Analyzing Missing Data

Introduction

Problems

Using Scripts

Missing data and data analysis

- Missing data is a problem in multivariate data because a case will be excluded from the analysis if it is missing data for any variable included in the analysis.
- If our sample is large, we may be able to allow cases to be excluded.
- If our sample is small, we will try to use a substitution method so that we can retain enough cases to have sufficient power to detect effects.
- In either case, we need to make certain that we understand the potential impact that missing data may have on our analysis.

Tools for evaluating missing data

- SPSS has a specific package for evaluating missing data, but it is included under the UT license.
- In place of this package, we will first examine missing data using SPSS statistics and procedures.
- After studying the standard SPSS procedures that we can use to examine missing data, we will use an SPSS script that will produce the output needed for missing data analysis without requiring us to issue all of the SPSS commands individually.

Key issues in missing data analysis

- We will focus on two key issues for evaluating missing data:
  - The number or proportion of cases missing for each variable
  - Whether or not cases with missing data had statistically significant differences from cases with valid data for the other variables included in the analysis.

- Further analysis may be required depending on the problems identified in these analyses.

Benchmark for evaluating missing data

- The text suggests that, in general, if no more than 5% of the cases in the sample were missing data for a variable and if the pattern of missing data is random, missing data is not especially problematic for the analysis.

Our strategy for evaluating missing data

- The criteria lead us to a two stage strategy for evaluating the pattern of missing data.
- First, we will identify variables that are missing data for more than 5% of the cases in the sample.
  - If no variables are missing more than 5% of the cases, we will assume that there is not a problematic pattern.
- Second, for each variable that is missing data for more than 5% of the cases, we create a dichotomous missing/valid variable that is coded 0 for cases missing data and 1 for cases with valid data and test for statistically significant differences between the valid and missing groups for all other variables in the analysis.
  - If significant differences are found, we will attach a caution to our analysis with a recommendation for further study of the problems.

Testing for differences in missing/valid groups

- If the variable to be tested is metric, we use a t-test to compare the missing and valid groups.
- If the variable is nonmetric, we use a chi-square test of independence to compare the missing and valid groups.
- In all tests, we will use the level of significance stated in the problem for evaluating missing data and assumptions.

Example

- For example, suppose we are testing the relationship between the independent variables sex and age, and the dependent variable respondent's income. A frequency distribution on income indicates that 37.8% of the cases did not answer the question, so we create a dichotomous variable that is coded 0 for missing income and 1 for valid income.
- Since sex is a nonmetric variable, we do a chi-square test of independence with the missing/valid income as the independent variable and sex as the dependent variable to see if there is a relationship.
- Since age is a metric variable, we do a t-test to see if the average age for subjects who answered the question is different than the average age for subjects who skipped the question.

Problem 1

In the dataset GSS2000R, is the following statement true, false, or an incorrect application of a statistic? Use a level of significance of 0.01 for evaluating missing data and assumptions.

In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

1. True
2. True with caution
3. False
4. Inappropriate application of a statistic

Checking level of measurement

GSS2000R, is the following statement true,
ect applicati level of
fo

Since we are pre-screening for a multiple regression problem, we should make sure we satisfy the level of measurement before proceeding.

"Total hours spent on the Internet" [netime] is interval, satisfying the metric level of measurement requirement for the dependent variable.

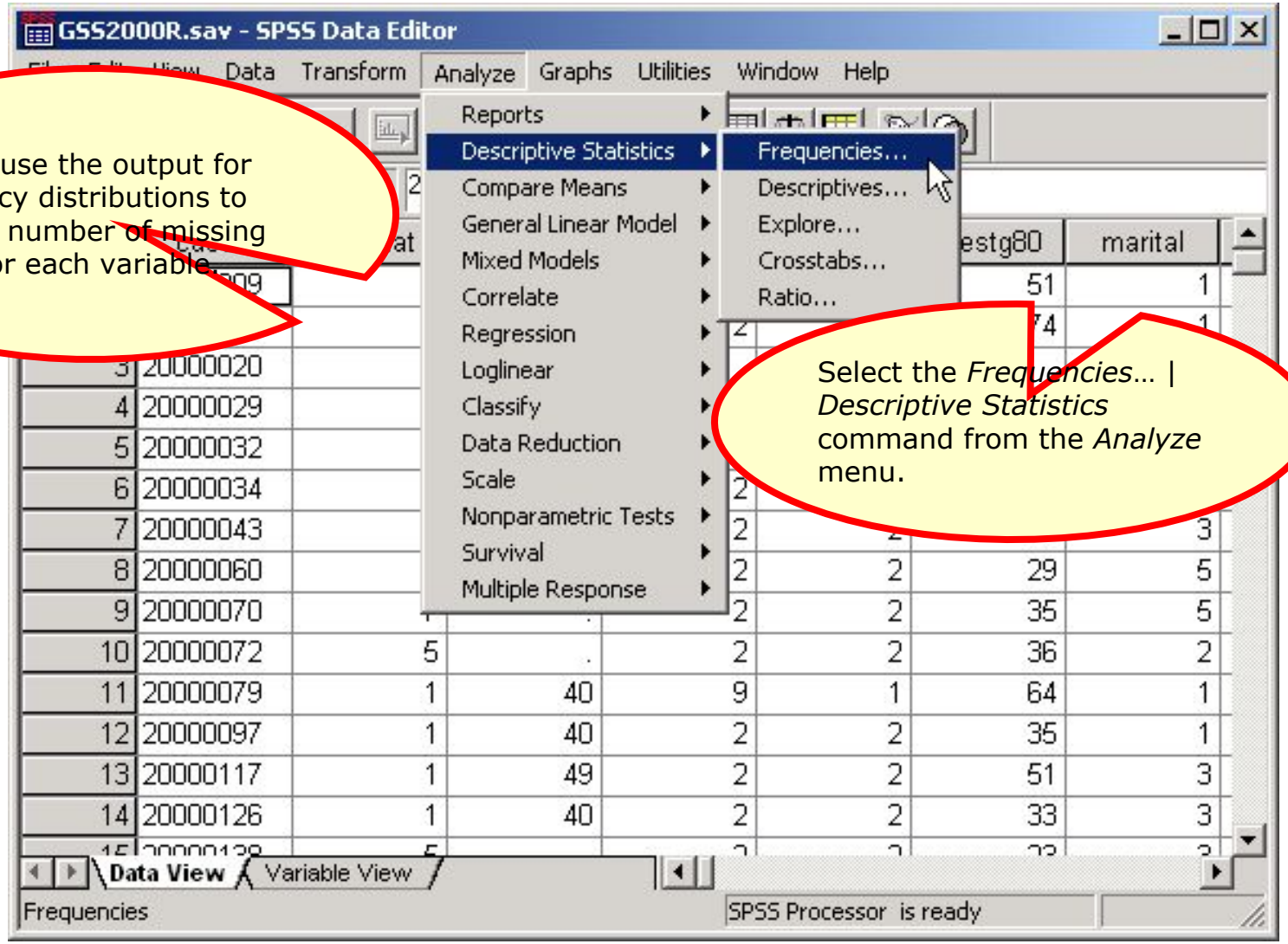
In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

"Age" [age] and "highest year of school completed" [educ] are interval, satisfying the metric or dichotomous level of measurement requirement for independent variables.

"Sex" [sex] is dichotomous, satisfying the metric or dichotomous level of measurement requirement for independent variables.

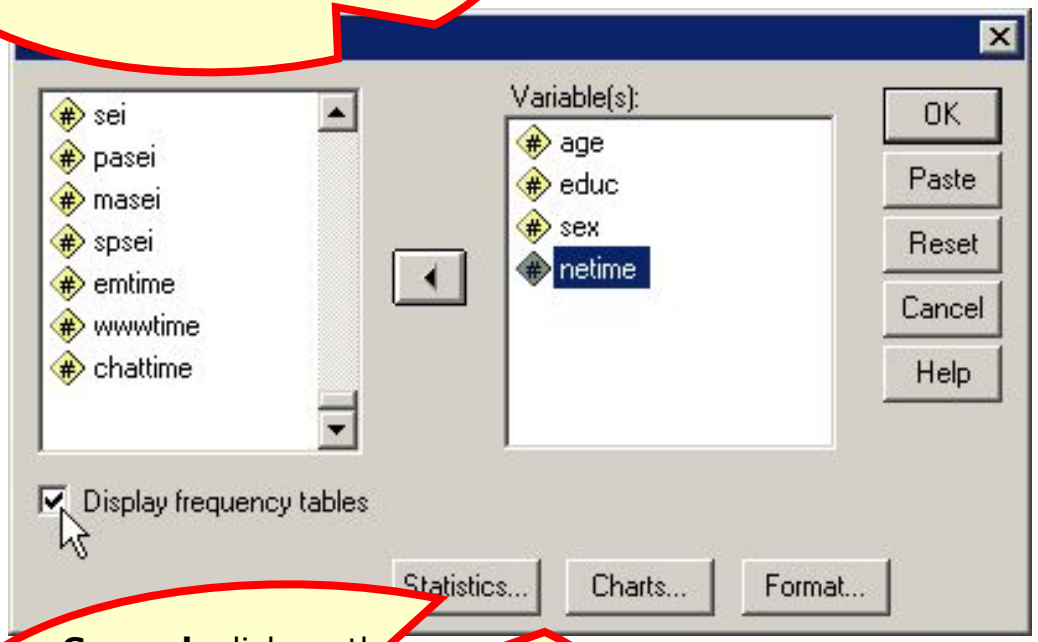
1.
2.
3. False
4. Inappropriate application of a statistic

Request frequency distributions



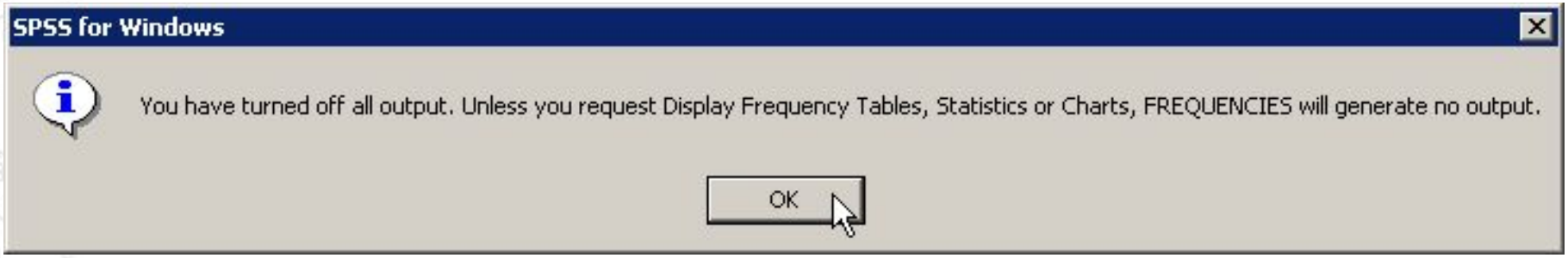
Completing specifications for frequencies - 1

First, move the four variables included in the problem statement to the list box for variables.



Second, click on the *Display frequency tables* check box to clear it, since all we want is the statistics for missing and valid cases.

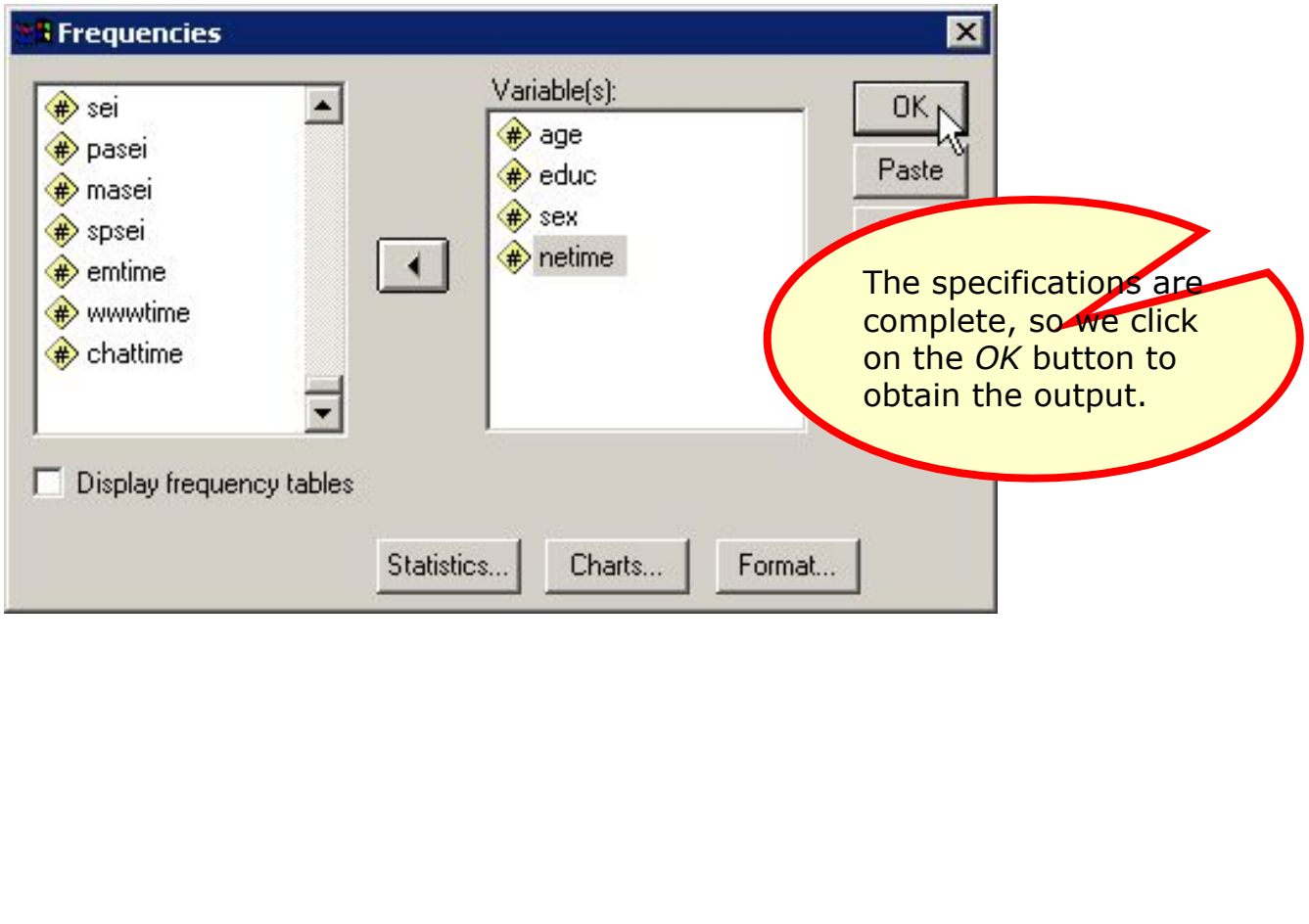
Completing specifications for frequencies - 2



