

2

Reading Data into Mplus

Mplus can process data in different formats. The two most relevant formats are discussed here: individual data and summary data. The *individual data* format is probably the most commonly used format in practice. Individual data are raw data in which the scores of all individuals on all variables are preserved. This is the case, for example, in our KFT data file (see Figure 1.11). The *summary data* format is used when one wants to analyze data in summary format, for example, a covariance or a correlation matrix (and potentially the means and standard deviations of the variables). The possibility to analyze summary data can be useful, for example, when a researcher wants to reanalyze data reported in a research report (e.g., a journal article, grant proposal). Many scholarly publications provide the covariance or correlation matrices that served as the basis for the analyses (e.g., path or structural equation models [SEMs]), whereas raw data are only rarely made available in publications. For many types of structural equation analyses, it is not necessary to analyze individual data (although using individual data may often be advantageous, e.g., when the user wants to take missing data into account or use specific estimation procedures, e.g., for clustered data). For many models, the covariance matrix (sometimes in combination with the mean vector) of the variables is sufficient as input for estimating the model. Let us first consider the case of reading individual data, using the data example from Chapter 1.

2.1 IMPORTING AND ANALYZING INDIVIDUAL DATA (RAW DATA)

Importing individual data into Mplus is usually unproblematic if the procedure described in Chapter 1 is followed. Nevertheless, it is recommended to check the proper import of the data into Mplus by first performing only basic statistical analyses (e.g., running descriptive statistics for all variables in the data set) and making sure that these statistics match the corresponding statistics calculated in at least one other program (e.g., SPSS). This approach can be used to check whether Mplus is reading the data correctly. This point cannot be emphasized enough. In practice, users often skip this step and then realize too late that Mplus has not processed the data correctly—leading to incorrect model estimates. For this reason users should take the time to first run a so-called **basic** analysis in Mplus to ensure the correct processing of the data before carrying out actual analyses in Mplus.

2.1.1 Basic Structure of the Mplus Syntax and BASIC Analysis

Mplus is an almost entirely syntax-based program. This means that the estimation of statistical models and many of the other functions in Mplus are executed via syntax commands and are not available through a point-and-click interface as, for example, in SPSS. Nonetheless, for the construction of the basic Mplus syntax, a point-and-click interface (the so-called *Mplus language generator*) is available. This option allows users to generate the most important basic syntax commands. Here I do not discuss the use of the Mplus language generator in detail, because users usually do not refer to this option any more after having learned the basic syntax rules in Mplus—which, fortunately, are not very difficult.

In the following material I demonstrate a useful strategy for reading data into Mplus and to check the correct processing of the data using the Mplus **basic** option. For this purpose we again refer to the sample data set **KFT.dat**. In the first step, the Mplus editor has to be opened (in MS Windows: **Start** → **Programs** → **Mplus** → **Mplus Editor**). An empty window becomes visible, into which one can either type the required syntax commands manually or use the Mplus language generator to get started. The syntax commands needed to run a **basic** analysis in Mplus on the KFT variables are shown in Figure 2.1.

