WEB APPENDIX
Sarstedt, M. & Mooi, E. (2019). *A concise guide to market research. The process, data, and methods using SPSS* (3rd ed.). Heidelberg: Springer.

Missing Value Analysis and Multiple Imputation in SPSS

Missing Value Analysis

We use the Oddjob dataset to illustrate how to run a missing value analysis in SPSS. First, let's check whether our data contain missing values and, if applicable, identify the underlying missing value pattern using Little's MCAR test. To do so, go to ► Analyze ► Missing Value Analysis. In the dialog box that opens (Fig. A5.1), move all the continuous variables (labelled Scale in SPSS) into the **Quantitative Variables** box and all the nominal and ordinal variables into the **Categorical Variables** box. Under **Estimation**, check the box next to **EM**, which is the abbreviation of Expectation Maximization. The EM method is an iterative, two-step procedure that can be used for imputing missing values. Each iteration consists of an E (expectation) step and an M (maximization) step. Given the observed values and current estimates of the parameters (e.g., the means, standard deviations, or correlations), the E step finds the missing data's expected value. In the M step, these parameters are re-estimated by assuming that the missing values have been replaced with the expected values. This way, the EM method finds a suitable value, which is then used to impute (i.e., substitute) each missing value. While the EM method is a good approach to missing value treatment, multiple imputation is generally considered superior in this respect. Nevertheless, we select the EM method, as this will also produce Little's MCAR test results, which we need to assess the missing value type.

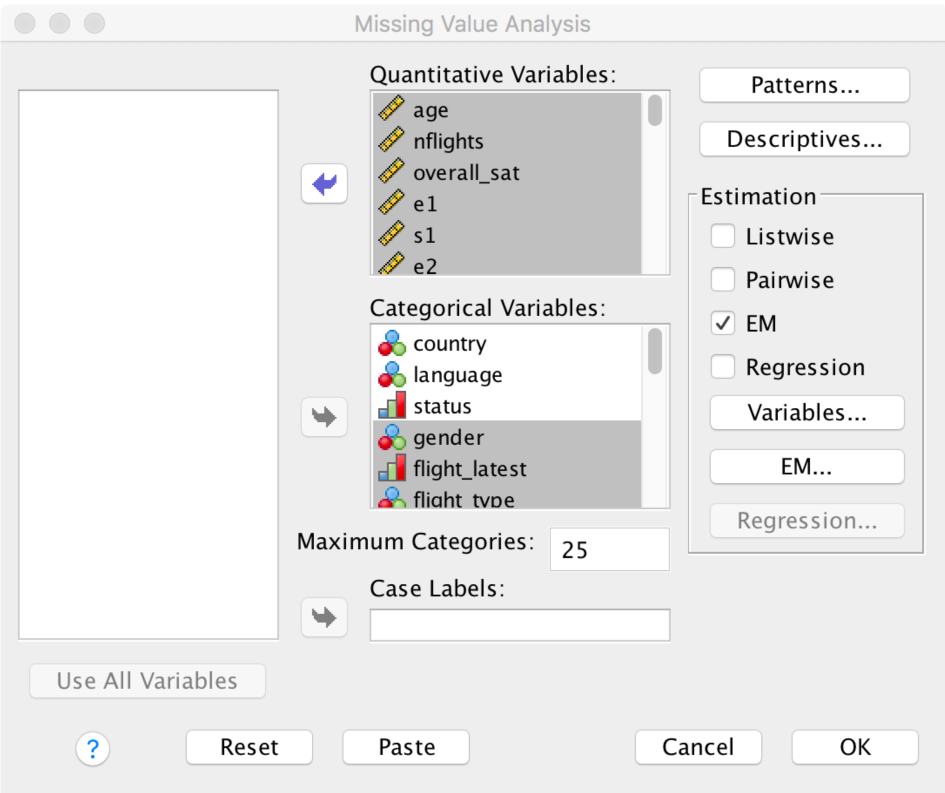


Fig. A5.1 Missing value analysis dialog box

For a basic analysis, there is no need to use the options **Variables…** or **EM…** and you can simply click on **OK**. SPSS will then produce output similar to that in Table A5.1. This table shows a series of descriptive statistics, including the number and percentage of missing values per variable. As can be seen, all of the expectation and satisfaction variables, except for *e23* and *s23* have missing values. While most of these variables have between 20 and 30 missing values, *e3* and *s3* ("... in case something does not work out as planned, Oddjob Airways will find a good solution.") have the most number of missing values (**111**, which correspond to **10.4%** of the entire data).