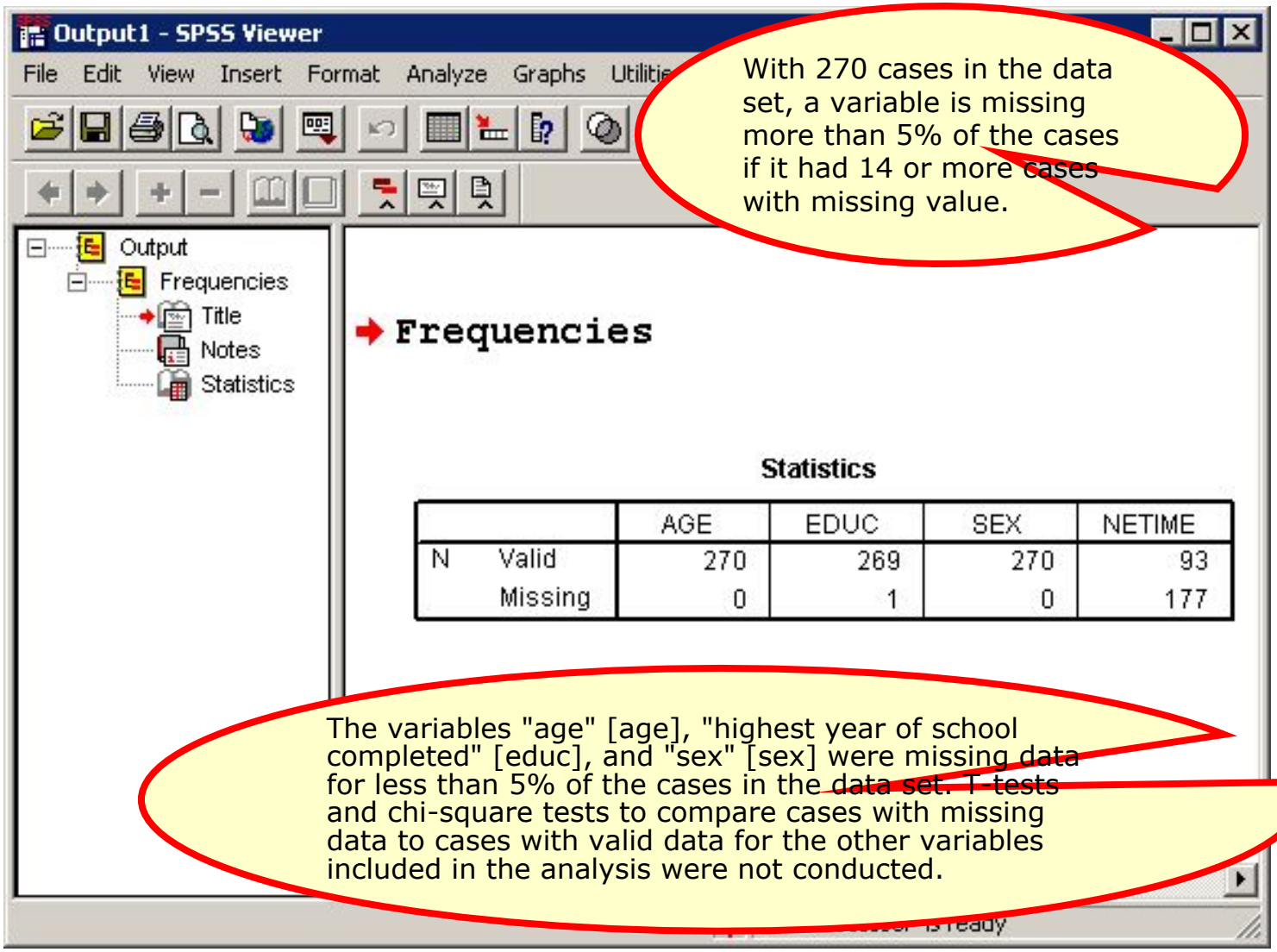
SPSS give us a warning message that we will not generate any output. However, it will produce the statistics for valid an missing data which is want we want.

Click on the *OK* button to close the warning.

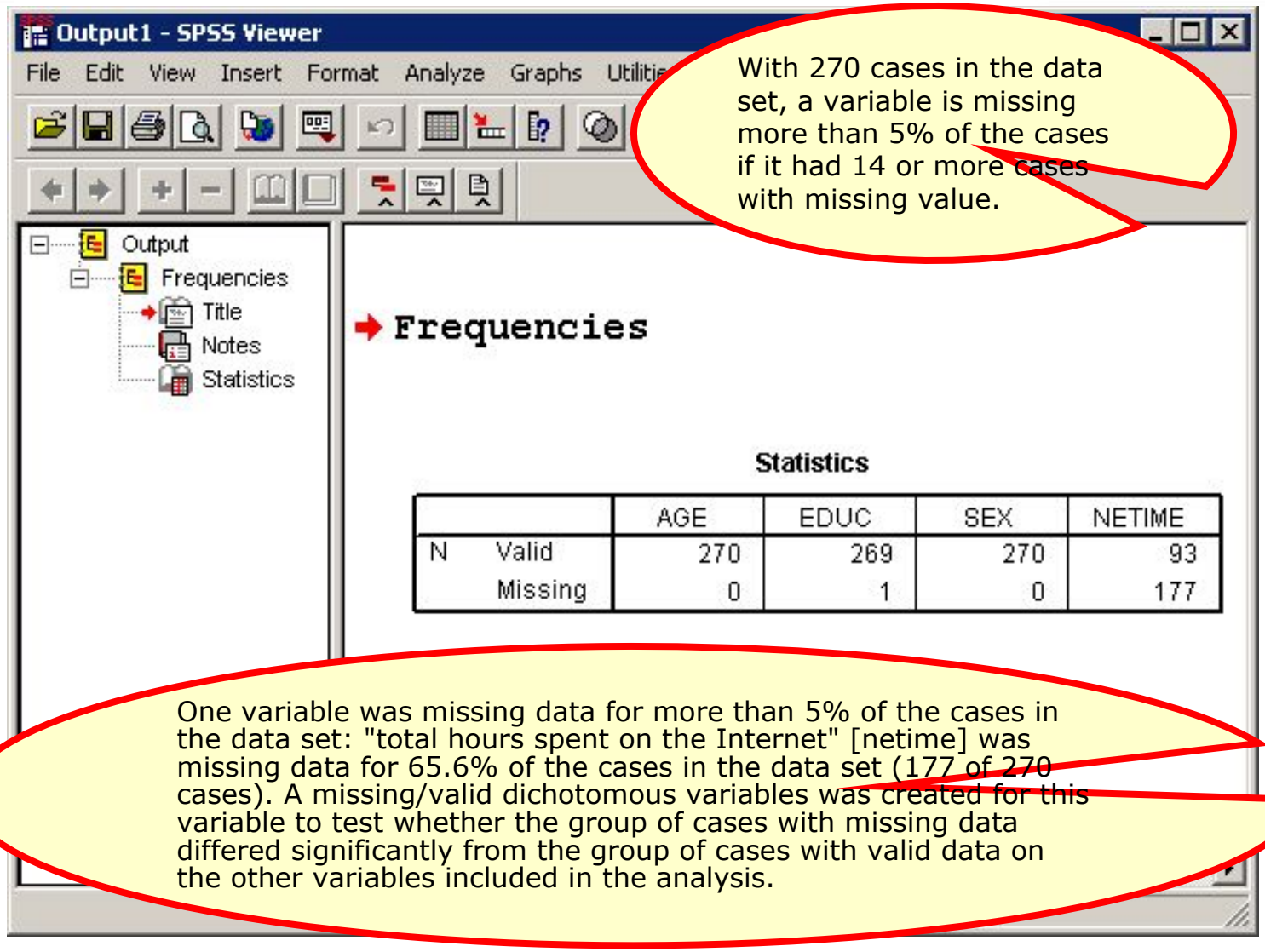
Completing specifications for frequencies - 3



Number of missing cases for each variable - 1



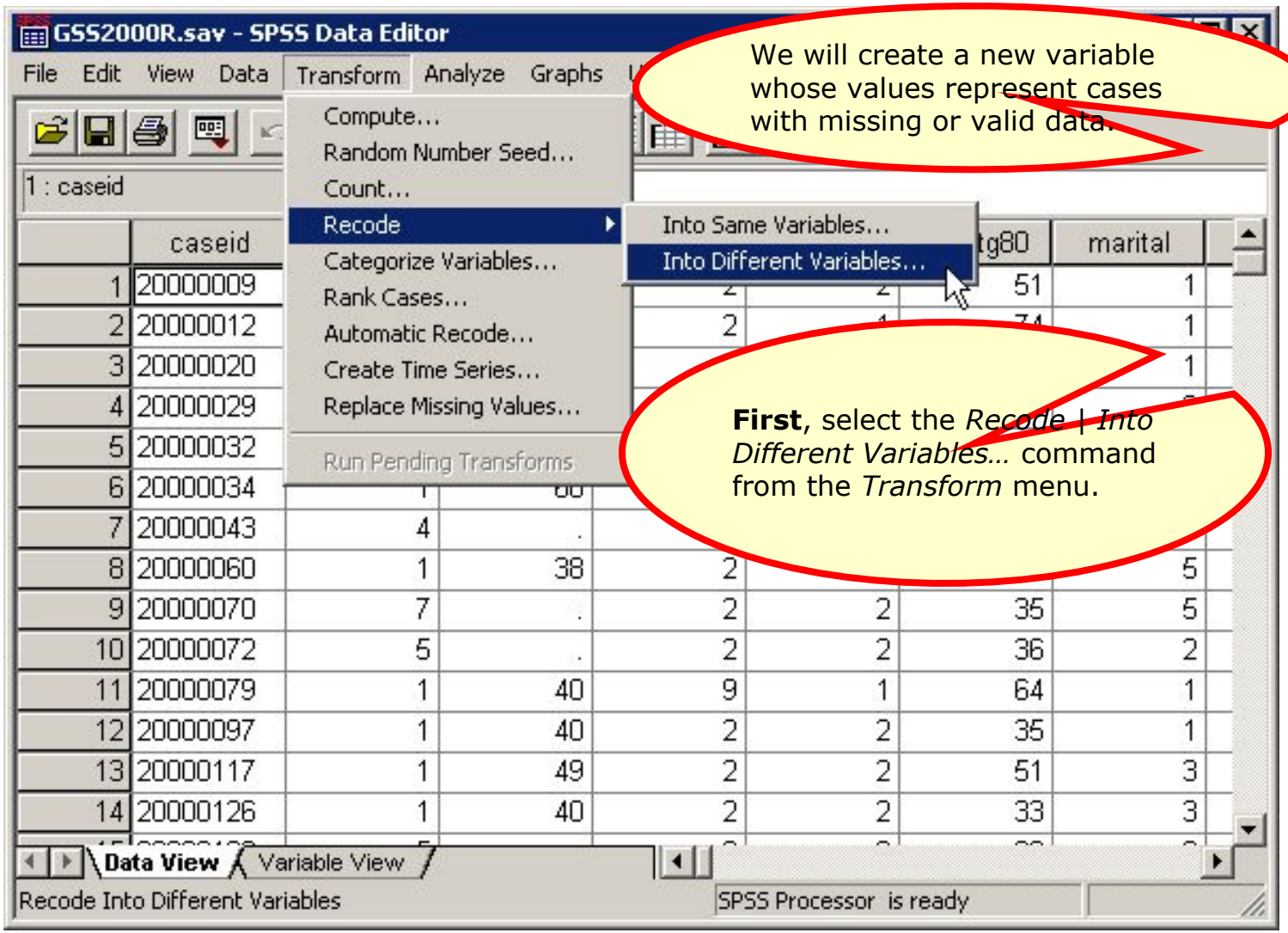
Number of missing cases for each variable - 2



With 270 cases in the data set, a variable is missing more than 5% of the cases if it had 14 or more cases with missing value.

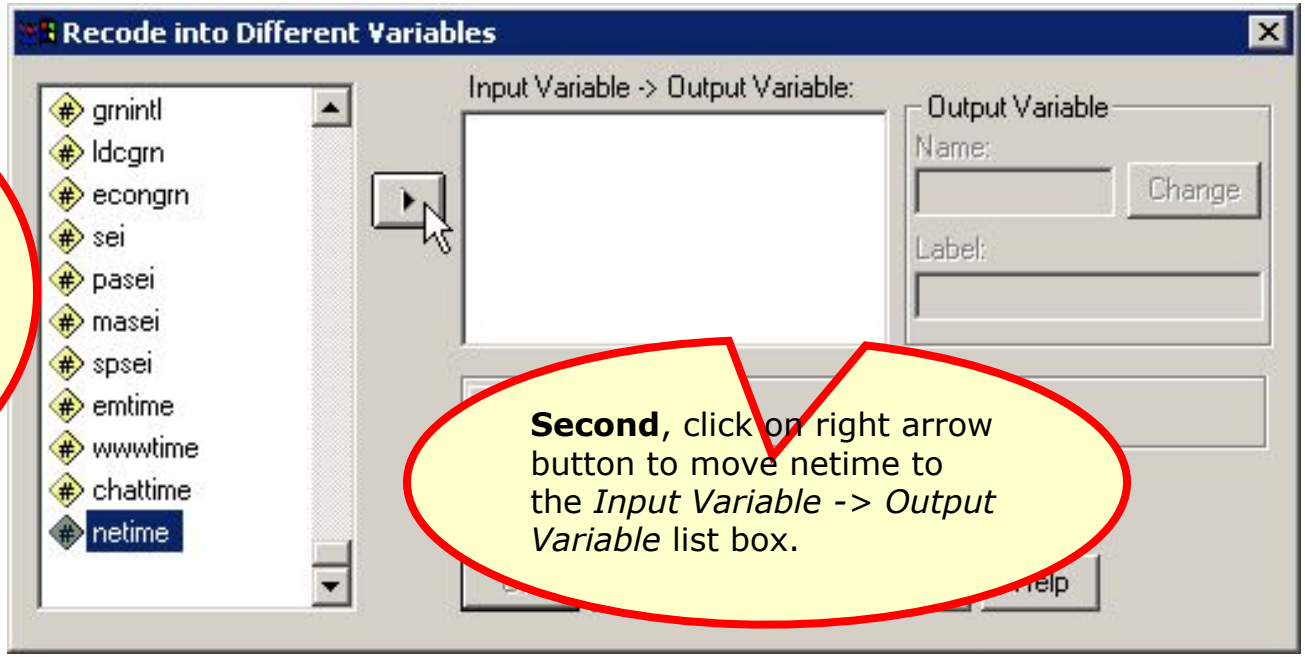
One variable was missing data for more than 5% of the cases in the data set: "total hours spent on the Internet" [netime] was missing data for 65.6% of the cases in the data set (177 of 270 cases). A missing/valid dichotomous variables was created for this variable to test whether the group of cases with missing data differed significantly from the group of cases with valid data on the other variables included in the analysis.

