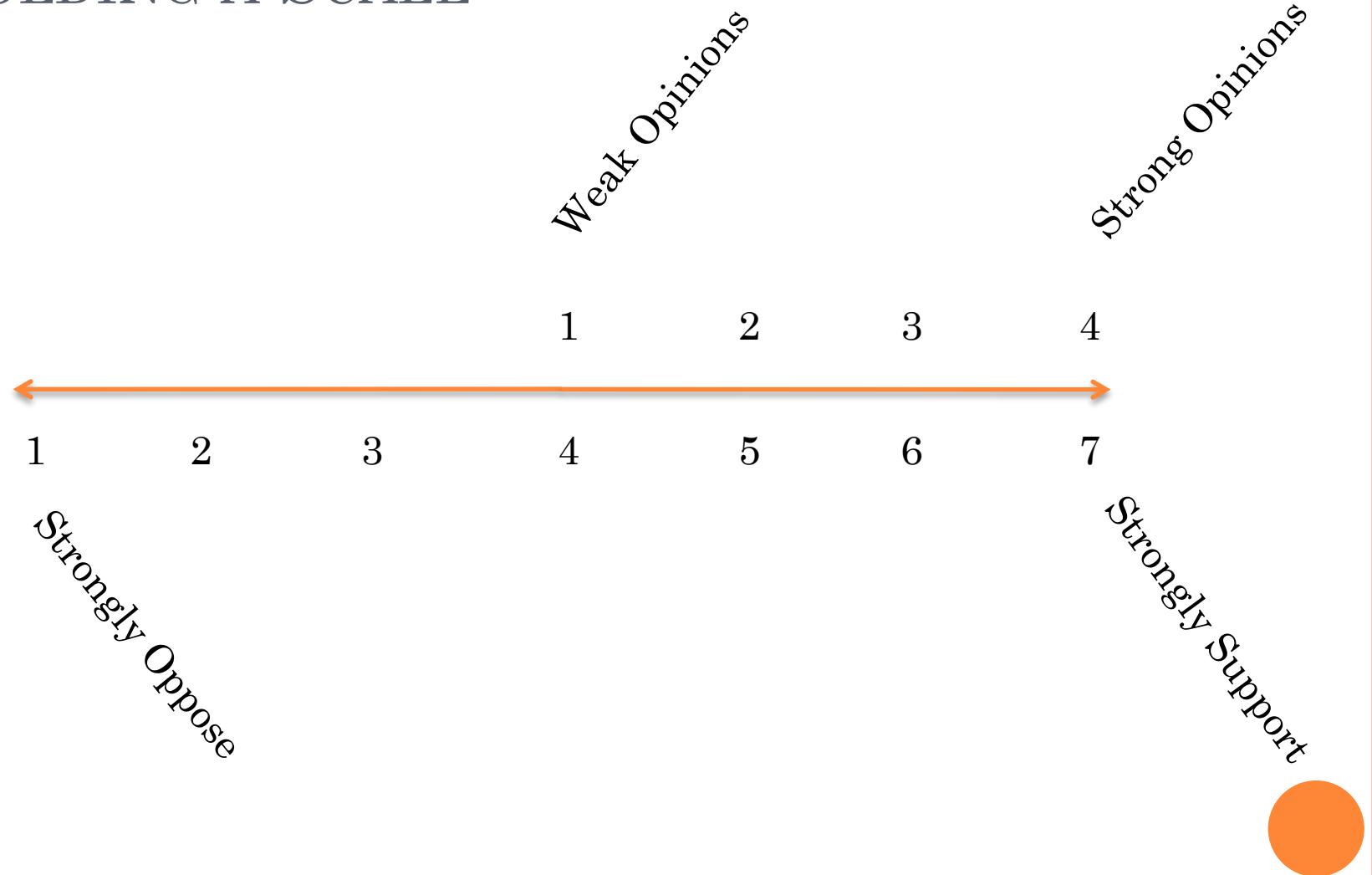


# Folding



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# FOLDING A SCALE



# EXAMPLE OF FOLDING A SCALE

## Original Scale

- 1 = much better
- 2 = somewhat better
- 3 = about the same
- 4 = somewhat worse
- 5 = much worse

## Recoded (Folded) Scale

- 1 = see no change
- 2 = see some change
- 3 = see a lot of change

QOL06. Would you say your own personal financial situation is now much better, somewhat better, about the same, somewhat worse or much worse now compared to two years ago?







# Rescaling to a Common Range



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# RESCALING TO A COMMON RANGE

## ○ Why?

- Allow for comparisons of size of relationship across variables and models
- Makes it easy to have similar number of decimal places when presenting statistical results in tables

## ○ Preferred Method

- Set variables to have same value for minimum and maximum (e.g., set all variables to range between values of 0 and 1)
- Alternative to *standardization*, which has been used for same purpose by many scholars and is more problematic for making comparisons about the size of relationships
  - Standardized scores = transform the original data to indicate the number of standard deviations an observation is above or below the mean.
  - Standard deviations are not an intuitive unit
  - Estimated relationships will change with the variance of the sample, even if the true underlying relationship between X and Y is unchanged



## EXAMPLE: RESCALING TO COMMON RANGE

### Original Variable

- 1 = much better
- 2 = somewhat better
- 3 = about the same
- 4 = somewhat worse
- 5 = much worse

### New (Recoded) Variable

- 0 = much better
- .25 = somewhat better
- .50 = about the same
- .75 = somewhat worse
- 1 = much worse

QOL06. Would you say your own personal financial situation is now much better, somewhat better, about the same, somewhat worse or much worse now compared to two years ago?





# Interactive Exercise

Hands-On Practice: Recoding with SPSS



معهد البحوث الاجتماعية والاقتصادية المسحية  
Social & Economic Survey Research Institute

# INTERACTIVE EXERCISE: RECODE QATARI BELIEFS ABOUT THE IDEAL NUMBER OF CHILDREN

Valid Values	.00		11	.5%
	2.00		30	1.4%
	3.00		26	1.2%
	4.00		213	10.0%
	5.00		96	4.5%
	6.00		154	7.2%
	7.00		23	1.1%
	8.00		31	1.5%
	9.00		4	.2%
	10.00		26	1.2%
	11.00		1	.0%
	12.00		7	.3%
	15.00		1	.1%
	20.00		1	.0%
	44.00	44. god's will	34	1.6%
	50.00		1	.0%
	77.00		1	.1%
	98.00	98. don't know	4	.2%
	99.00	99. refused	0	.0%
Missing Values	System		1475	69.0%

Say you want to examine Qatari preferences for the ideal number of children. The values of the original variable range from 0 to 99.

Recode this variable so that it will be useful for your analyses.

You will need to determine which values refer to an actual number of children preferred by the respondent and which do not. Hint: Looking back the questionnaire and closely related questions can help a researcher interpret answers to a question.

Original Question (GEN01a): In your opinion what is the ideal number of children to have?