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of Extremes[a] Low	No. of Extremes[a] High
age	1065	50.42	12.275	0	.0	19	26
nflights	1065	13.42	20.226	0	.0	0	26
nps	1065	8.28	2.516	0	.0	77	0
e1	1038	86.08	19.395	27	2.5	42	0
s1	1038	60.91	26.022	27	2.5	48	0
e2	1040	86.47	19.292	25	2.3	41	0
s2	1040	59.64	25.750	25	2.3	55	0
e3	954	84.03	21.006	111	10.4	48	0
s3	954	55.62	25.072	111	10.4	51	0
e4	1035	87.48	19.002	30	2.8	49	0
s4	1035	57.27	27.543	30	2.8	49	0
e5	1041	77.56	21.183	24	2.3	41	0
s5	1041	56.61	22.518	24	2.3	31	0
e6	1041	78.72	20.761	24	2.3	41	0
s6	1041	56.21	22.150	24	2.3	31	0
e7	1048	80.31	21.890	17	1.6	50	0
s7	1048	51.76	24.646	17	1.6	33	0
e8	1034	78.30	19.732	31	2.9	31	0
s8	1034	57.42	21.402	31	2.9	29	0
e9	1036	87.80	17.042	29	2.7	68	0
s9	1036	72.23	20.713	29	2.7	14	0
e10	1025	84.40	19.738	40	3.8	41	0
s10	1025	64.54	21.408	40	3.8	21	0
e11	1045	84.58	18.336	20	1.9	35	0
s11	1045	64.49	22.066	20	1.9	23	0
e12	999	75.98	22.184	66	6.2	39	0
s12	999	67.19	19.168	66	6.2	9	0
e13	1031	83.21	19.594	34	3.2	43	0
s13	1031	63.18	22.152	34	3.2	27	0
e14	976	80.70	20.810	89	8.4	38	0
s14	976	55.16	24.869	89	8.4	51	0
e15	1036	76.75	21.920	29	2.7	46	0
s15	1036	56.04	24.136	29	2.7	36	0
e16	1027	70.43	24.681	38	3.6	36	0
s16	1027	56.24	23.077	38	3.6	38	0
e17	1041	83.17	19.162	24	2.3	41	0
s17	1041	63.15	23.632	24	2.3	31	0
e18	1034	82.34	20.338	31	2.9	49	0
s18	1034	59.07	24.362	31	2.9	45	0
e19	1013	73.67	22.255	52	4.9	34	0
s19	1013	57.21	21.661	52	4.9	38	0
e20	1030	81.56	19.779	35	3.3	38	0
s20	1030	62.44	23.144	35	3.3	35	0
e21	1028	80.39	20.628	37	3.5	37	0
s21	1028	58.96	22.684	37	3.5	41	0
e22	1012	70.70	23.643	53	5.0	34	0
s22	1012	57.59	20.644	53	5.0	28	40
e23	1065	76.83	23.096	0	.0	44	0
s23	1065	48.94	22.711	0	.0	38	42
commitment	1065	4.1637	1.73922	0	.0	0	0

country	1065			0	.0		
language	1065			0	.0		
status	1065			0	.0		
gender	1065			0	.0		
flight_latest	1065			0	.0		
flight_type	1065			0	.0		
flight_purpose	1065			0	.0		
flight_class	1065			0	.0		
reputation	1065			0	.0		
sat1	1065			0	.0		
sat2	1065			0	.0		
sat3	1065			0	.0		
overall_sat	1065			0	.0		
loy1	1065			0	.0		
loy2	1065			0	.0		
loy3	1065			0	.0		
loy4	1065			0	.0		
loy5	1065			0	.0		
com1	1065			0	.0		
com2	1065			0	.0		
com3	1065			0	.0		

a. Number of cases outside the range (Mean - 2*SD, Mean + 2*SD).

Table A5.1 Univariate statistics table

Scroll down and SPSS shows a series of further tables related to the EM method. These include **Summary of Estimated Means**, **Summary of Estimated Standard Deviations**, and **EM Estimated Statistics**. Below these tables, SPSS shows the results of Little's MCAR test, The test produces a very high χ^2 value of **6,168.002**, which is significant at a 1% level (**Sig. = .000**). Hence, we conclude that the missing values are not MCAR (see Fig. 5.2 in Chap. 5).

Following the procedure outlined in Fig. 5.2 in Chap. 5, we need to carry out further tests to establish whether the missingness in variables *e1* to *e22* and *s1* to *s22* is related to another variable in the dataset. While we could principally test all variables included in our dataset, we focus on the respondents' gender. Specifically, we run a series of χ^2*-tests* by comparing whether or not an observation is missing with the respondent's gender in order to identify potential relationships.