Creating the missing/valid variable - 1



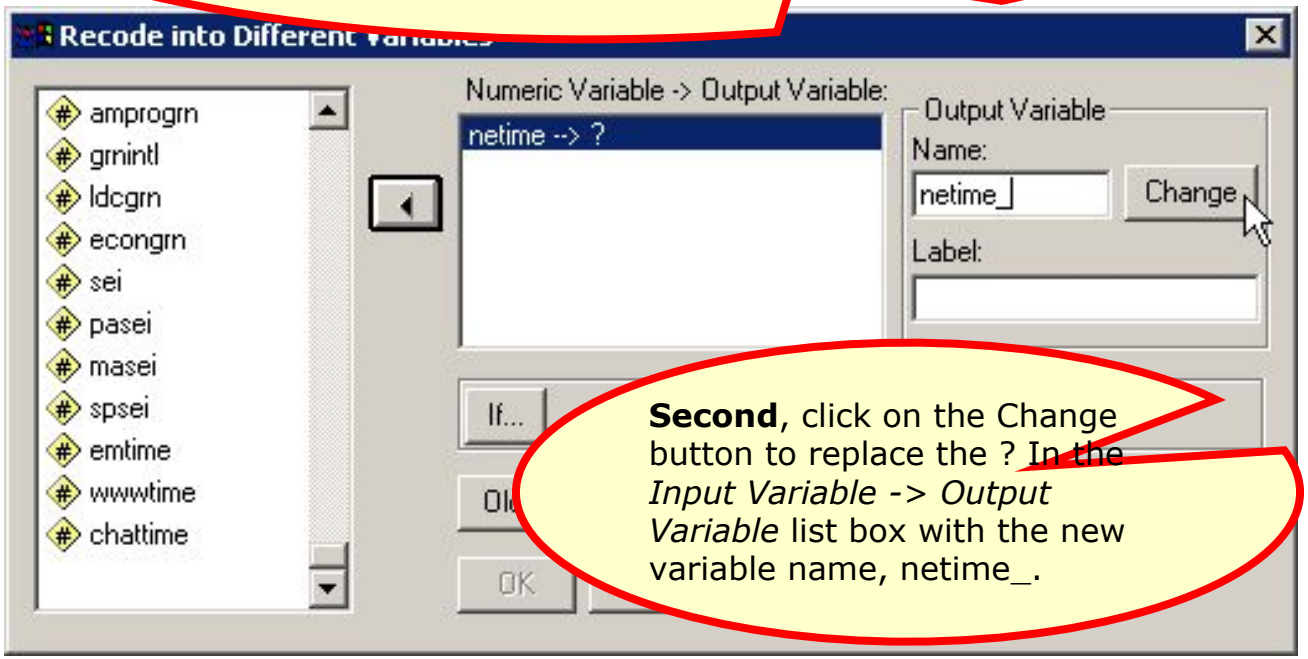
Creating the missing/valid variable - 2

First, highlight the variable netime, which is the variable which had more than 5% missing data, for which we want to create the missing/valid variable.



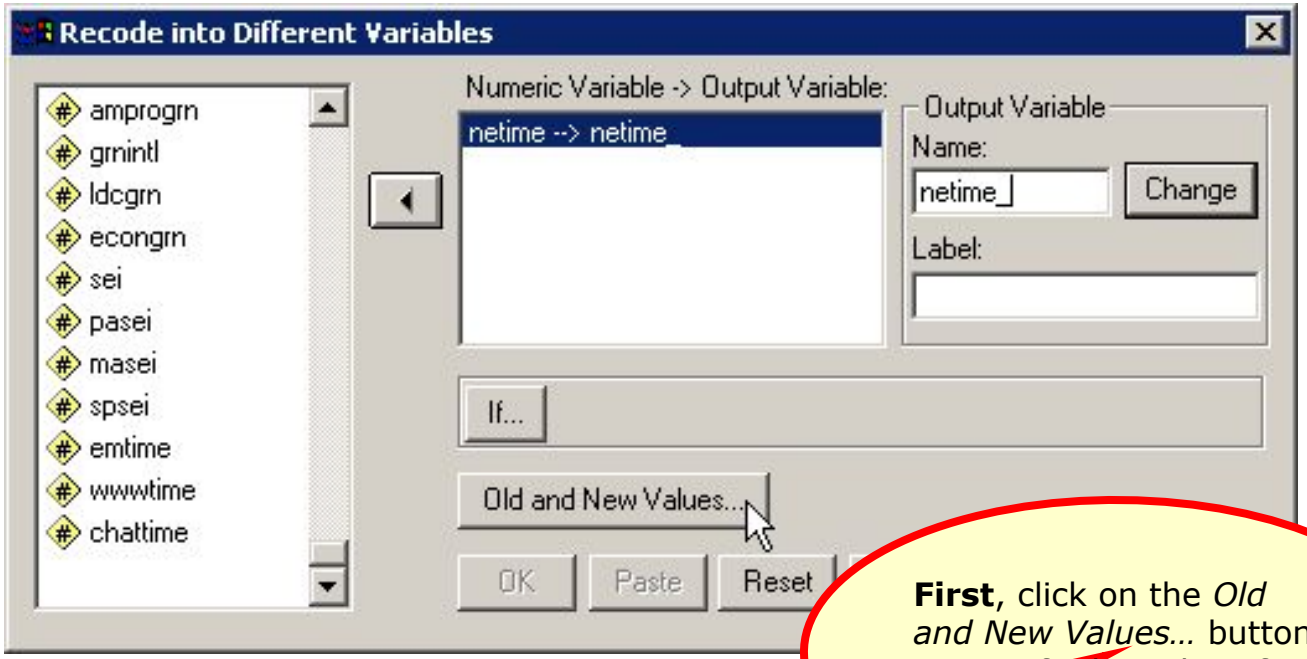
Creating the missing/valid variable - 3

First, type a name for the new variable into the Name: text box. I usually just add an underscore to the variable name if the original variable name is 7 letters or less. If the variable is 8 letters, I delete the last letter so that I do not exceed the SPSS requirement that a variable name be 8 characters or less.



Second, click on the Change button to replace the ? In the *Input Variable -> Output Variable* list box with the new variable name, netime_.

Creating the missing/valid variable - 4

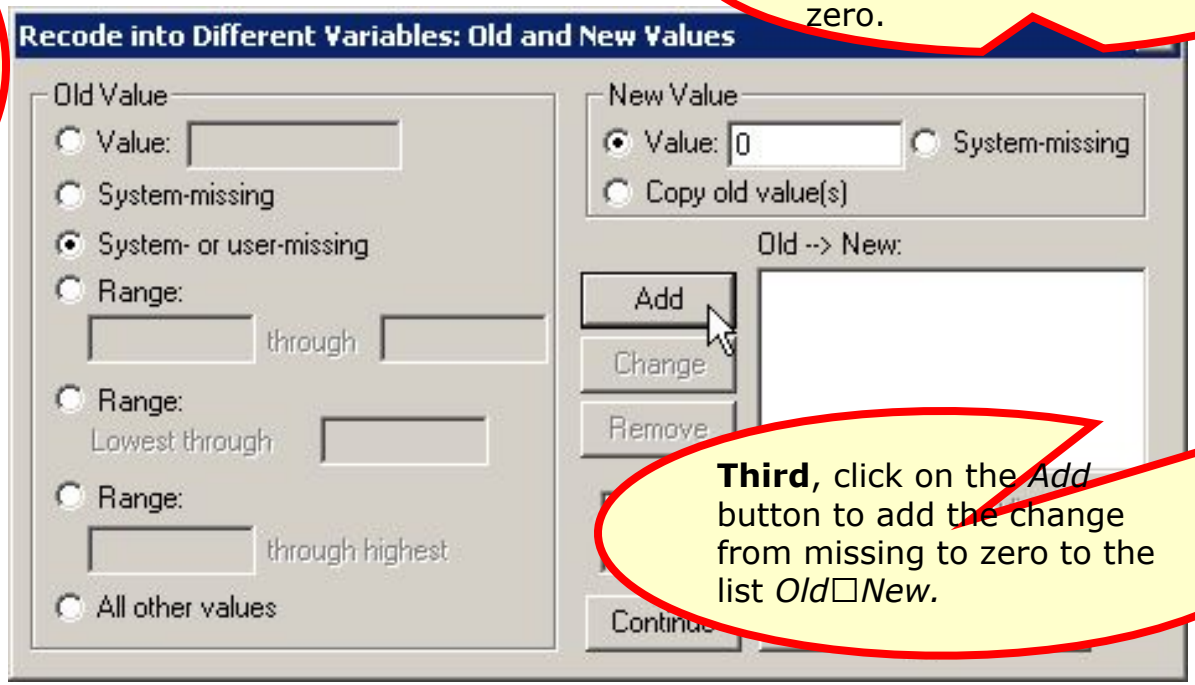


First, click on the *Old and New Values…* button to specify the values for the new variable.

Creating the missing/valid variable - 5

First, to create the code 0 for missing data, we mark the *System- or user-missing* option button on the *Old Value* panel.

Second, in the Value: text box in the New Value panel, we type a zero.

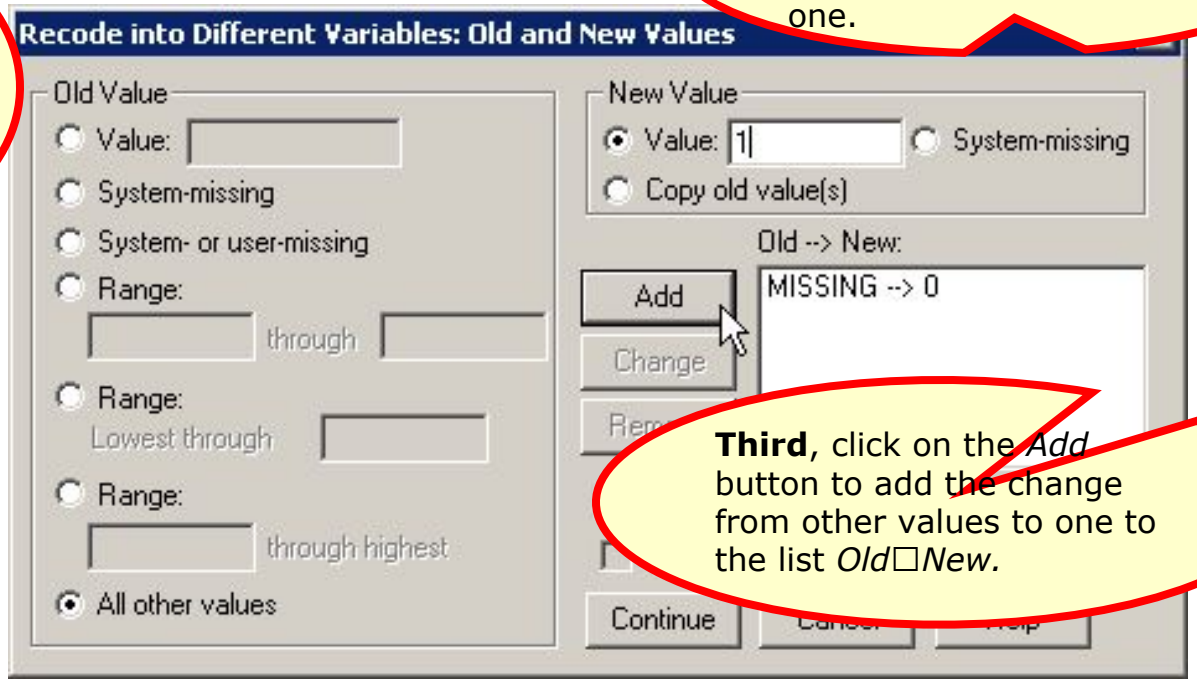


Third, click on the *Add* button to add the change from missing to zero to the list *Old*☐*New.*

Creating the missing/valid variable - 6

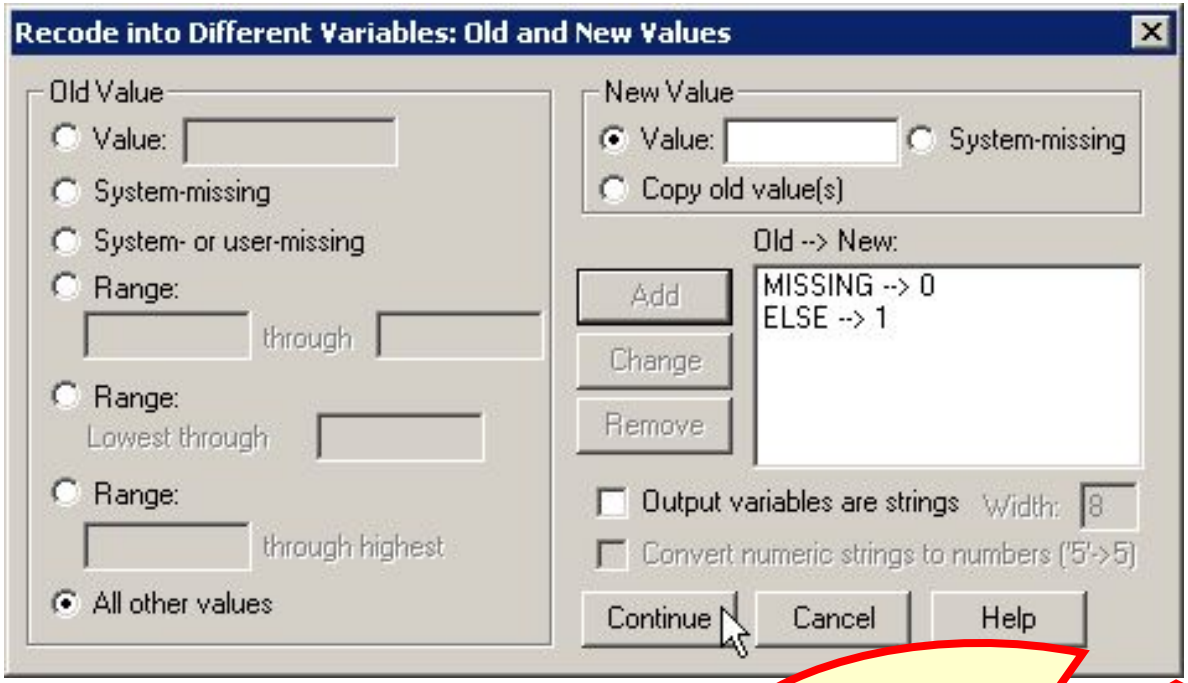
First, to create the code 1 for valid data, we mark the *All other values* option button on the *Old Value* panel.

Second, in the Value: text box in the New Value panel, we type a one.