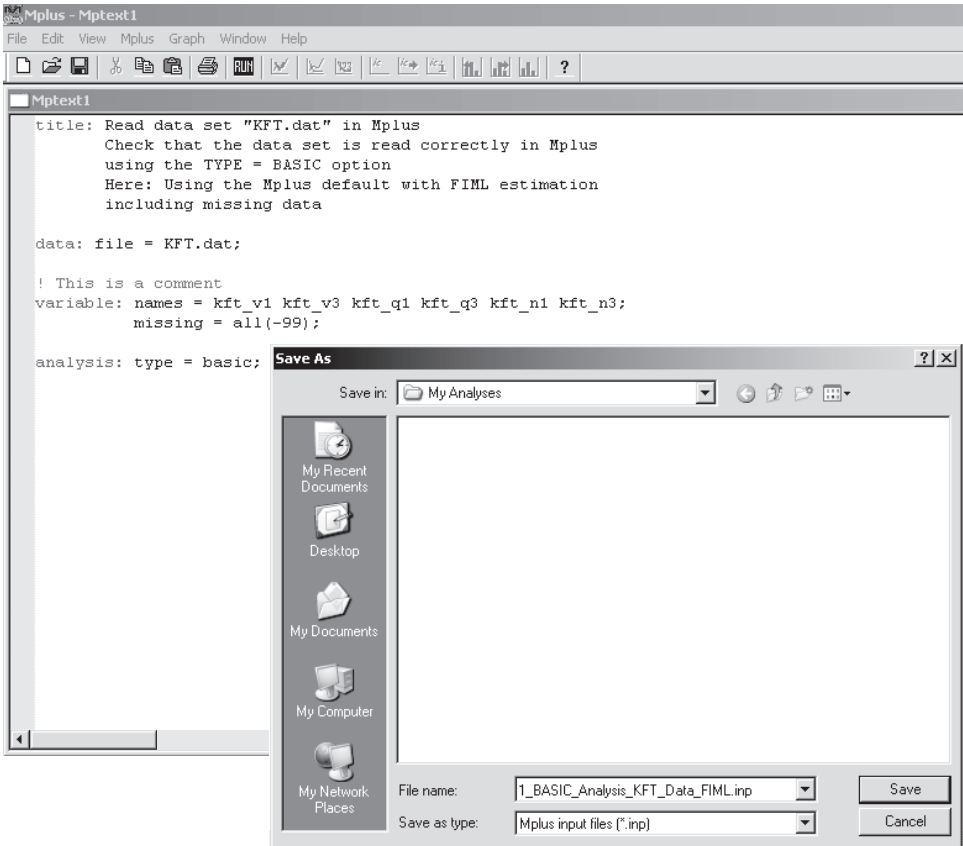


FIGURE 2.1. Mplus syntax file for checking the correct data import using the option `type = basic`. The menu option **File** → **Save as** is used here to save the input as a *.inp file. It is most convenient to save the input file to the same directory that also contains the data file to be analyzed. That way, no specific path has to be included under `data: file =`.

A new syntax file should immediately be saved as an Mplus input file (see Figure 2.1). In addition, when writing or editing a syntax file, one should not forget to save changes regularly, by either using the save symbol in the menu or hitting the combination CTRL + S so that work is not lost if the computer crashes, etc. Mplus input files contain the file ending *.inp. Unlike, for example, SPSS syntax, Mplus requires that a separate input file be specified for each statistical model.

The `title` command is used to label the analysis that is performed and to provide explanations on what statistical analysis is carried out (and

potentially in which way the specific model or analysis is different from others in a series of analyses). Although this command is not required, it is recommended to include a meaningful and informative title section so as to clearly document what has been done. An additional useful option is the use of comments. Each comment has to begin with an exclamation mark (!). No specific ending of the line is required, but every new line of comment has to begin with an exclamation mark again. Comments will appear in green font. Lines of comment that exceed 80 characters (in older versions of Mplus) or 90 characters (in newer versions) will be truncated at the end, so that comments may be incompletely shown in the output. For comments (unlike actual commands) this may not be critical; however, comments would also be incomplete on printouts, etc.

The `data` command is used to inform Mplus about the name, type, and location of the data file to be used. If the data set is saved in the same folder as the Mplus input file, no specific path to the data set needs to be specified in the Mplus input. The `variable` command is used to define the names of the variables in the data set as well as to define the missing value code (see also Section 1.1). Using the subcommand `names =`, the variable names are defined. It is important to list the variable names in the correct order in which they appear in the data set. Furthermore, it is important to know that in Mplus, variable names cannot be longer than eight characters. It is convenient to import variable names directly from SPSS using the SPSS option **Utilities** → **Variables**. This option is illustrated in Figures 2.2 and 2.3.