Before proceeding with this step, we need to create dummy variables for the missing observations in each of the variables *e1* to *e22* and *s1* to *s22*. This can be done by going to ► Transform ► Recode into Different Variables. In the dialog box that opens (Fig A5.2), move *e1* into the **Numeric Variable → Output Variable box** and click on **Old and New Values**. In the following dialog box (Fig. A5.3), select **System- or user-missing** under **Old Value**, enter **1** under **New Value**, and click on **Add**. Next, select **All other values** under **Old Value**, enter **0** under **New Value**, and click on **Add**. Click on **Continue**, which will return you to the initial dialog box. Before finishing the recoding, we need to specify a name for the new dummy-coded variable. To do so, enter *e_dummy1* under **Output Variable** and click on **Change**. When clicking on **OK**, SPSS will add a new variable labelled *e_dummy1* to the dataset, which takes the value 1 if a value in *e1* is missing and 0 else. We now need to redo these steps for all other variables (i.e., *e2 – e22* and *s1 – s22*). However, by using the syntax, we can facilitate this analysis substantially. To do so, click on **Paste** and SPSS will open a syntax window similar to the one shown in Fig A5.4.

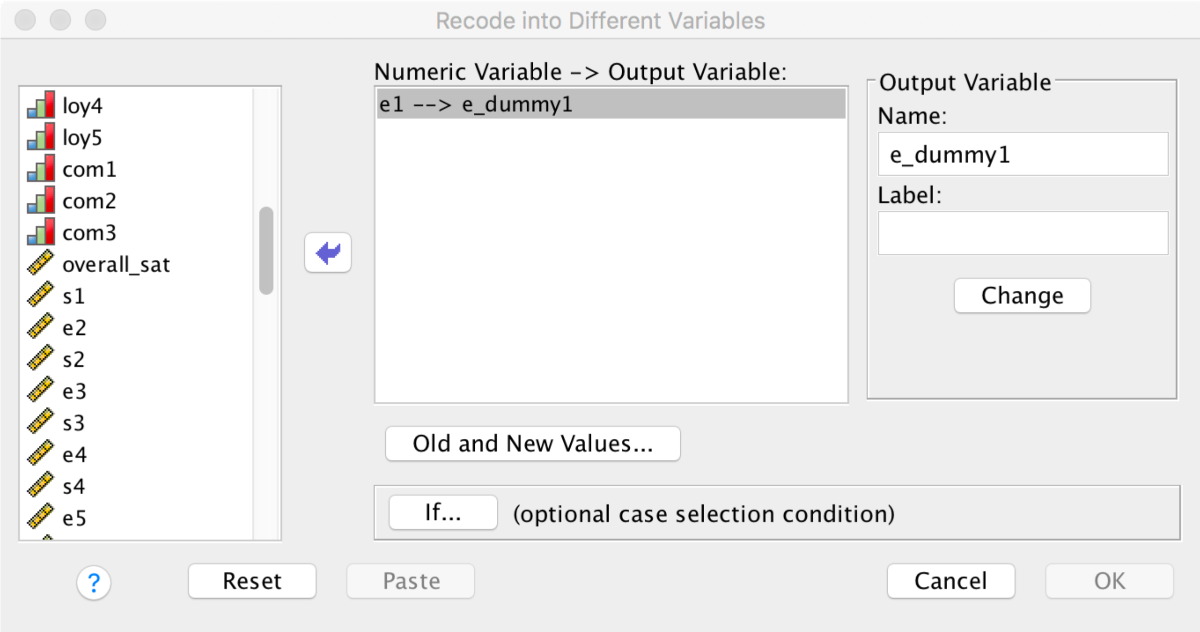


Fig A5.2 Recode into different variables dialog box

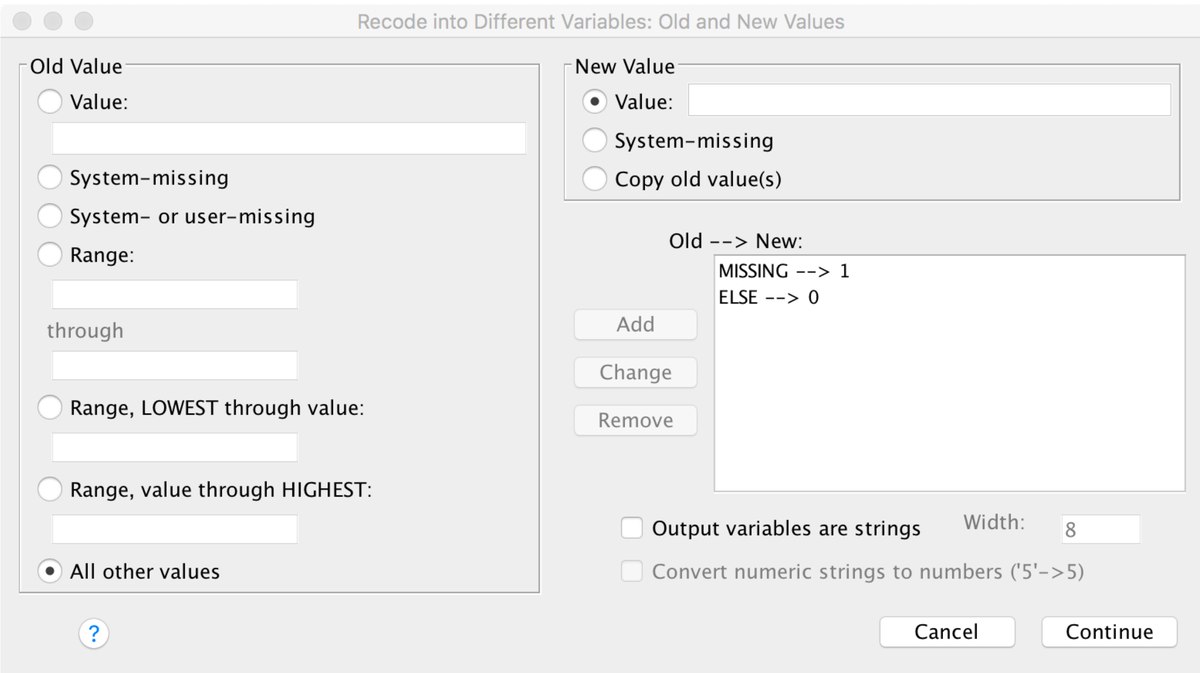


Fig A5.3 Recode into different variables dialog box: Old and new values

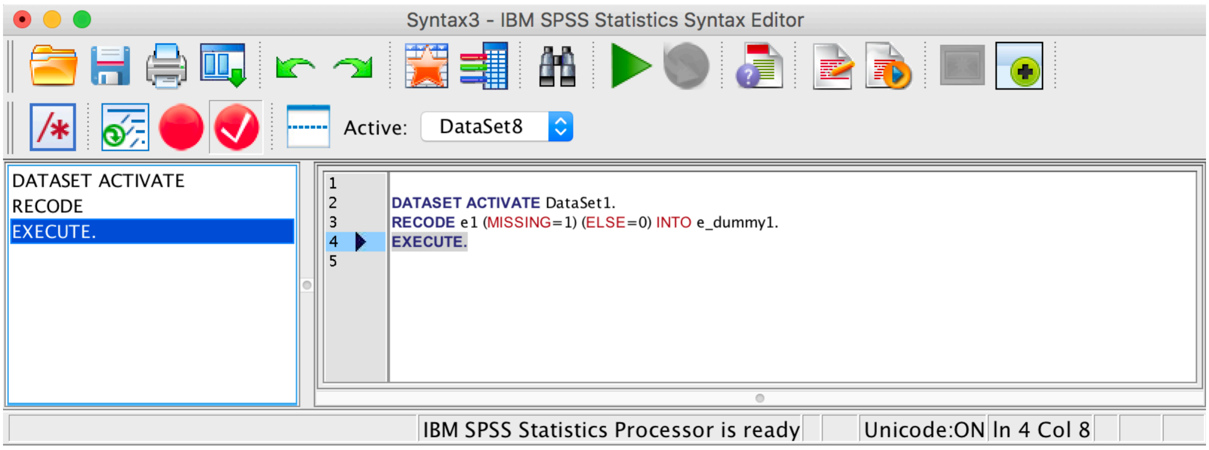


Fig A5.4 Syntax window