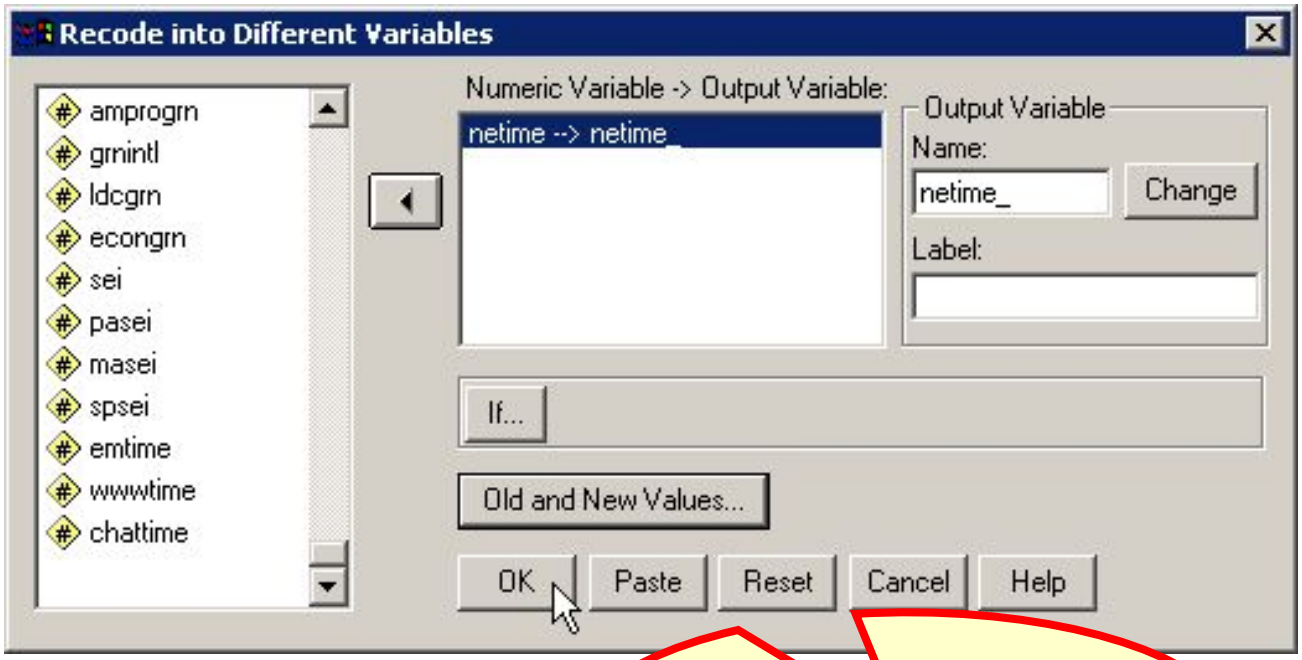
Third, click on the *Add* button to add the change from other values to one to the list *Old*☐*New.*

Creating the missing/valid variable - 7



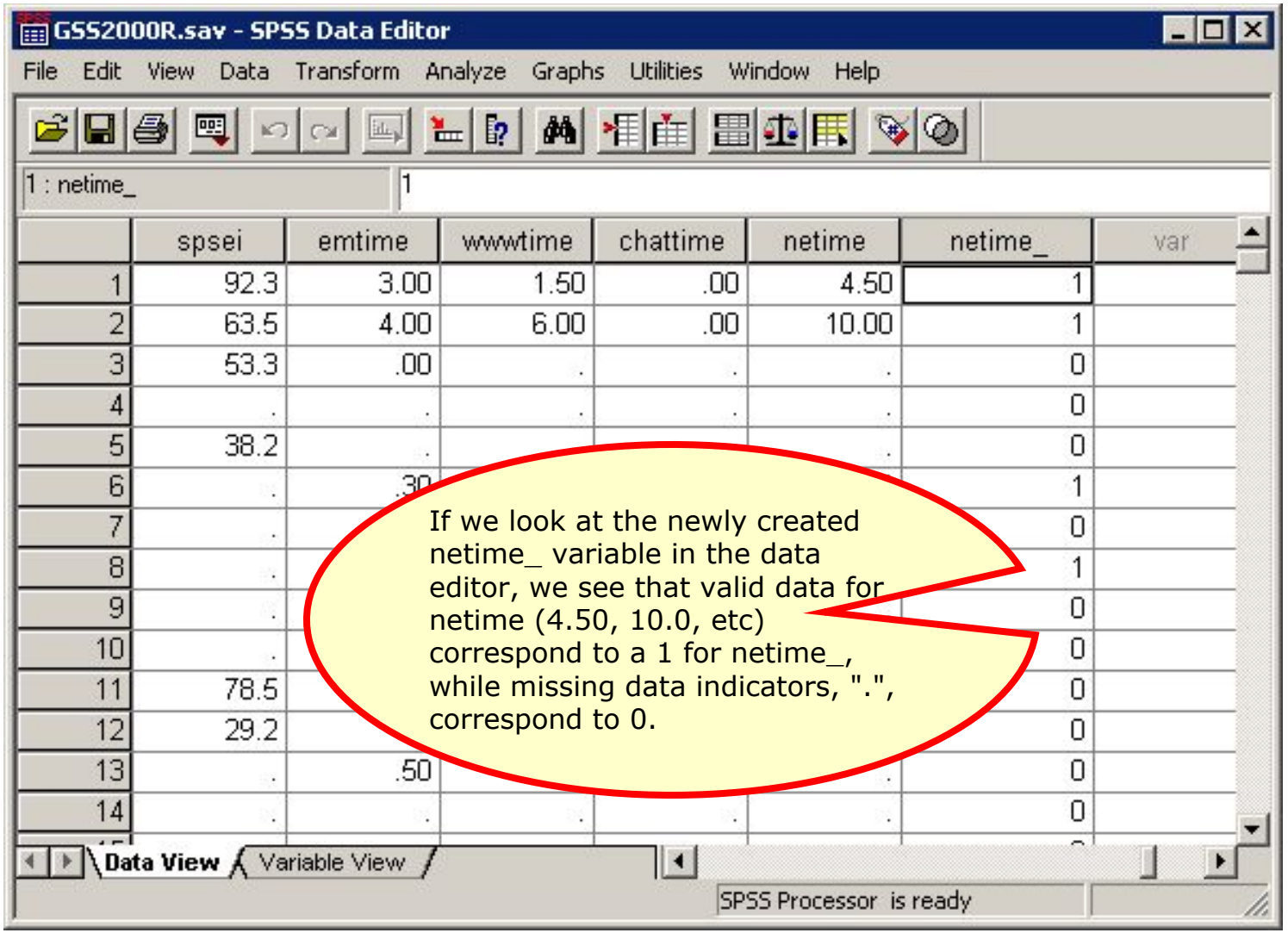
Having completed the changes, we click on the *Continue* button to close the dialog box.

Creating the missing/valid variable - 8

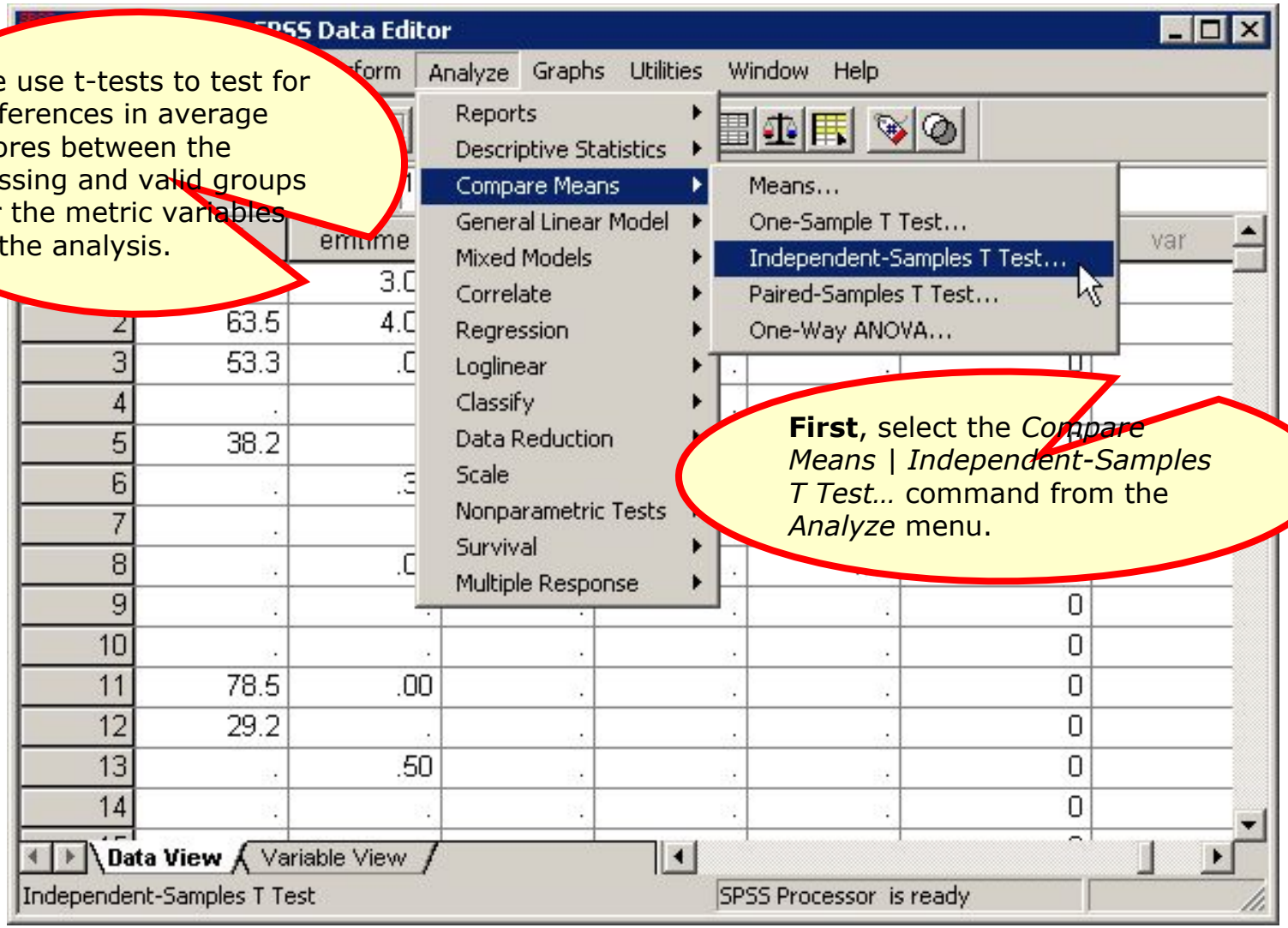


Click on the *OK* button to indicate the completion of the specifications for the new variable.

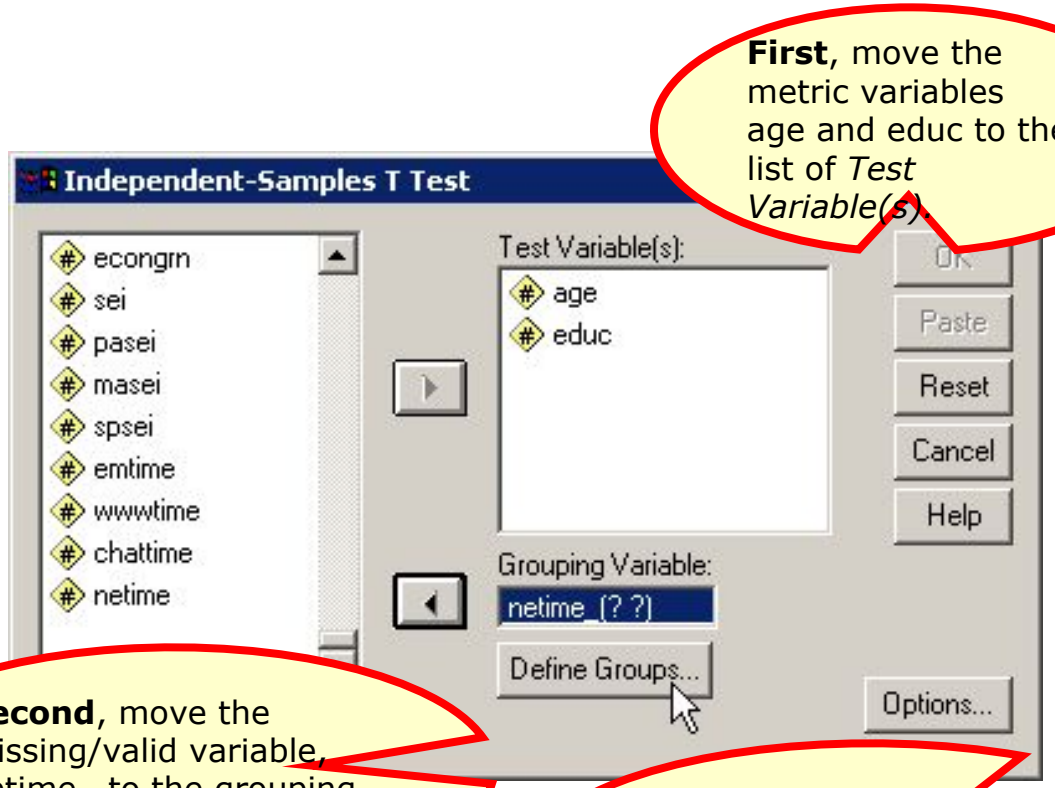
The missing/valid variable in the data editor



T-tests comparing missing and valid cases - 1



T-tests comparing missing and valid cases – 2

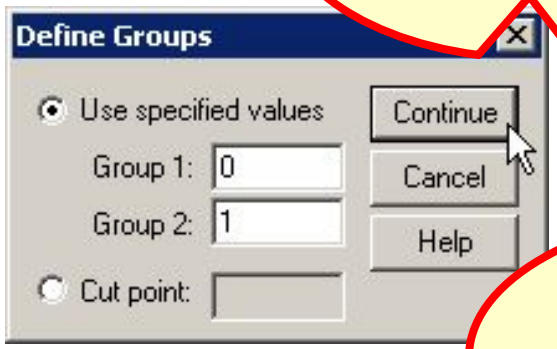


First, move the metric variables age and educ to the list of *Test Variable(s)*.

Second, move the missing/valid variable, netime_ to the grouping variable text box.

Third, click on the *Define Groups…* button to specify the codes for the groups to compare in the analysis.