Another advantage of this procedure is that the variable names will be identical in both programs. Of course, this requires that variable names be defined that are fewer or equal to eight characters in SPSS as well. It is also important to check whether SPSS provides the variable names in the correct order when using the **Utilities** → **Variables** option. Furthermore, it is important to know that each subcommand in Mplus has to end with a semicolon (;). Missing semicolons are probably the most common cause of error messages in Mplus. Using the `missing` subcommand, we tell Mplus how missing values are coded. In our example, we coded missing values as -99 (cf. Section 1.1). We add the following additional subcommand under `variable`: `missing = all (-99);`. Using the command `analyses: type = basic;` we request descriptive statistics that we can then compare to the descriptive statistics in SPSS. By clicking on *run* (or hitting the combination CTRL + R), the basic analysis is executed. Box 2.1 gives an overview of some basic rules of Mplus syntax. Subsequently, we discuss the results of the **BASIC** analysis.

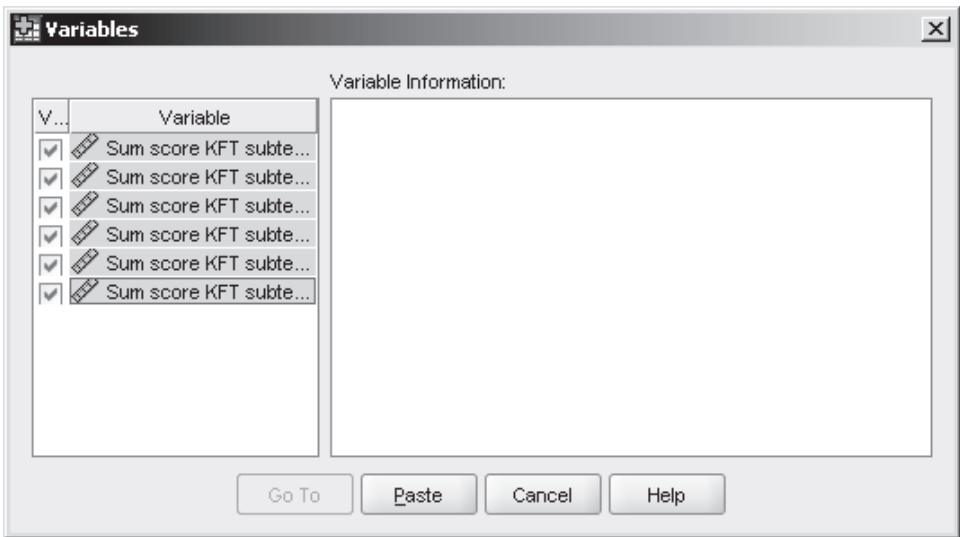


FIGURE 2.2. Exporting variable names from SPSS via the menu option **Utilities** → **Variables** in SPSS. On the left-hand side, all variable names are highlighted. By clicking on **Paste**, the variable names are added to an SPSS syntax window (see Figure 2.3).

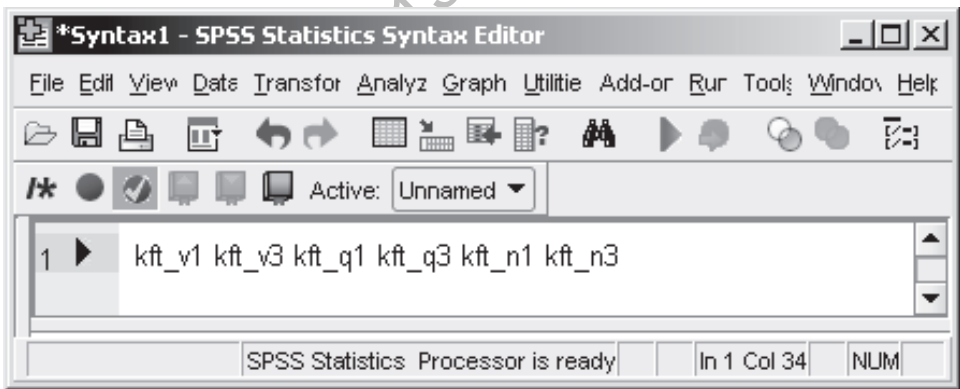


FIGURE 2.3. Variable names in the SPSS syntax window. The names can now be copied and pasted into an Mplus syntax under `variable: names =`. Note that the correct order of the variable names in the SPSS syntax file should be checked before transferring the names to Mplus.