The second line in the syntax represents the Recode into Different Variables command from the graphical user interface. As you can see, the code is very intuitive. We can now duplicate the line and replace *e1* with *e2* and *e_dummy1* with *e_dummy2*. The resulting line then looks like this:

```
RECODE e2 (MISSING=1) (ELSE=0) INTO e_dummy2.
```

Selecting the entire code and clicking on the play symbol (the green triangle) in the toolbar will initiate the recoding of *e1* and *e2* into *e_dummy1* and *e_dummy2* (make sure that the command `EXECUTE`. appears behind the `RECODE` command). We can now duplicate and adjust the `RECODE` command for all the remaining expectation and satisfaction items to create a total of 44 dummy variables. However, the syntax allows us to further simplify such recurring commands by using macros. Macros function as a "mini program" within the SPSS syntax. These mini programs are written in a combination of a special SPSS macro language and the standard SPSS syntax language. We can automate the recoding of the remaining variables by using the following macro:

```
DEFINE !dummycode (DESIG1 = !TOKENS(1) / DESIG2 = !TOKENS(1))
!DO !i = !DESIG1 !TO !DESIG2
RECODE !CONCAT(s,!i) (MISSING=1) (ELSE=0) INTO
!CONCAT(s_dummy,!i).
RECODE !CONCAT(e,!i) (MISSING=1) (ELSE=0) INTO
!CONCAT(e_dummy,!i).
!DOEND
!ENDDEFINE.
!dummycode DESIG1=1 DESIG2=22.
EXE.
```

Simply copy and paste this into the syntax editor, select the code, and click on the play symbol (the green triangle) in the toolbar. Discussing the details of the syntax macros is clearly beyond the scope of this book, but the interested reader can find further information in the SPSS help option.