T-tests comparing missing and valid cases – 3

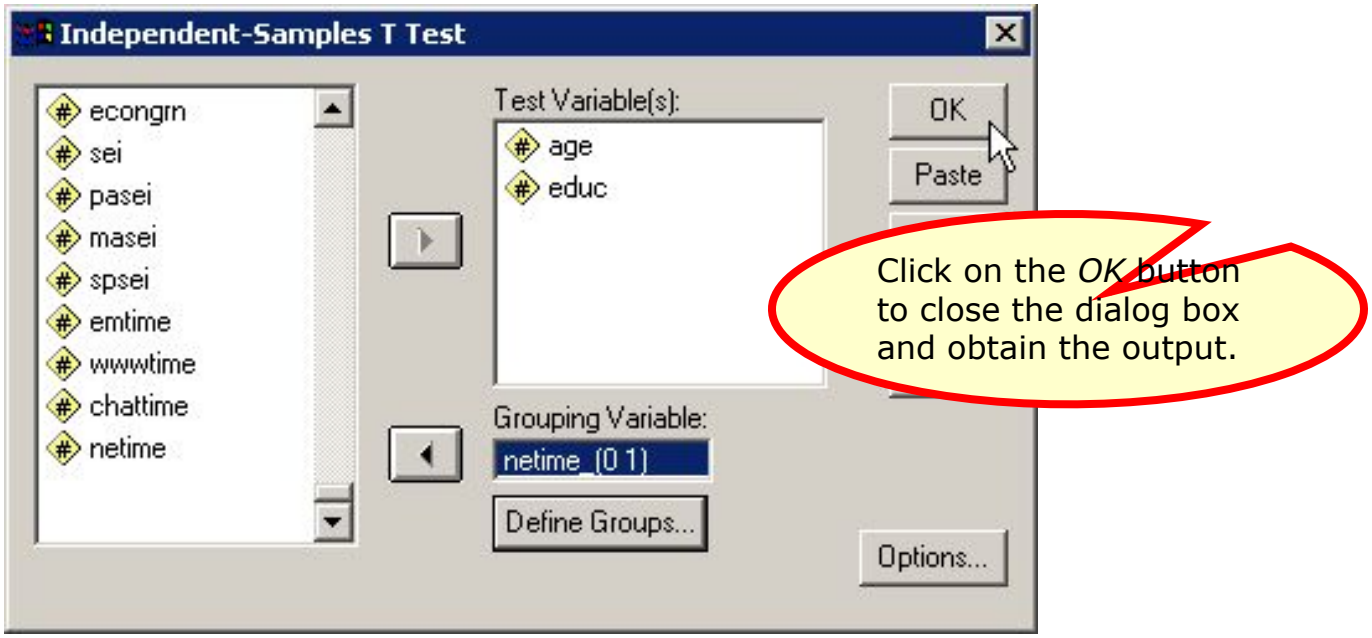


First, type the number 0 for the missing group into the *Group 1* text box.

Second, type the number 1 for the valid group into the *Group 2* text box.

Third, click on the *Continue* button complete the definition of the groups for the independent variable.

T-tests comparing missing and valid cases – 4



Output for the t-tests - 1

Group Statistics

	NETIME_	N	Mean	Std. Deviation	Std. Error Mean
AGE	0	177	48.32	18.414	1.384
	1	93	41.55	12.117	1.256
EDUC	0	177	12.34	2.812	.2
	1	92	14.62	2.554	

There were significant differences in the statistical tests comparing cases with missing data to cases with valid data.

Cases who had missing data for the variable "total hours spent on the Internet" [netime] had an average score on the variable "age" [age] that was 6.77 units higher than the average for cases who had valid data (t=3.624, p<0.001).

Indepe

		Levene's Test for Equality of Variances					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
AGE	Equal variances assumed	22.640	.000	3.201	268	.002	6.77
	Equal variances not assumed			3.624	254.703	.000	6.77
EDUC	Equal variances assumed	.050	.823	-6.507	267	.000	-2.28
	Equal variances not assumed			-6.708	200.606	.000	-2.28

Output for the t-tests - 2

Group Statistics

	NETIME_	N	Mean	Std. Deviation	Std. Error Mea
AGE	0	177	48.32	18.414	
	1	93	41.55	12.117	
EDUC	0	177	12.34	2.812	
	1	92	14.62	2.554	

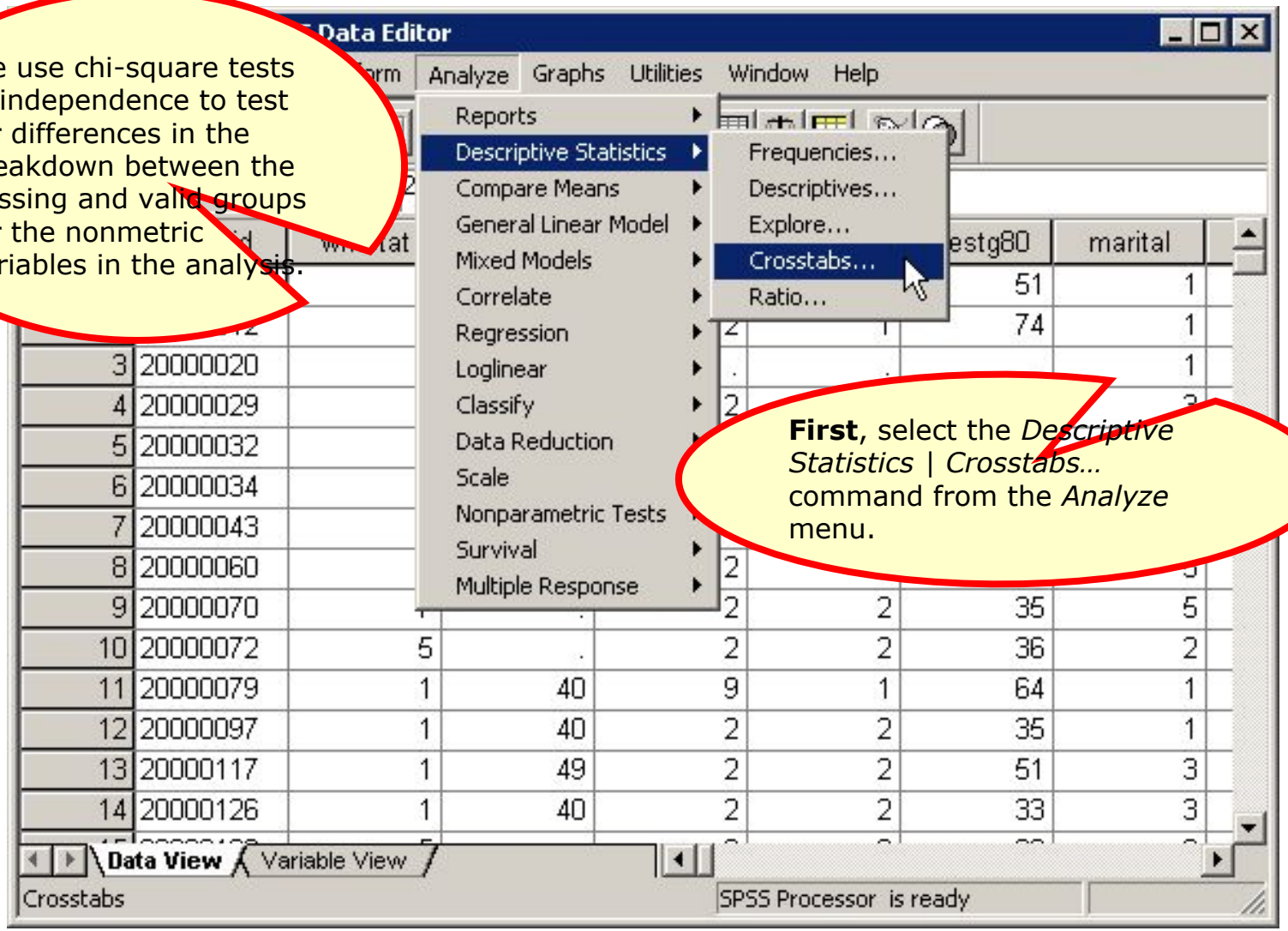
Cases who had missing data for the variable "total hours spent on the Internet" [netime] had an average score on the variable "highest year of school completed" [educ] that was 2.28 units lower than the average for cases who had valid data (t=-6.708, p<0.001).

Indep

		Levene's Test for Equality of Variances		t-t	of Me		
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
AGE	Equal variances assumed	22.640	.000	3.201	268	.002	6.77
	Equal variances not assumed			3.624	254.703	.000	6.77
EDUC	Equal variances assumed	.050	.823	-6.507	267	.000	-2.28
	Equal variances not assumed			-6.708	200.606	.000	-2.28

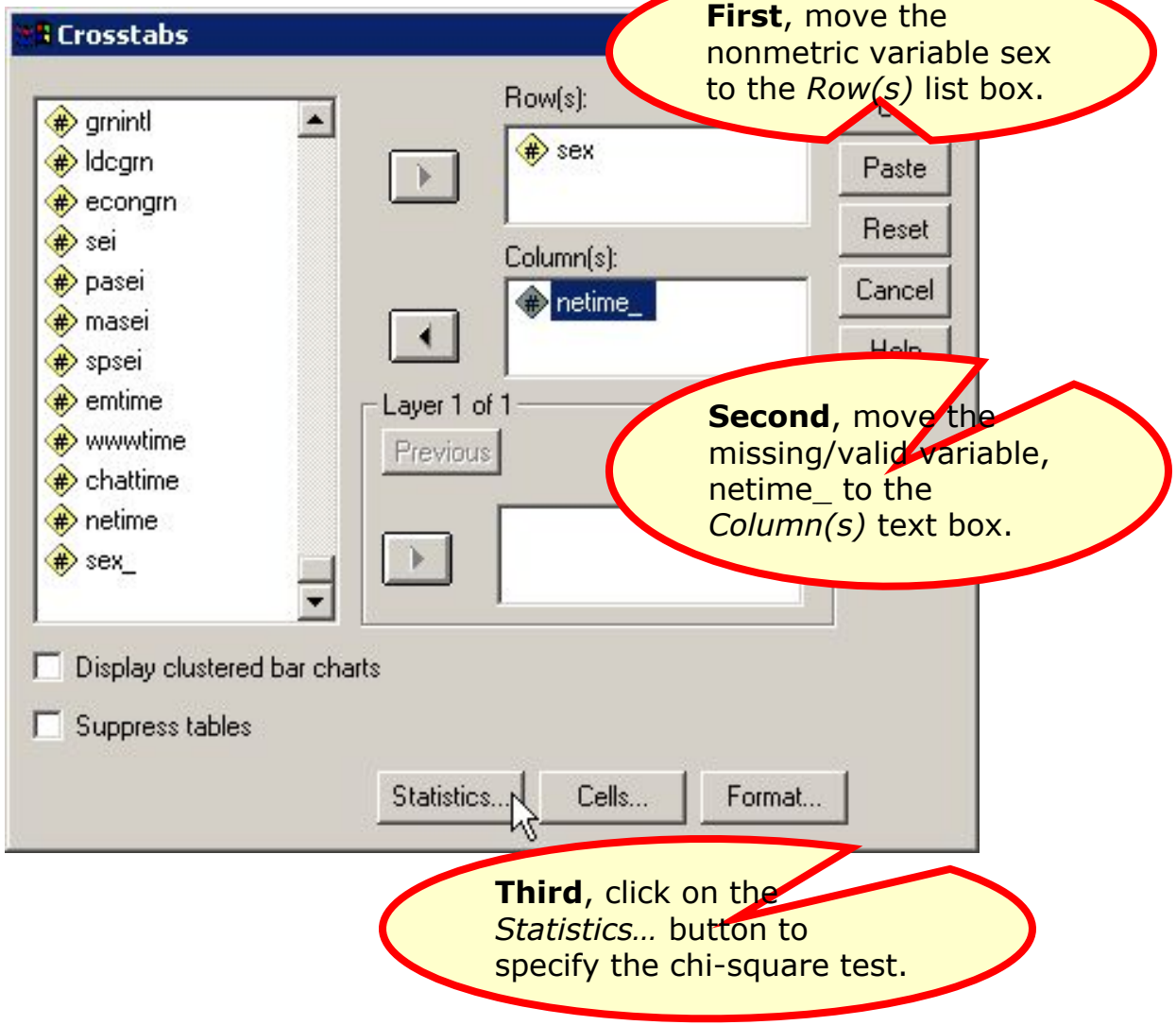
Chi-square tests comparing missing and valid cases - 1

We use chi-square tests of independence to test for differences in the breakdown between the missing and valid groups for the nonmetric variables in the analysis.

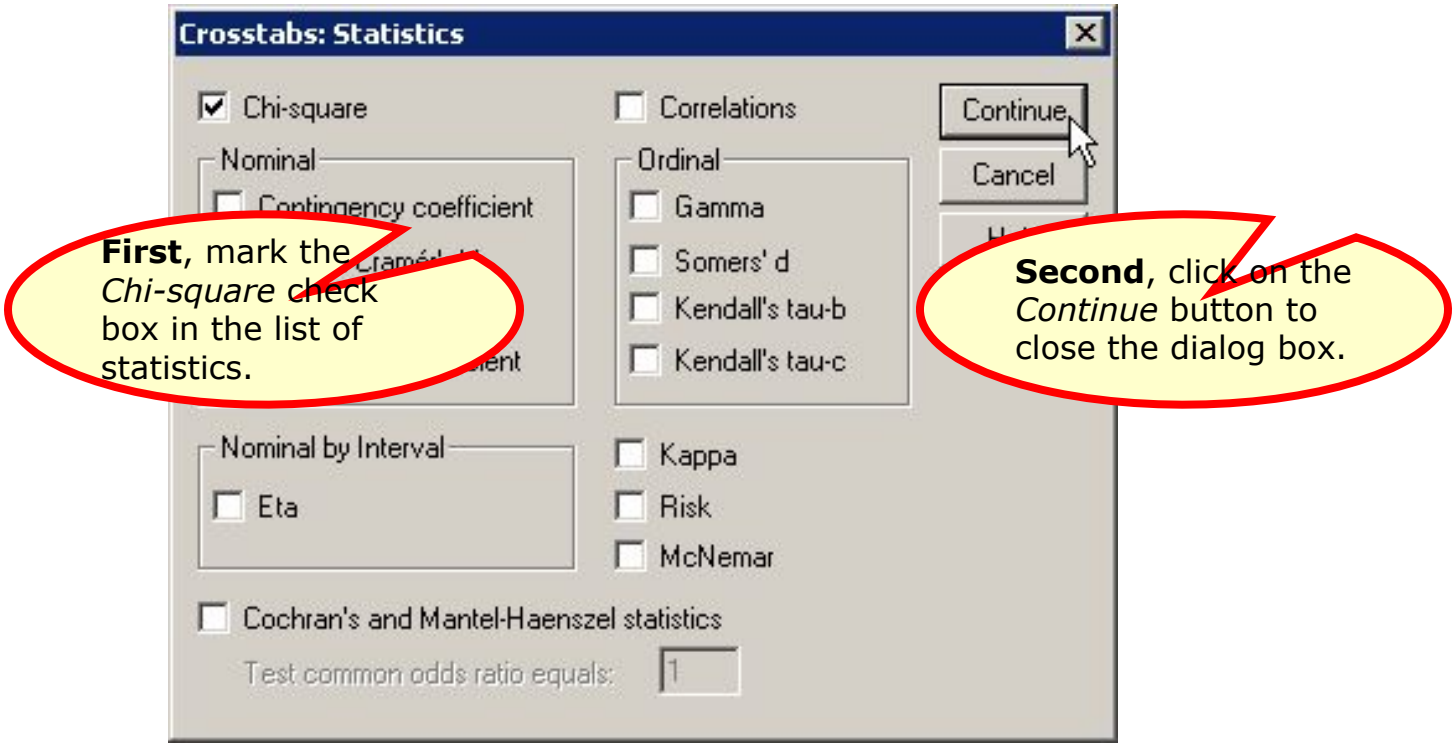


First, select the *Descriptive Statistics | Crosstabs…* command from the *Analyze* menu.