BOX 2.1. Some Basic Mplus Syntax Rules

- In the Mplus syntax no differences are made between upper- and lower-case letters. (In other words, Mplus is not case sensitive.)
- In addition, the order in which different commands are listed is largely arbitrary.
- Every command line has to end with a semicolon (;).
- A single command line should not exceed 90 characters (in older Mplus versions, only 80 characters were allowed per line). Longer lines can occur, for example, when there is a long list of variables to read into Mplus or when a long file path has to be specified. Lines that are too long will be cut after 80/90 characters (depending on the program version), and Mplus will ignore all following specifications in that line. Clearly, this action can lead to very significant errors. Mplus will inform users about this problem in the output by means of an error message. One should not overlook or ignore this error message. A simple way to deal with the problem of overly long lines is to simply break the lines using the ENTER key on the keyboard.
- Variable names cannot be longer than eight characters.
- Each line of comment has to start with an exclamation mark. Comments appear in green font in the input and output. Comments do not have to end with a specific symbol; however, every new line of comment has to start with an exclamation mark again.

2.1.2 Mplus Output for BASIC Analysis

After running the input file for the **basic** analysis, a new window automatically pops up containing the output for the analysis. The corresponding output file is automatically saved to the same folder that contains the input file. The output file has the same file name as the input file but can be distinguished from the input file through the different file ending (*.out instead of *.inp). The output file for our **BASIC** analysis is shown in the following. First of all, the commands used to specify the analysis are reproduced in the output. This is useful because it allows us to review the input specifications to check if everything is set up correctly. In addition, that way, the input specifications will be included in any printout of the results.

```
Mplus VERSION 6.1
MUTHEN & MUTHEN
11/10/2011    4:36 PM
```

INPUT INSTRUCTIONS

```

title: Read data set "KFT.dat" in Mplus
      Check that the data set is read correctly in Mplus
      using the TYPE = BASIC option
      Here: Using the Mplus default with FIML estimation
      including missing data

data: file = KFT.dat;

! This is a comment
variable: names = kft_v1 kft_v3 kft_q1 kft_q3 kft_n1 kft_n3;
          missing = all(-99);

analysis: type = basic;

```

Next, we receive the following warning message, which is caused by the fact that in our data set, 131 students had missing values on *all* six variables:

```

*** WARNING
Data set contains cases with missing on all variables.
These cases were not included in the analysis.
Number of cases with missing on all variables: 131
1 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS

```

This message appears because Mplus, by default, uses FIML estimation with missing data (e.g., see Enders, 2010). This procedure cannot be used for participants who have no valid scores on any of the variables, because these individuals do not provide any information about the variables.

Following is the title and some technical information about the analysis with regard to, for example, the sample size, the variables used in the analysis, and the data structure in general:

```

Read data set "KFT.dat" in Mplus
Check that the data set is read correctly in Mplus
using the TYPE = BASIC option
Here: Using the Mplus default with FIML estimation
including missing data

SUMMARY OF ANALYSIS

```

Number of groups	1
Number of observations	456
Number of dependent variables	6
Number of independent variables	0
Number of continuous latent variables	0

```
Observed dependent variables

Continuous
  KFT_V1      KFT_V3      KFT_Q1      KFT_Q3      KFT_N1      KFT_N3

Estimator                                ML
Information matrix                      OBSERVED
Maximum number of iterations              1000
Convergence criterion                    0.500D-04
Maximum number of steepest descent iterations 20
Maximum number of iterations for H1        2000
Convergence criterion for H1              0.100D-03

Input data file(s)
  KFT.dat

Input data format  FREE
```

This information is useful to check, among other things, that the correct data set was used as well as the intended variables. For `Number of observations` we can see that 456 students contributed one or more values (here, those cases that have missing data on all six variables are already excluded). Next is a summary of the missing data patterns that occurred in the present application:

```
SUMMARY OF DATA

      Number of missing data patterns              2

SUMMARY OF MISSING DATA PATTERNS

      MISSING DATA PATTERNS (x = not missing)

      1  2
KFT_V1  x  x
KFT_V3  x  x
KFT_Q1  x  x
KFT_Q3  x  x
KFT_N1  x  x
KFT_N3  x

      MISSING DATA PATTERN FREQUENCIES

      Pattern  Frequency      Pattern  Frequency
           1           455           2           1
```