Next, we separately perform a χ^2-test on the respondents' gender and on the 44 dummy variables. To run the χ^2-test, go to ► Analyze ► Descriptive Statistics ► Crosstabs. In the dialog box that opens, move *gender* into the **Row(s)** box and the first dummy variable *e_dummy1* into the **Column(s)** box. Next, click on **Statistics**, check the box **Chi-Square**, and click on **Continue**. Initiate the analysis by clicking on **OK**.

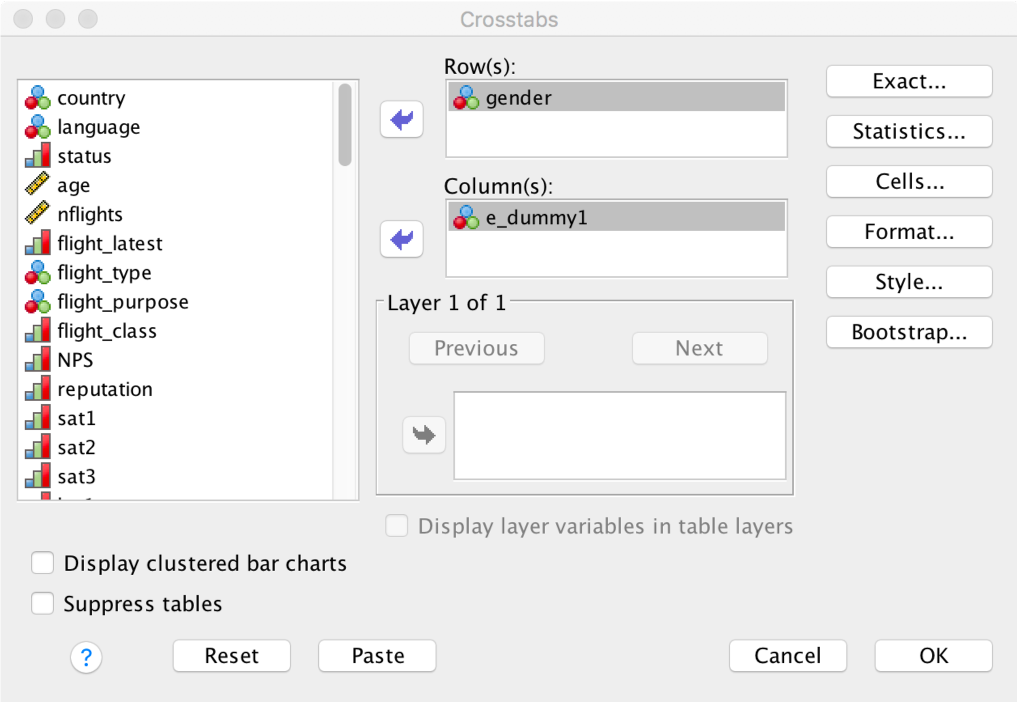


Fig A5.4 Crosstabs dialog box

The *p*-value of **0.627** in Table A5.2 indicates that there is no significant relationship between the respondents' gender and the missingness of observations in *e_dummy1*.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.237[a]	1	.627		
Continuity Correction[b]	.070	1	.791		
Likelihood Ratio	.245	1	.621		
Fisher's Exact Test				.825	.408
Linear-by-Linear Association	.236	1	.627		
N of Valid Cases	1065				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.10.
b. Computed only for a 2x2 table

Table A5.2 χ^2-test output

We would now have to repeat this test for the remaining 43 variables; however, the SPSS syntax facilitates these analyses greatly by means of the following macro:

```
DEFINE !chisquaretest (DESIG1 = !TOKENS(1) / DESIG2 =
!TOKENS(1))
!DO !i = !DESIG1 !TO !DESIG2
CROSSTABS
  /TABLES=gender BY !CONCAT(s_dummy,!i)
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
  CROSSTABS
```

```
  /TABLES=gender BY !CONCAT(e_dummy,!i)
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
!DOEND
!ENDDEFINE.
!chisquaretest DESIG1=1 DESIG2=22.
EXE.
```

The results of all the separate 44 χ^2-tests (not shown here) yield significant relationships for only two variables: *e_dummy18* and *s_dummy18*. Considering that we carried out 44 tests at a significance level of 5%, we can expect $44 \cdot 0.05 \approx 2$ erroneous rejections of the (true) null hypothesis (i.e., type I errors; see Chap. 6). Hence, the two significant results in the χ^2-tests are statistically expected and we can conclude that the data are MNAR—at least with regard to the respondents' *gender*. In principle, we could proceed by testing the relationships between the variables with missing values and other variables, such as *status* or *gender*.