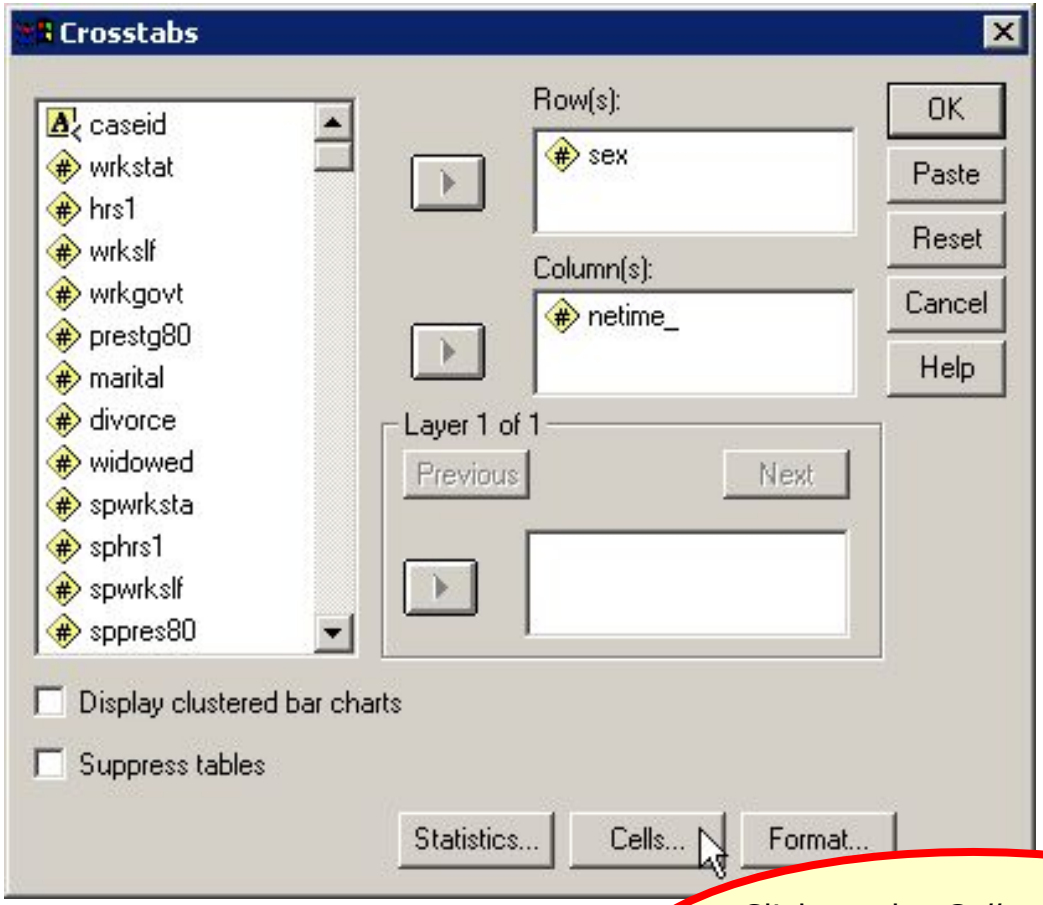
Chi-square tests comparing missing and valid cases - 2



Chi-square tests comparing missing and valid cases - 3



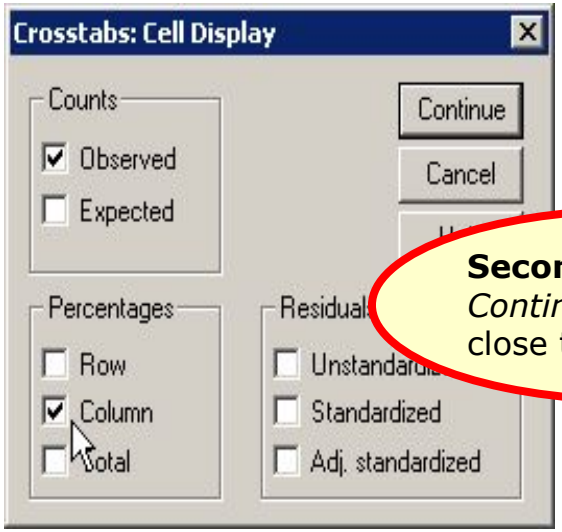
Chi-square tests comparing missing and valid cases - 4



Click on the *Cells..* button to request that column percentages be included in the cross tabulated table.

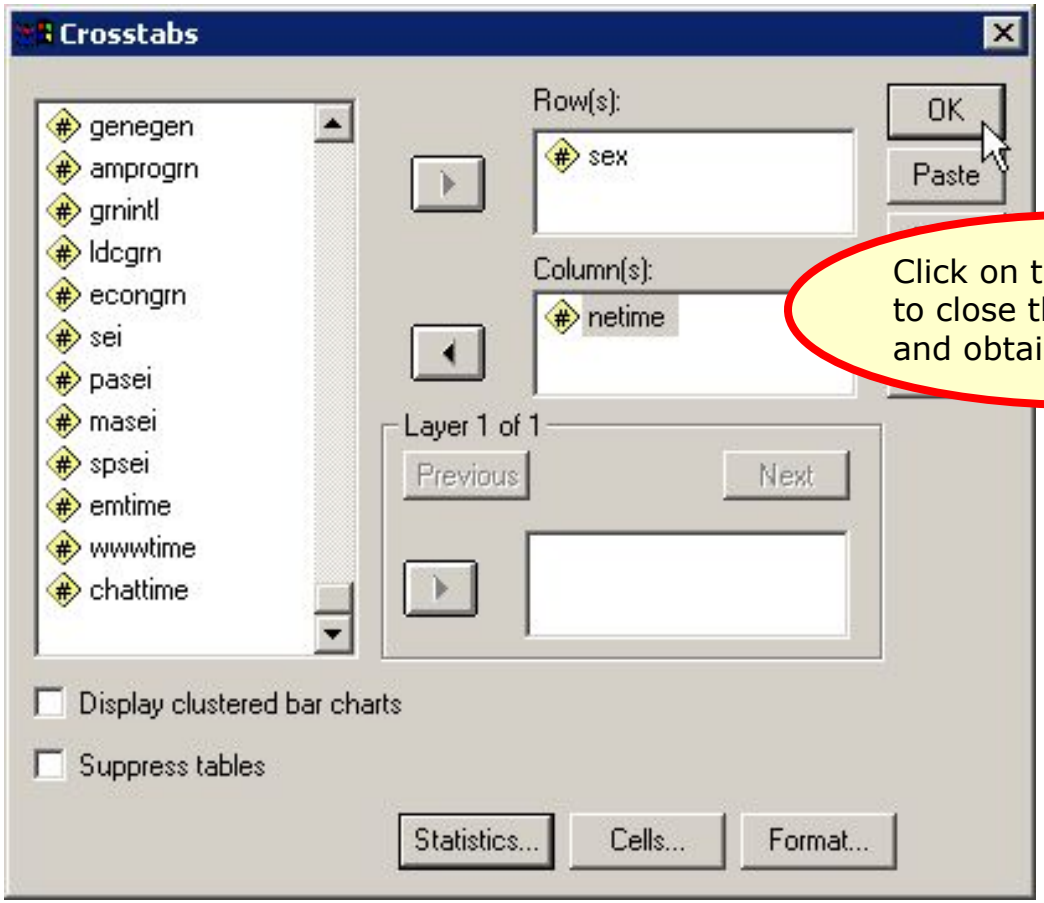
Chi-square tests comparing missing and valid cases - 5

First, mark the *Column* check box in the *Percentages* panel.



Second, click on the *Continue* button to close the dialog box.

Chi-square tests comparing missing and valid cases - 6



Output for the chi-square test

RESPONDENTS SEX * Valid/Missing TOTAL TIME SPENT ON THE INTERNET Crosstabulation

			Valid/Missing TOTAL TIME SPENT ON THE INTERNET		
			0	1	Total
RESPONDENTS SEX	1	Count	73	38	111
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET	41.2%	40.9%	41.1%
	2	Count	104	55	159
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET	58.8%		
Total		Count			
		% within Valid/Missing TOTAL TIME SPENT ON THE INTERNET			

On the chi-square test, the difference in the breakdown for the missing cases is not statistically different from the breakdown for the valid cases.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.004[b]	1	.952		
Continuity Correction[a]	.000	1	1.000		
Likelihood Ratio	.004	1	.952		
Fisher's Exact Test				1.000	.529
Linear-by-Linear Association	.004	1	.952		
N of Valid Cases	270				

Answer 1

In the dataset GSS2000R, is the following statement true, false, or an incorrect application of a statistic? Use a level of significance of 0.01 for evaluating missing data and assumptions.

In pre-screening the data for use in a multiple regression of the dependent variable "total hours spent on the Internet" [netime] with the independent variables "age" [age], "highest year of school completed" [educ], and "sex" [sex], the missing data analysis did not indicate any need for caution or further analysis for a problematic pattern of missing data.

1. True
2. True with
3. False
4. Inappropriate ap

Since there were significant differences in the statistical tests comparing cases with missing data to cases with valid data, a caution was added to the interpretation of any findings, pending further analysis of the missing data pattern.

The answer to the question is false.

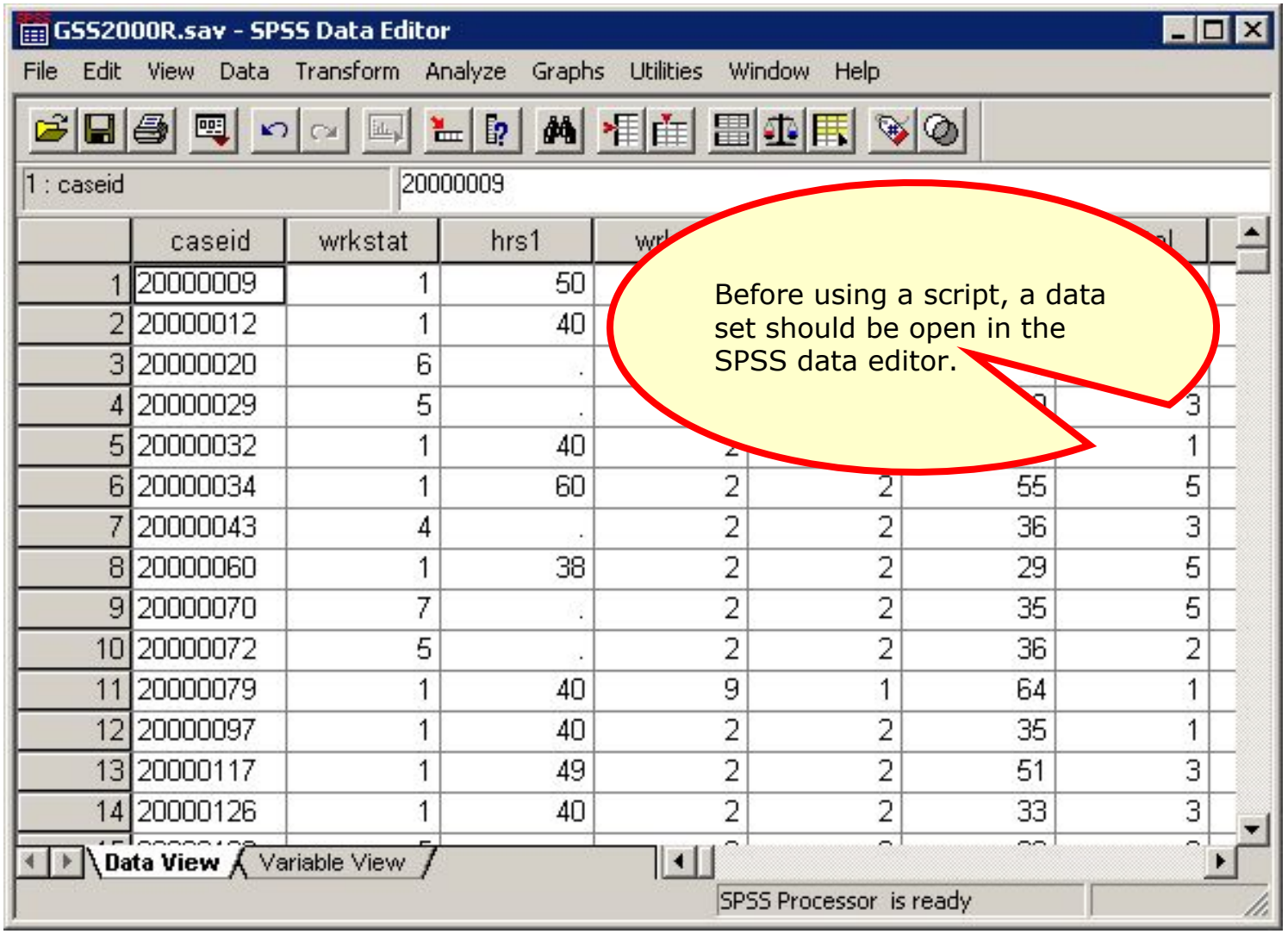
Using scripts

- The process of evaluating missing data requires numerous SPSS procedures and outputs that are time consuming to produce.
- These procedures can be automated by creating an SPSS script. A script is a program that executes a sequence of SPSS commands.
- Though writing scripts is not part of this course, we can take advantage of scripts that I use to reduce the burdensome tasks of evaluating missing data.