Under `SUMMARY OF DATA` we see that two distinct missing data patterns occurred in this example. Under `SUMMARY OF MISSING DATA PATTERNS` Mplus shows what those patterns were. Each missing data pattern is represented by a separate column (with the exception of

the pattern in which all values are missing, which is not shown). Column 1 (missing data pattern 1) contains only “x’s,” indicating that this is the pattern in which *no* missing data occurred. The second column contains “x’s” except for the last variable KFT_N3. This means that individuals with pattern 2 have valid scores on all variables except for the variable KFT_N3.

The section on MISSING DATA PATTERN FREQUENCIES shows that of the 456 students, all but one student contributed complete data (values on all six variables): 455 students showed missing data pattern 1 (frequency = 455), whereas only one student showed missing data pattern 2 (frequency = 1), that is, a missing value on KFT_N3.

The covariance coverage shows us the proportion of cases that contributes values for the calculation of each variance or covariance. In our case, for the variances and covariances, those are 100% of cases, with the exception of the variance and covariances that are associated with variable KFT_N3 (only 99.8% of cases contribute to this variable because of the one student who has a missing value on this variable). The minimum acceptable value for the covariance coverage is 10% according to the Mplus default (“Minimum covariance coverage value 0.100”). If the covariance coverage fell below this value, Mplus would no longer estimate a model by default because the coverage would be seen as too weak.

COVARIANCE COVERAGE OF DATA					
Minimum covariance coverage value 0.100					
PROPORTION OF DATA PRESENT					
	Covariance Coverage				
	KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1
KFT_V1	1.000				
KFT_V3	1.000	1.000			
KFT_Q1	1.000	1.000	1.000		
KFT_Q3	1.000	1.000	1.000	1.000	
KFT_N1	1.000	1.000	1.000	1.000	1.000
KFT_N3	0.998	0.998	0.998	0.998	0.998
Covariance Coverage					
KFT_N3					
KFT_N3	0.998				

Note that the covariance coverage does not refer to the actual variances and covariances of the variables, but instead simply informs us about the “completeness” of the data (the amount of data available to calculate those statistics). The actual estimated sample variances and covariances

are obtained next, under the header RESULTS FOR BASIC ANALY-
SIS/ESTIMATED SAMPLE STATISTICS:

RESULTS FOR BASIC ANALYSIS					
ESTIMATED SAMPLE STATISTICS					
Means					
KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1	KFT_N3
11.904	8.978	12.377	7.730	11.088	8.277
Covariances					
	KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1
KFT_V1	21.592				
KFT_V3	10.761	17.938			
KFT_Q1	6.010	5.267	11.235		
KFT_Q3	4.645	4.406	4.027	6.754	
KFT_N1	11.072	11.644	6.870	6.111	29.826
KFT_N3	6.347	7.190	4.415	3.986	9.298
Covariances					
	KFT_N3				
KFT_N3	11.782				
Correlations					
	KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1
KFT_V1	1.000				
KFT_V3	0.547	1.000			
KFT_Q1	0.386	0.371	1.000		
KFT_Q3	0.385	0.400	0.462	1.000	
KFT_N1	0.436	0.503	0.375	0.431	1.000
KFT_N3	0.398	0.495	0.384	0.447	0.496
Correlations					
	KFT_N3				
KFT_N3	1.000				
MAXIMUM LOG-LIKELIHOOD VALUE FOR THE UNRESTRICTED (H1) MODEL IS -7152.289					
Beginning Time: 16:36:14					
Ending Time: 16:36:15					
Elapsed Time: 00:00:01					
MUTHEN & MUTHEN					
3463 Stoner Ave.					
Los Angeles, CA 90066					
Tel: (310) 391-9971					
Fax: (310) 391-8971					
Web: www.StatModel.com					
Support: Support@StatModel.com					
Copyright (c) 1998-2010 Muthen & Muthen					

The descriptive statistics obtained for each of the six observed KFT variables (i.e., the estimated means, variances, covariances, and product-moment correlations of the variables) were estimated using the FIML procedure. Therefore, the statistics in this form are not directly comparable to the corresponding statistics calculated in SPSS. The reason is that SPSS statistics are based either on pairwise or listwise deletion of cases rather than on FIML estimation.