Multiple Imputation

While our prior analyses indicated that the data are MNAR when considering *gender*, we nevertheless proceed by illustrating the use of multiple imputation in SPSS. To initiate multiple imputation, go to ► Analyze ► Multiple Imputation ► Impute Missing Data Values. In the dialog box that opens (Fig. A5.5), move all variables that you wish to include in your subsequent analysis into the **Variables in Model** box. For example, if you want to run a regression of *overall_sat* on *s1*, *s2*, *s3*, *s4*, and *s5*, you need to include these six variables in the multiple imputation procedure (Enders 2010). In addition, you should include other variables that potentially explain (or have been shown to explain; see previous step) the missingness in the variables' observations, such as the respondents' demographics. In our example, we include *overall_sat*, *s1-s5*, *age*, *gender*, and *status* in the multiple imputation procedure.

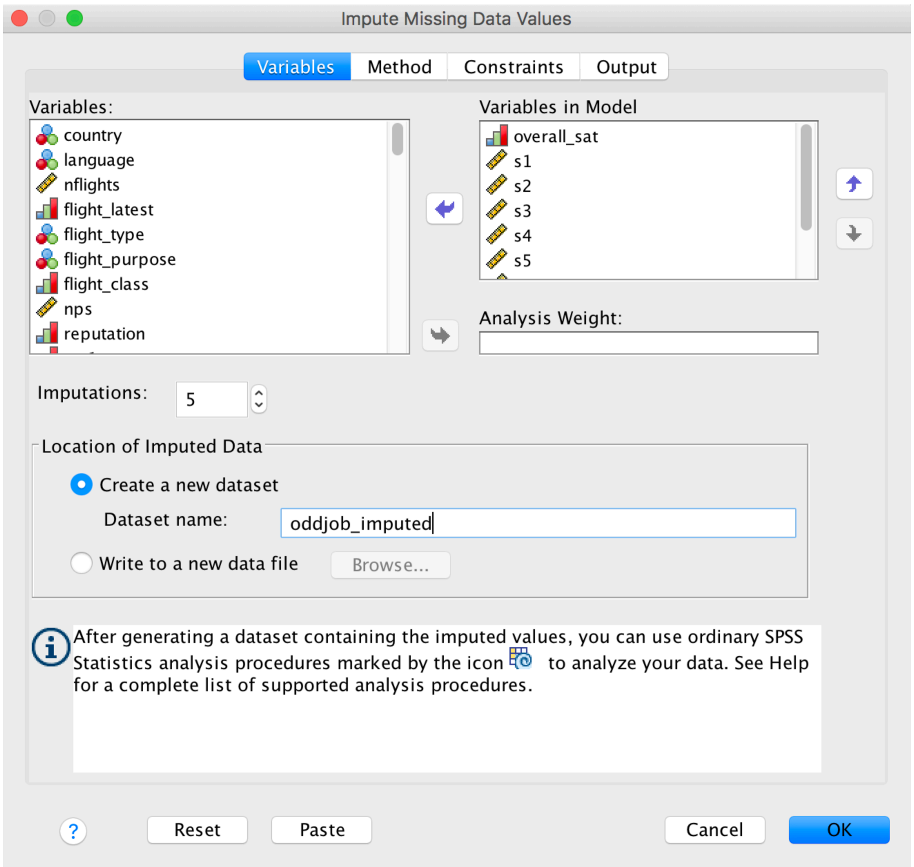


Fig A5.5 Multiple imputation dialog box

Next, specify the number of times the missing values should be replaced (i.e., *m*=**5**) under **Imputations** and indicate a name for the new dataset, such as *oddjob_imputed*, next to **Dataset name**. When clicking on **OK**, SPSS will produce an output similar to Table A5.3.