Using a script for missing data

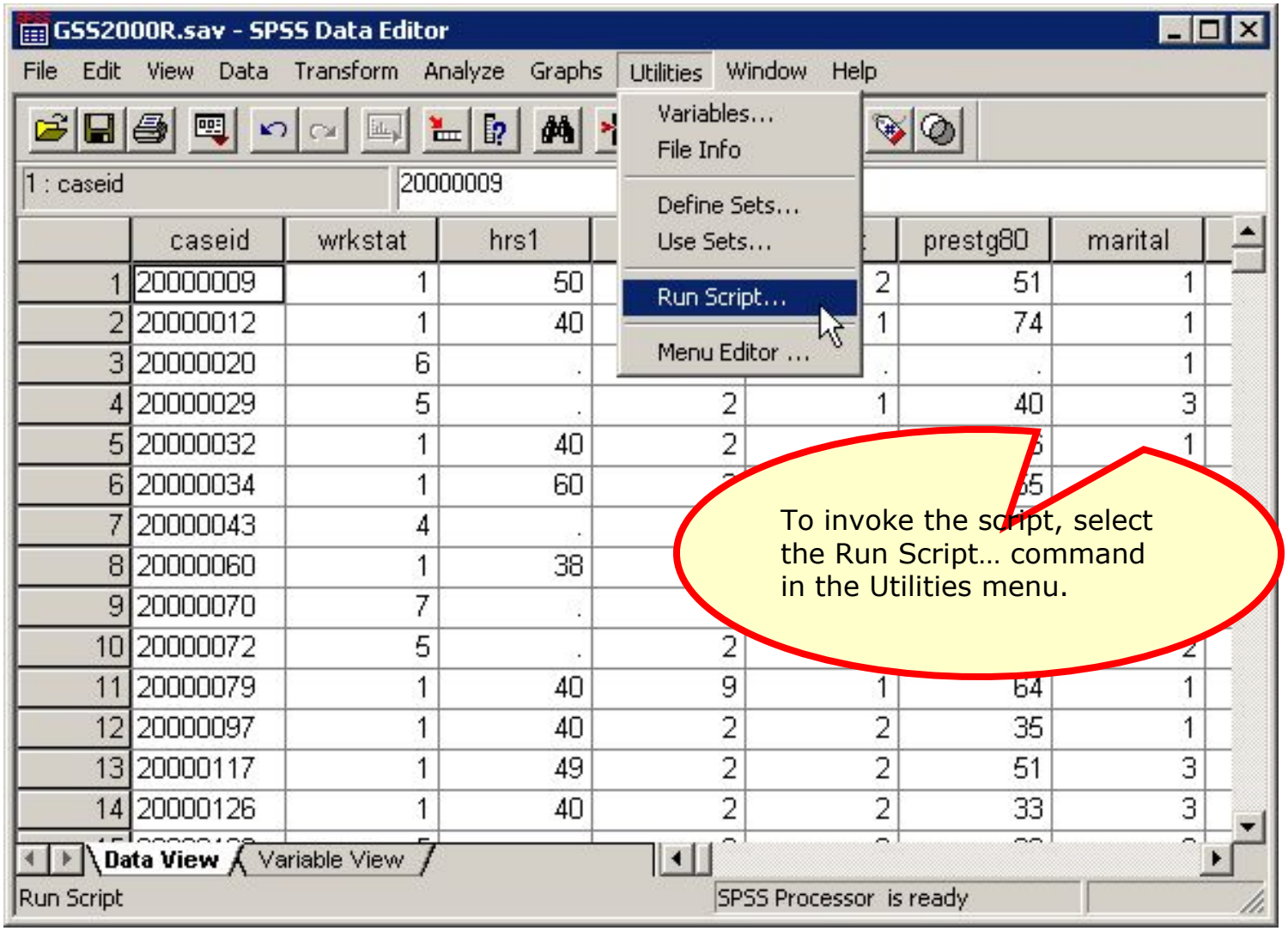
- The script “EvaluatingAssumptionsAndMissingData.exe” will produce all of the output we have used for evaluating missing data (as well as output for testing assumptions).
- Navigate to the link “SPSS Scripts and Syntax” on the course web page.
- Download the script file “EvaluatingAssumptionsAnd MissingData.exe” to your computer and install it, following the directions on the web page.

Open the data set in SPSS



	caseid	wrkstat	hrs1				
1	20000009	1	50				
2	20000012	1	40				
3	20000020	6	.				
4	20000029	5	.				3
5	20000032	1	40				1
6	20000034	1	60	2	2	55	5
7	20000043	4	.	2	2	36	3
8	20000060	1	38	2	2	29	5
9	20000070	7	.	2	2	35	5
10	20000072	5	.	2	2	36	2
11	20000079	1	40	9	1	64	1
12	20000097	1	40	2	2	35	1
13	20000117	1	49	2	2	51	3
14	20000126	1	40	2	2	33	3

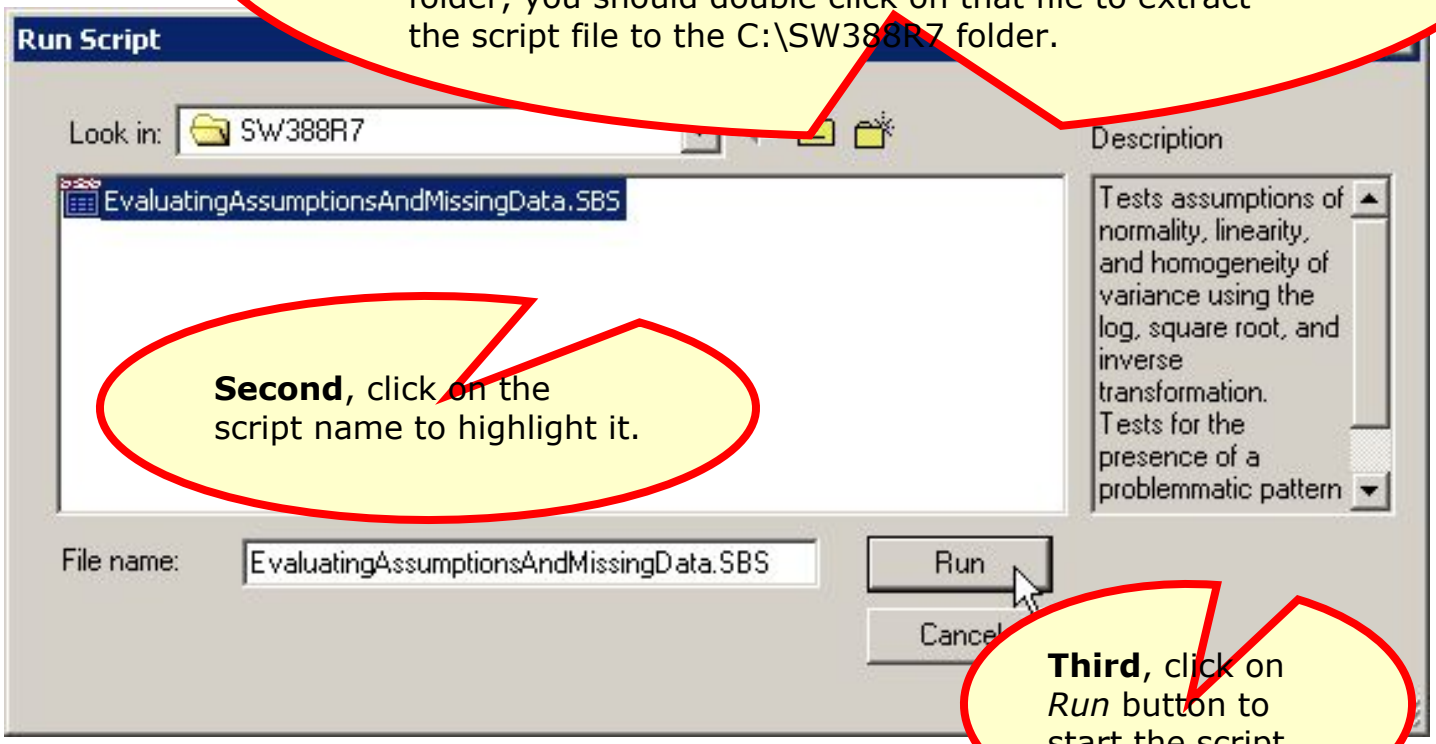
Invoke the script



Select the missing data script

First, navigate to the folder where you put the script. If you followed the directions, you will have a file with an ".SBS" extension in the C:\SW388R7 folder.

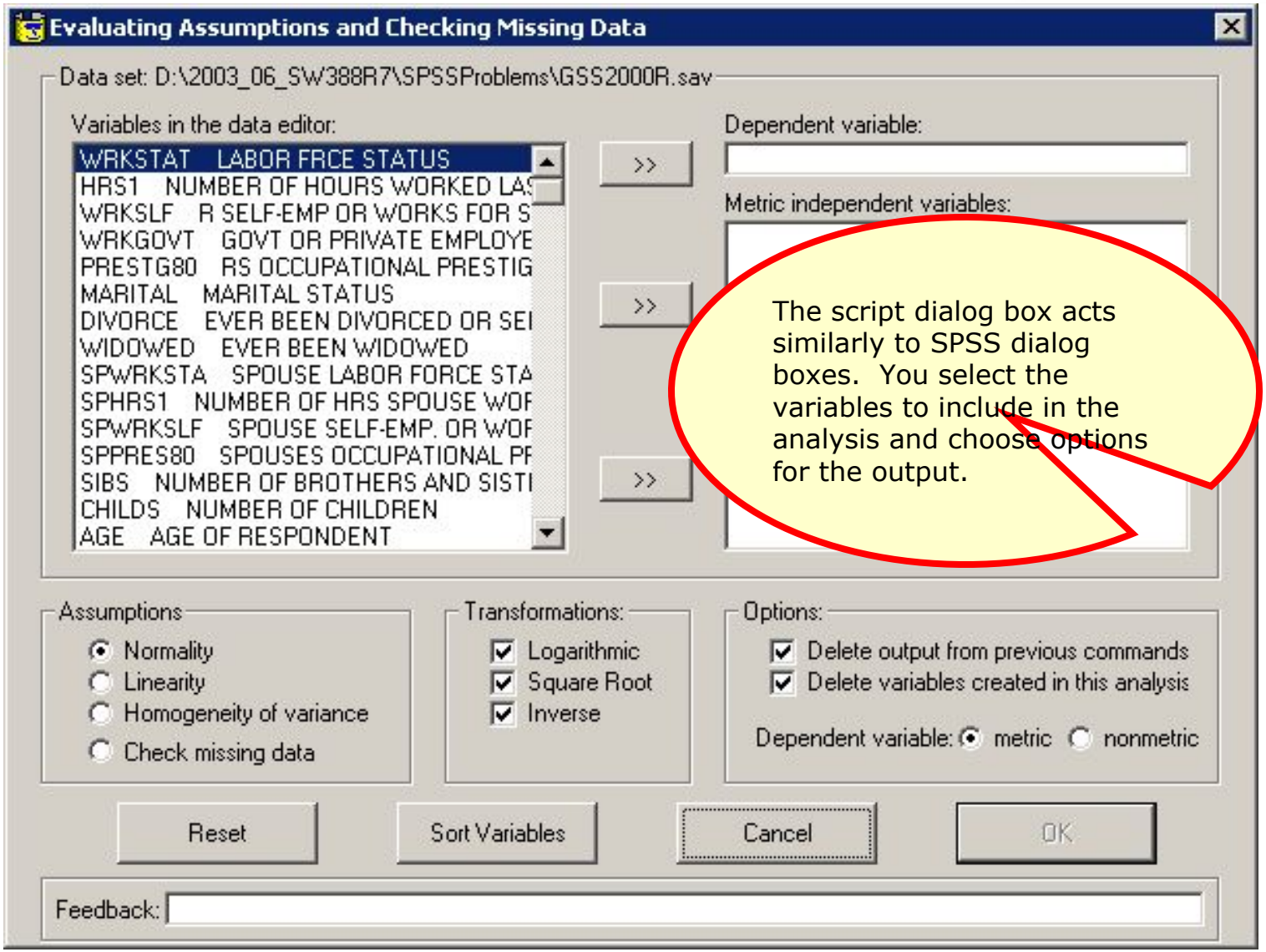
If you only see a file with an ".EXE" extension in the folder, you should double click on that file to extract the script file to the C:\SW388R7 folder.



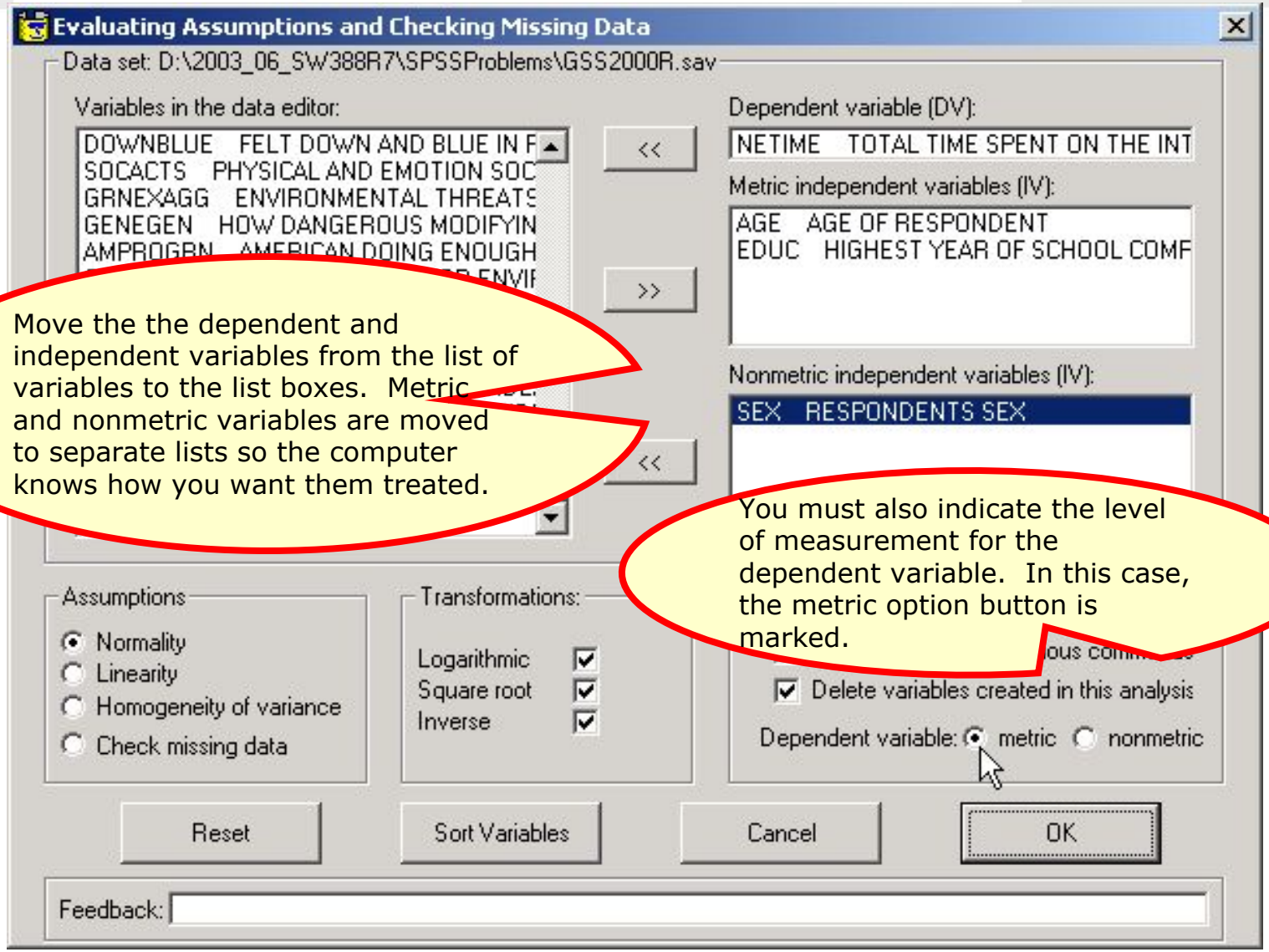
Second, click on the script name to highlight it.

Third, click on *Run* button to start the script.