In order to directly compare Mplus and SPSS descriptive statistics, one can request the use of listwise deletion in Mplus rather than the default (FIML). Note that we discourage the use of listwise deletion in later analysis steps that involve actual model fitting. The procedure is used here for technical reasons only, in order to facilitate the process of checking the proper data entry into Mplus. In later analysis steps, researchers should consider using more advanced missing data analytic techniques, such as FIML, as described by Enders (2010) or by Schafer and Graham (2002), because these techniques are usually more reasonable than listwise deletion.

Listwise deletion of cases is obtained in Mplus by adding the subcommand `listwise = on;` under the `data` command. The `listwise = on;` subcommand deactivates the FIML procedure and excludes all cases that have missing data on at least one of the variables included in the analysis. Note that FIML has been the default since Mplus version 5. In previous versions of the program through Mplus version 4, listwise deletion was the default.

The extended Mplus syntax using listwise deletion is shown in Figure 2.4. The resulting descriptive statistics are shown below. For comparison, the corresponding SPSS statistics (with listwise deletion) are shown in Figure 2.5. The SPSS statistics were generated using the SPSS option **Analyze → Scale → Reliability Analysis → Statistics → Descriptives for Item Inter-Item Covariances/Correlations**. The results are identical to the Mplus results, rounded to three decimals, which shows us that the data set **KFT.dat** was apparently correctly read by Mplus. In addition, SPSS returns the same sample size ($N = 455$ listwise cases). After this check, we can start the first actual model in Mplus (see Chapter 3).

SAMPLE STATISTICS [Now based on listwise deletion of cases]					
Means					
KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1	KFT_N3
11.899	8.974	12.387	7.732	11.112	8.281

Covariances					
	KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1
KFT_V1	21.677				
KFT_V3	10.799	18.008			
KFT_Q1	6.057	5.310	11.242		
KFT_Q3	4.669	4.429	4.038	6.783	
KFT_N1	11.172	11.745	6.794	6.120	29.686
KFT_N3	6.383	7.230	4.417	4.001	9.294

Covariances					
	KFT_N3				
KFT_N3	11.811				

Correlations					
	KFT_V1	KFT_V3	KFT_Q1	KFT_Q3	KFT_N1
KFT_V1	1.000				
KFT_V3	0.547	1.000			
KFT_Q1	0.388	0.373	1.000		
KFT_Q3	0.385	0.401	0.462	1.000	
KFT_N1	0.440	0.508	0.372	0.431	1.000
KFT_N3	0.399	0.496	0.383	0.447	0.496

Correlations					
	KFT_N3				
KFT_N3	1.000				

The screenshot shows the Mplus software window titled "Mplus - [2_BASIC_Analysis_KFT_Data_listwise_deletion.inp]". The menu bar includes File, Edit, View, Mplus, Graph, Window, and Help. The toolbar contains icons for file operations, running, and viewing. The main text area displays the following input commands:

```

title: Read data set "KFT.dat" in Mplus
      Check that the data set is read correctly in Mplus
      using the TYPE = BASIC option
      Here: Using listwise deletion of cases

data: file = KFT.dat;
      listwise = on; ! This command turns FIML estimation off

variable: names = kft_v1 kft_v3 kft_q1 kft_q3 kft_n1 kft_n3;
          missing = all(-99);

analysis: type = basic;
  
```

The status bar at the bottom indicates "Ready" and "Ln 1, Col 1".

FIGURE 2.4. Modified Mplus input file for the basic analysis, now with FIML estimation turned off. This file generates descriptive statistics based on a listwise deletion of cases that can be more easily compared to results in SPSS.