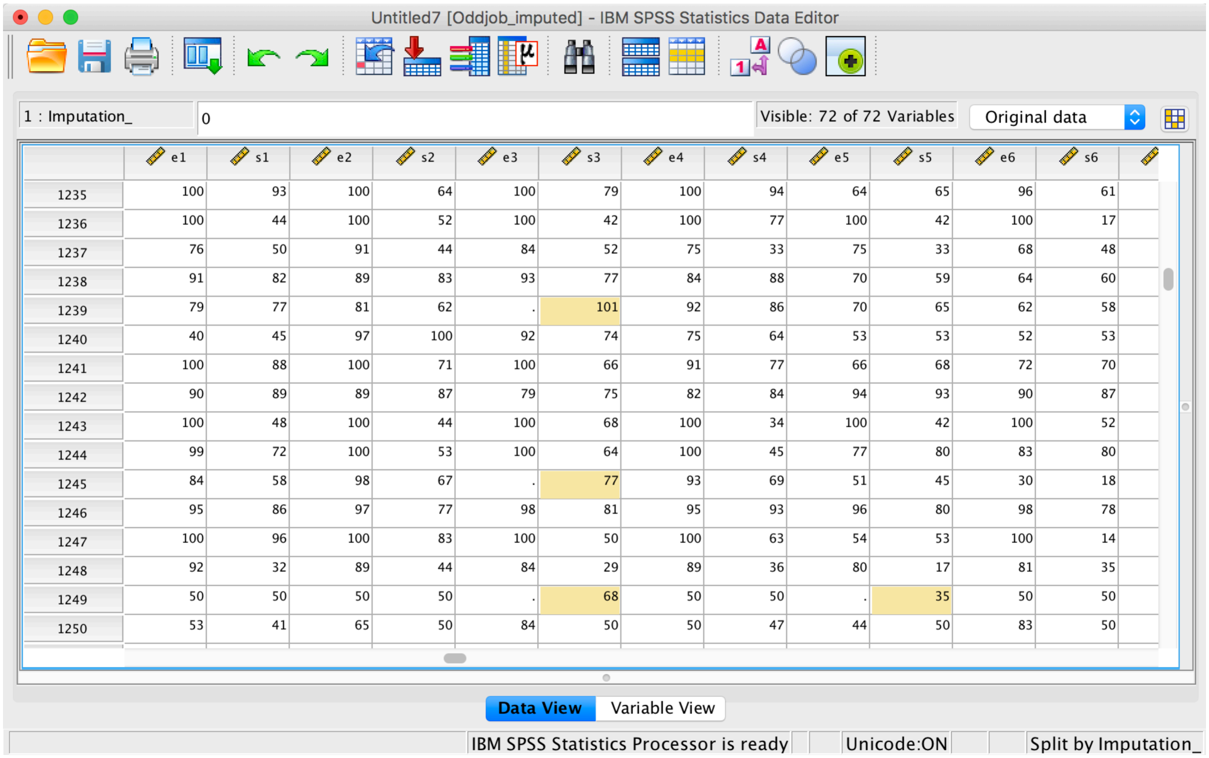
Imputation Models

	Model		Missing Values	Imputed Values
	Type	Effects		
s5	Linear Regression	overall_sat,status,gender,age,s2,s1,s4,s3	24	120
s2	Linear Regression	overall_sat,status,gender,age,s5,s1,s4,s3	25	125
s1	Linear Regression	overall_sat,status,gender,age,s5,s2,s4,s3	27	135
s4	Linear Regression	overall_sat,status,gender,age,s5,s2,s1,s3	30	150
s3	Linear Regression	overall_sat,status,gender,age,s5,s2,s1,s4	111	555

Table A5.3 Multiple imputation output

For each variable with missing values, Table A5.3 shows the number of missing values in the original dataset and the total number of imputed values, which is simply *m* times the number of missing values. The procedure doesn't look as if it has done much for us, but it has, in fact, created five datasets containing imputed values, which are in *Oddjob_imputed*. If you go to this dataset, you will notice that it looks similar to the original dataset, but with 6,390 observations. This is because SPSS did not produce m=5 separate datasets, but merged them with the original observations in *Oddjob_imputed*, producing 6 · 1,065 = 6,390 observations. You will also notice a new variable *Imputation_* at the beginning of the variable list. This variable takes values from 0 to 5, which refer to the particular imputation session. The value 0 indicates the original dataset. The multiple imputation procedure automatically defines the *Imputation_* variable as a split variable when the output dataset is created.

Fig. A5.6 shows an excerpt of the dataset. The areas shaded in yellow are imputed values where the value was missing in the original dataset. At the bottom right of the screen, SPSS displays **Split by Imputation_**, indicating that the Split File command (see Chap. 5) is in effect.



	e1	s1	e2	s2	e3	s3	e4	s4	e5	s5	e6	s6
1235	100	93	100	64	100	79	100	94	64	65	96	61
1236	100	44	100	52	100	42	100	77	100	42	100	17
1237	76	50	91	44	84	52	75	33	75	33	68	48
1238	91	82	89	83	93	77	84	88	70	59	64	60
1239	79	77	81	62	.	101	92	86	70	65	62	58
1240	40	45	97	100	92	74	75	64	53	53	52	53
1241	100	88	100	71	100	66	91	77	66	68	72	70
1242	90	89	89	87	79	75	82	84	94	93	90	87
1243	100	48	100	44	100	68	100	34	100	42	100	52
1244	99	72	100	53	100	64	100	45	77	80	83	80
1245	84	58	98	67	.	77	93	69	51	45	30	18
1246	95	86	97	77	98	81	95	93	96	80	98	78
1247	100	96	100	83	100	50	100	63	54	53	100	14
1248	92	32	89	44	84	29	89	36	80	17	81	35
1249	50	50	50	50	.	68	50	50	.	35	50	50
1250	53	41	65	50	84	50	50	47	44	50	83	50