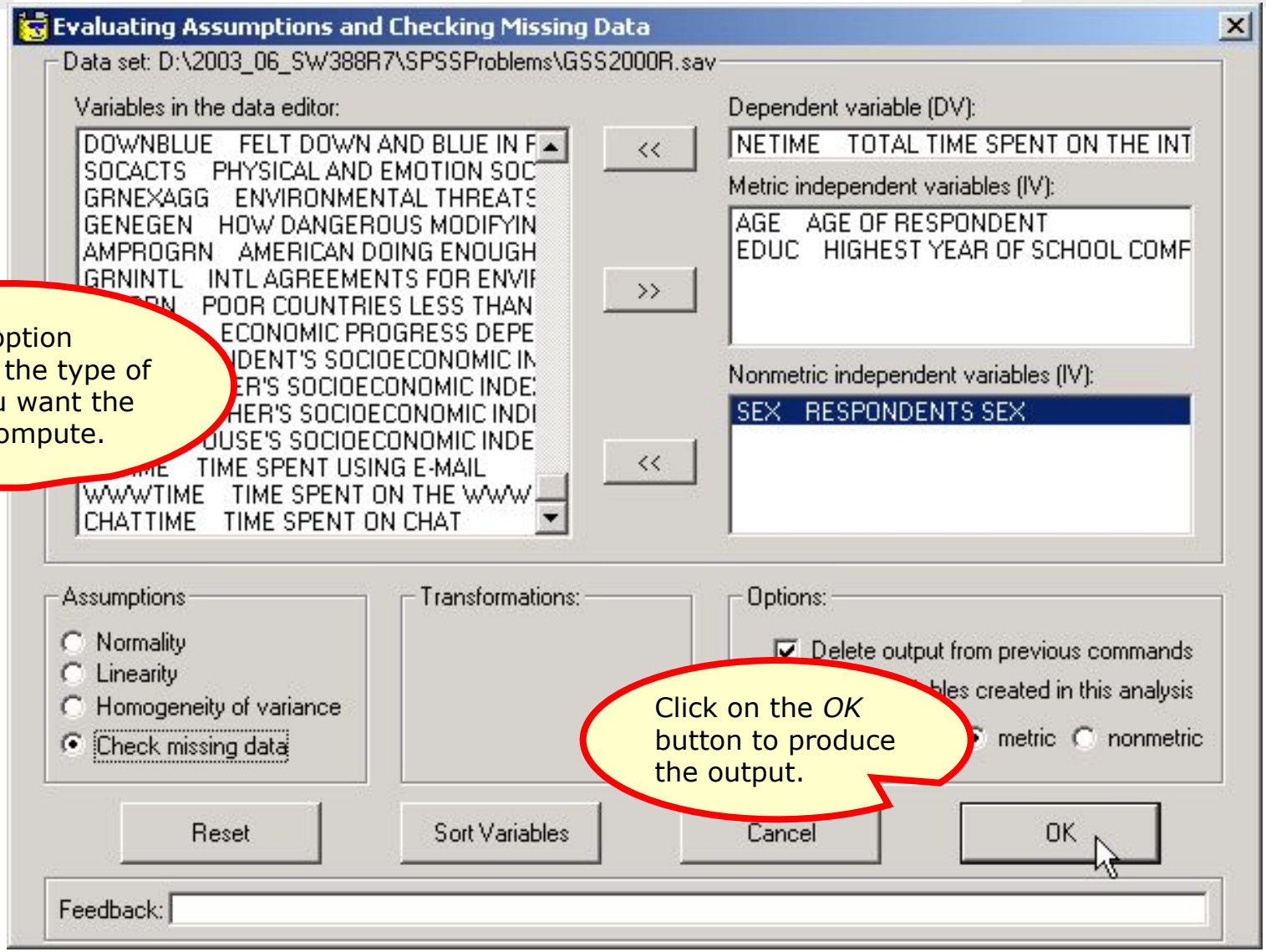
The script dialog



Complete the specifications - 1



Complete the specifications - 2



Mark the option button for the type of output you want the script to compute.

Click on the *OK* button to produce the output.

The script finishes

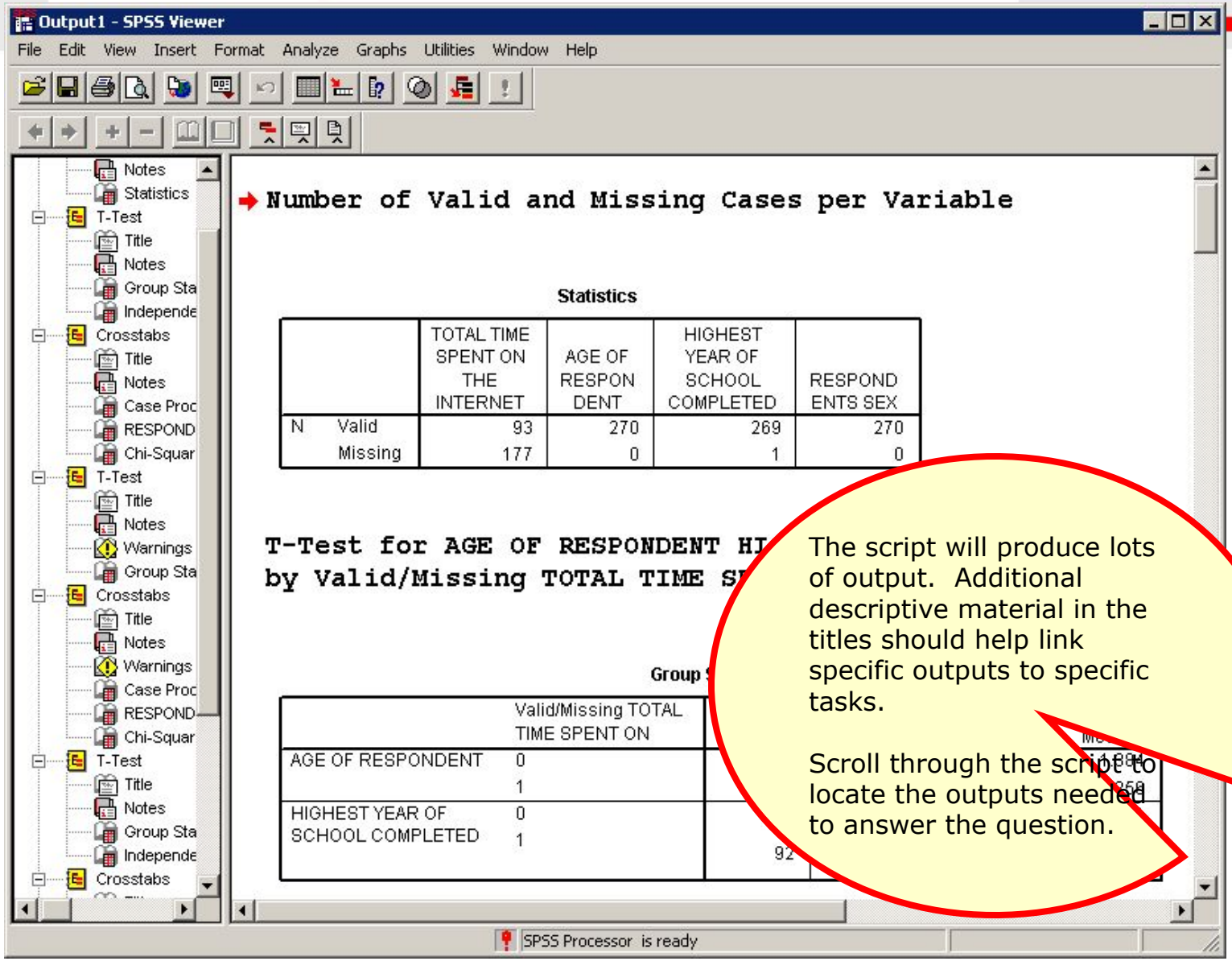
If you SPSS output viewer is open, you will see the output produced in that window.

Since it may take a while to produce the output, and since there are times when it appears that nothing is happening, there is an alert to tell you when the script is finished.

Unless you are absolutely sure something has gone wrong, let the script run until you see this alert.

When you see this alert, click on the OK button.

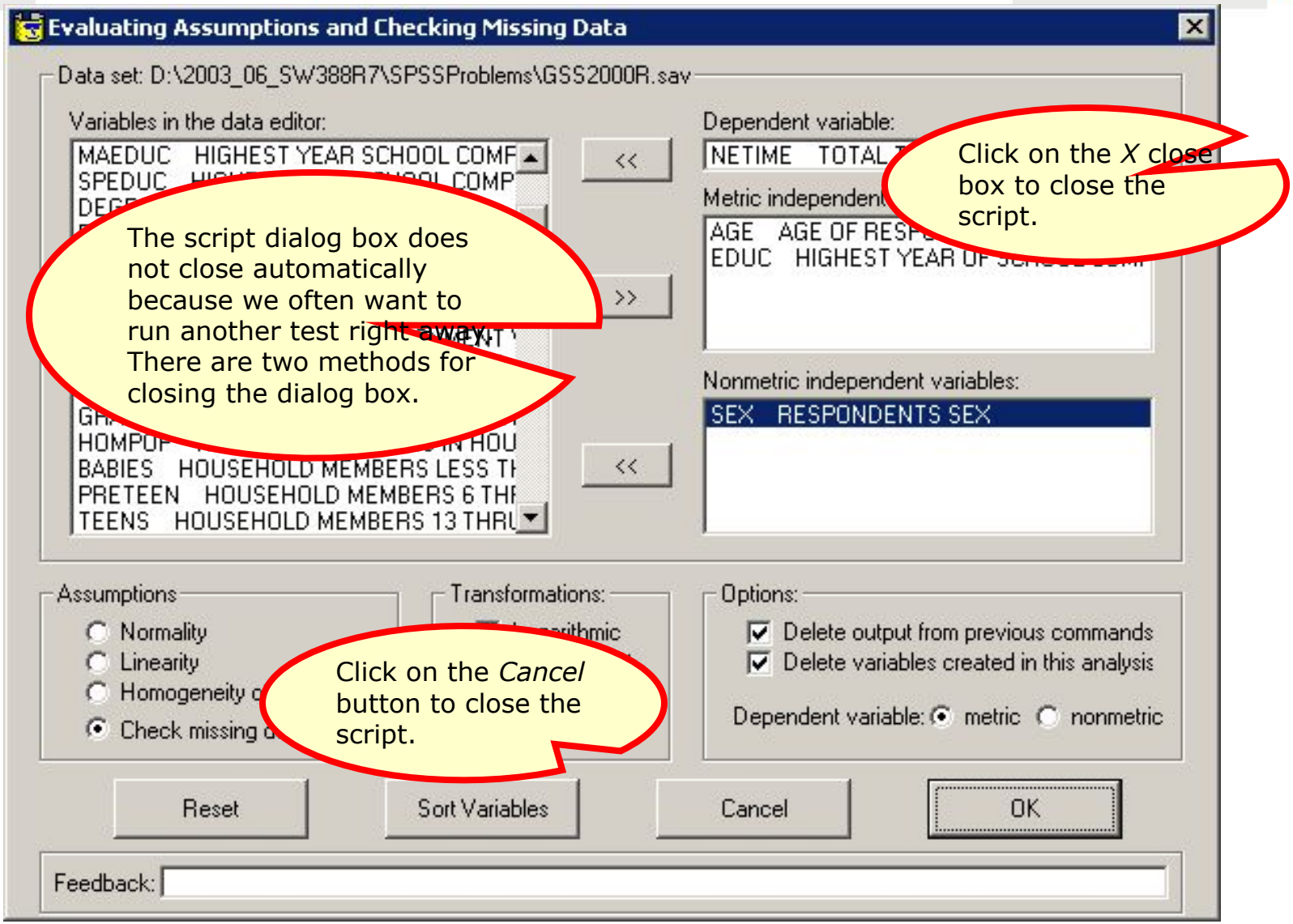
Output from the script - 1



The script will produce lots of output. Additional descriptive material in the titles should help link specific outputs to specific tasks.

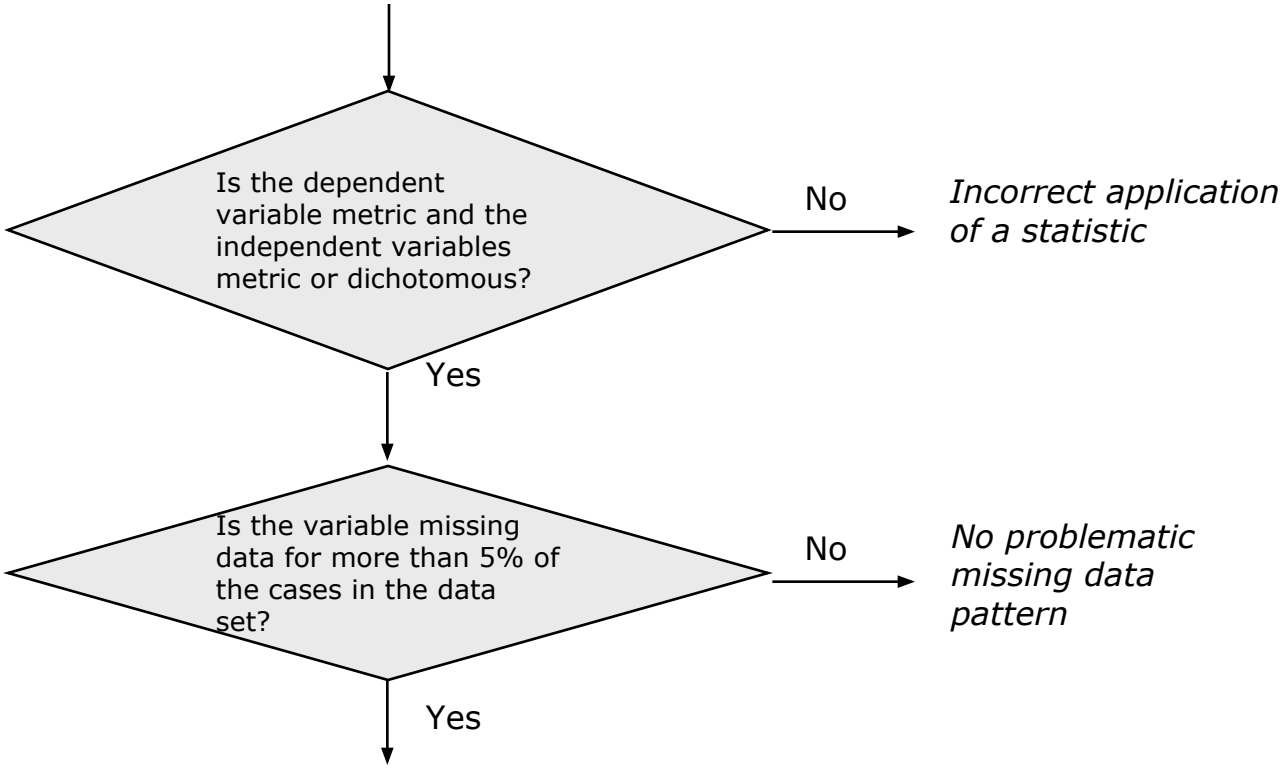
Scroll through the script to locate the outputs needed to answer the question.

Complete the specifications – 2



Steps in analyzing missing data

The following is a guide to the decision process for answering problems about problematic patterns of missing data:



Steps in analyzing missing data