Item Statistics						
	Mean	Std. Deviation	N			
Sum score KFT subtest kft_v1	11.8989	4.65585	455			
Sum score KFT subtest kft_v3	8.9736	4.24360	455			
Sum score KFT subtest kft_q1	12.3868	3.35293	455			
Sum score KFT subtest kft_q3	7.7319	2.60434	455			
Sum score KFT subtest kft_n1	11.1121	5.44845	455			
Sum score KFT subtest kft_n3	8.2813	3.43665	455			

Inter-Item Covariance Matrix						
	Sum score KFT subtest kft_v1	Sum score KFT subtest kft_v3	Sum score KFT subtest kft_q1	Sum score KFT subtest kft_q3	Sum score KFT subtest kft_n1	Sum score KFT subtest kft_n3
Sum score KFT subtest kft_v1	21.677	10.799	6.057	4.669	11.172	6.383
Sum score KFT subtest kft_v3	10.799	18.008	5.310	4.429	11.745	7.230
Sum score KFT subtest kft_q1	6.057	5.310	11.242	4.038	6.794	4.417
Sum score KFT subtest kft_q3	4.669	4.429	4.038	6.783	6.120	4.001
Sum score KFT subtest kft_n1	11.172	11.745	6.794	6.120	29.686	9.294
Sum score KFT subtest kft_n3	6.383	7.230	4.417	4.001	9.294	11.811

Inter-Item Correlation Matrix						
	Sum score KFT subtest kft_v1	Sum score KFT subtest kft_v3	Sum score KFT subtest kft_q1	Sum score KFT subtest kft_q3	Sum score KFT subtest kft_n1	Sum score KFT subtest kft_n3
Sum score KFT subtest kft_v1	1.000	.547	.388	.385	.440	.399
Sum score KFT subtest kft_v3	.547	1.000	.373	.401	.508	.496
Sum score KFT subtest kft_q1	.388	.373	1.000	.462	.372	.383
Sum score KFT subtest kft_q3	.385	.401	.462	1.000	.431	.447
Sum score KFT subtest kft_n1	.440	.508	.372	.431	1.000	.496
Sum score KFT subtest kft_n3	.399	.496	.383	.447	.496	1.000

FIGURE 2.5. Descriptive statistics for the six KFT variables produced in SPSS through the option **Analyze** → **Scale** → **Reliability** → **Statistics** → **Descriptive Statistics for Items/Inter-Items** → **Covariances/Correlations**. The values match the Mplus estimates obtained under the `listwise = on;` option.

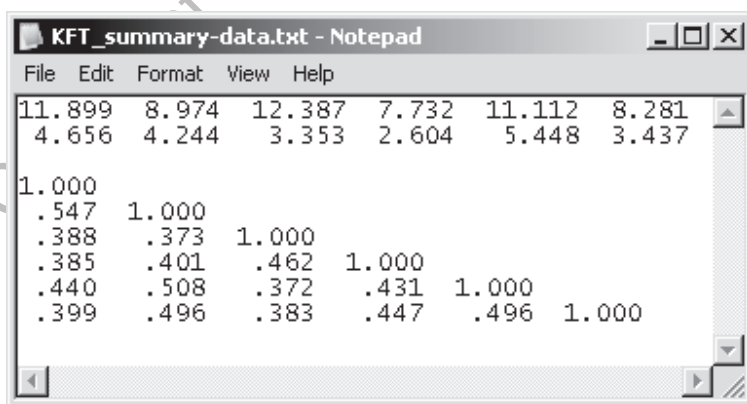
2.2 IMPORTING AND ANALYZING SUMMARY DATA (COVARIANCE OR CORRELATION MATRICES)

As already mentioned at the beginning of this chapter, it is sometimes practical to use summary data rather than individual data for an analysis. Summary data are, for example, covariance or correlation matrices, sometimes supplemented by the means and/or standard deviations of all variables. We now show the procedure of reading summary data into Mplus

using the six KFT variables as an example. The simplest way to enter summary data is to copy and paste the covariance or correlation matrix into a simple text file. For this purpose one can use, for example, the simple Editor program in Windows (**Start** → **Programs** → **Accessories**) or WordPad.