Fig A5.6 Imputed dataset

When initiating an analysis, SPSS now separately produces an output for the original dataset (where *Imputation_*=0) and the five imputed datasets. Many procedures also support the pooling of results from the analysis of multiply imputed datasets. On the Descriptive Statistics submenu of the Analyze menu, for example, Frequencies, Descriptives, Explore, and Crosstabs all support pooling. However, several of procedures that generally support pooling do not produce pooled results for all the statistics. For example, running a regression of *overall_sat* on *s1*-*s5* will produce the outputs in Tables A5.4 and A5.5. As you can see, the **ANOVA** output in Table A5.4 only shows the results of the original and the five imputed datasets, as indicated in the first column labelled **Imputation_**. Conversely, the **Coefficients** output in Table A5.5 also shows the pooled data's unstandardized coefficients, as well as their significances, at the bottom of the output.

As you can see, the differences in results between the original data and the pooled data are rather marginal. Even with regard to *s3*, which had the most missing values, the unstandardized coefficient differs only at the third decimal place, with no change in its significance. In the context of this regression analysis, these results suggest that we could likewise use the original data using listwise deletion.

ANOVA[a]

Imputation_	Model		Sum of Squares	df	Mean Square	F	Sig.
Original data	1	Regression	642.285	5	128.457	67.828	.000[b]
		Residual	1768.859	934	1.894		
		Total	2411.144	939			
1	1	Regression	784.153	5	156.831	81.974	.000[b]
		Residual	2026.057	1059	1.913		
		Total	2810.210	1064			
2	1	Regression	787.162	5	157.432	82.411	.000[b]
		Residual	2023.049	1059	1.910		
		Total	2810.210	1064			
3	1	Regression	760.894	5	152.179	78.640	.000[b]
		Residual	2049.317	1059	1.935		
		Total	2810.210	1064			
4	1	Regression	761.589	5	152.318	78.738	.000[b]
		Residual	2048.621	1059	1.934		
		Total	2810.210	1064			
5	1	Regression	769.212	5	153.842	79.823	.000[b]
		Residual	2040.999	1059	1.927		
		Total	2810.210	1064			

a. Dependent Variable: overall_sat
b. Predictors: (Constant), s5, s4, s3, s1, s2

Table A5.4 ANOVA table

Coefficients[a]

Imputation_	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
Original data	1	(Constant)	1.927	.135		14.282	.000			
		s1	.007	.003	.114	2.473	.014			
		s2	-.001	.003	-.018	-.349	.727			
		s3	.010	.003	.157	3.740	.000			
		s4	.001	.003	.019	.393	.694			
		s5	.024	.003	.330	9.336	.000			
1	1	(Constant)	1.903	.128		14.848	.000			
		s1	.006	.003	.104	2.481	.013			
		s2	-.001	.003	-.013	-.275	.783			
		s3	.011	.003	.173	4.466	.000			
		s4	.000	.003	.003	.074	.941			
		s5	.025	.002	.347	10.560	.000			
2	1	(Constant)	1.888	.129		14.677	.000			
		s1	.007	.003	.107	2.546	.011			
		s2	-.002	.003	-.025	-.518	.604			
		s3	.012	.003	.188	4.840	.000			
		s4	.000	.003	.008	.180	.857			
		s5	.025	.002	.338	10.310	.000			
3	1	(Constant)	1.920	.129		14.825	.000			
		s1	.007	.003	.109	2.575	.010			
		s2	.000	.003	-.007	-.138	.891			
		s3	.010	.003	.155	3.961	.000			
		s4	.001	.003	.019	.417	.677			
		s5	.024	.002	.335	10.190	.000			
4	1	(Constant)	1.931	.129		14.948	.000			
		s1	.006	.003	.100	2.378	.018			
		s2	.000	.003	.002	.035	.972			
		s3	.009	.003	.146	3.740	.000			
		s4	.001	.003	.013	.295	.768			
		s5	.025	.002	.346	10.507	.000			

5	1	(Constant)	1.915	.129		14.866	.000			
		s1	.007	.003	.107	2.524	.012			
		s2	.000	.003	.002	.052	.959			
		s3	.010	.003	.147	3.744	.000			
		s4	.000	.003	.003	.057	.955			
		s5	.025	.002	.350	10.609	.000			
Pooled	1	(Constant)	1.911	.130		14.694	.000	.019	.019	.996
		s1	.007	.003		2.492	.013	.007	.007	.999
		s2	-.001	.003		-.164	.870	.065	.068	.987
		s3	.011	.003		3.670	.000	.236	.279	.955
		s4	.001	.003		.202	.840	.028	.028	.994
		s5	.025	.002		10.215	.000	.043	.044	.992

a. Dependent Variable: overall_sat

Table A5.5 Regression coefficients table

Reference

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.