Figure 2.6 shows an example of a text file that contains the means (first row), standard deviations (second row), and product moment correlation matrix of the six observed KFT variables (the file is located on the companion website and is named **KFT_summary-data.txt**). The data shown in this text file can easily be used by Mplus to estimate, for example, an SEM. The Mplus syntax to read the data from the file **KFT_summary-data.txt** is shown in Figure 2.7. The **data** command is again used to define the name of the data set. In addition, it has to be specified what *type* of summary statistics are being read from this file. In our case the subcommand **type = means std corr**; means that the data set **KFT_summary-data.txt** contains the means (**means**) followed by the standard deviations (**std**) and correlations (**corr**) of the variables. Note that the order of the subcommands is important. (They have to be in line with the order in which the summary statistics appear in the data file.) To read a covariance matrix, one would use the subcommand **cova** instead of **corr**.

The additional subcommand **nobservations = 455**; is used to specify the sample size (number of individuals on which the summary data set is based). In this case, the data are based on $N = 455$ listwise



11.899	8.974	12.387	7.732	11.112	8.281
4.656	4.244	3.353	2.604	5.448	3.437
1.000					
.547	1.000				
.388	.373	1.000			
.385	.401	.462	1.000		
.440	.508	.372	.431	1.000	
.399	.496	.383	.447	.496	1.000

FIGURE 2.6. Text file with summary data for the six KFT variables. The first line contains the means of the variables. The second line contains their standard deviations. The matrix contains the product-moment correlations of the variables. Figure 2.7 shows how these data can be properly processed in Mplus.

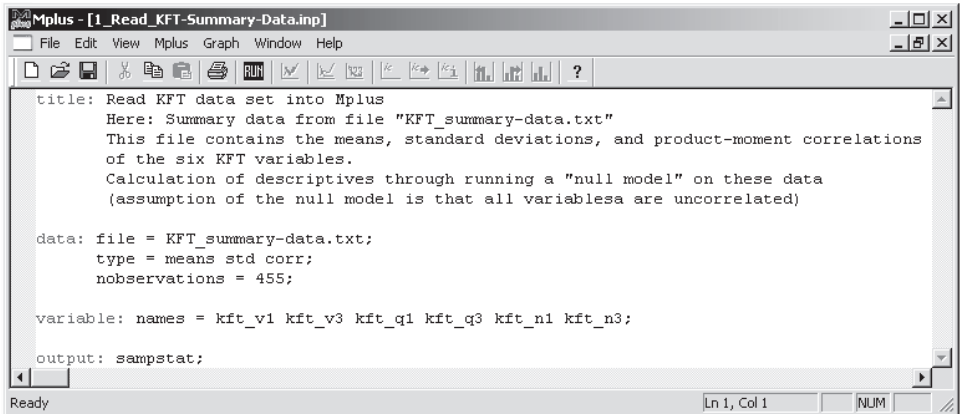


FIGURE 2.7. Mplus input file for reading the summary data shown in Figure 2.6. The `basic` option is not available for summary data. Therefore, descriptive statistics for data checking are requested via the `output: sampstat;` option. Given that we do not explicitly specify a model for the data, by default Mplus estimates a so-called *null model* for all variables listed under `variable: names =`.

cases. This information has to be provided, because it is not possible for Mplus to infer the number of observations from the summary data set (whereas this is possible for individual data). The option `variable: names =` is again used to define variable names—which also are not given in the data file.

The `BASIC` option is not available for summary data in Mplus. For this reason, descriptive statistics for data checking are now requested using the option `output: sampstat;`. The abbreviation `sampstat` stands for *sample statistics* and provides us with descriptive statistics for the six variables (in this case, the observed means and the covariance matrix will be provided). Given that no model is explicitly specified, by default Mplus estimates a so-called *null model* (sometimes referred to as an *independence model*). The null model assumes that there are no relationships among any of the six variables. The only model parameters to be estimated in this model are the means and variances of the observed variables. The output for this analysis is not shown here but can be found on the companion website.

The next chapter provides an introduction to the basics of model specification in Mplus using linear SEMs as an example. I first consider simple linear regression models and subsequently more complex SEMs such as confirmatory factor analysis and latent path analysis.