7EU VET Project

Work Package 7

Attachment 7.35

Manual for Entering and Cleaning Data

Number of pages: 7

Project name: Detailed Methodological Approach to understanding the VET educational system in 7 European countries

Project Number: 505480-LLP-1-2009-1-SI-KA 1-KA 1SCR

Grant Agreement: 2009 – 12029/001-001

Education and Culture DG
Lifelong Learning Programme

7 EU
Detailed Methodological Approach to Understanding Vocational Education and Training Systems

This project has been funded with support from the European Commission. This communication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Manual for Entering and Cleaning Data

Before you can enter the information from every questionnaire into SPSS, you have to get familiar with the codebook. The codebook summarizes all instructions and information you will use to convert the information obtained from each student into a format that SPSS can understand. The codebook has been made up of our master questionnaire. The labeling of the variables in the dataset corresponds to the variable names in the questionnaire. For example, the first question in the questionnaire A1 is labeled as A1 in the dataset and so on. Questions in the questionnaire which produce more than one variable per question are numbered chronologically, because each question or item must have a unique variable name (A4_1, A4_2, A4_3 and so on).

The master questionnaire might not be identical in all records with your national version so please check where differences occur. Questions or items that exist in the master questionnaire but not in your national version can be kept blank. Please always use the codes given in the master questionnaire and do not change them.

You find every variable name and the information needed for coding the questions in the codebook. Before you can begin to enter the questionnaire you have to fill in some variables regarding country, schools, classes and students.The first variable in our dataset is *Country.*

Please enter your country number for every case. Next, you have to enter the *SchoolID*. The *SchoolID* consists of

- one digit country number (for example 3 for Greece)
- three digits school ID (for example 402).

Country numbers:

1 - Lithuania

2 - Austria

3 - Greece

Following our example the *SchoolID* for school 402 in country 3 would be 3402.

The SchoolID is identical to the School Tracking ID in the field work monitoring tool.

In the next variable *Class* please fill in the class number (e.g. 1, 2 or 3). It is not necessary to type zero into the class variable.

Please enter the *StudentID* - the unique number that identifies each student.

The *StudentID* consists of

- one digit country number (for example 3 for Greece)
- three digits school ID (for example 402)
- two digits class number (for example **0**1 or **0**2 or **0**3) Attention: the class number must have two digits in the variable *StudentID*
- two digits student number (for example 04)

Following our example the final ID for one student would be 34020204.

Every partner also has to enter the teacher/headmaster questionnaire "class characteristics" into the provided entry mask. There you also have to enter the *SchoolID* and the *class*.

Coding responses and question types

Each response must be assigned a numerical code before it can be entered into SPSS.

In our questionnaire there are closed, partial open and open-ended questions. Closed questions have to be converted to the numerical format that is required for SPSS. Every possible response is numbered. You find the corresponding numbers in the provided codebook.

For example, *Males* are coded as 1, *Females* as 2 (Question G1). Closed questions may also involve a range of different choices (answer categories). In Question A6 ('*Have you considered any alternative programme when you were selecting your current one?*'), the number corresponding to the response ticked by the respondent would be entered. For example, if the respondent ticked '*I was considering one alternative programme'*, this would be coded as 2. If the respondent ticked '*I was considering more than three alternative programmes'*, this would be coded as 5.

In questions with multiple answer options every answer option will be treated as own variable. If the respondent ticks an answer, the variable will be coded as 1 ("Quoted"). If the respondent did not tick the answer option, it will be coded as 0 ("Not quoted"). For example, if the respondent ticked '*I have difficulties with some of my teachers'* (Question B9), variable B9_3 would be coded as 1. If the respondent didn't tick '*I have problems with some of my school-mates'*, variable B9_4 will be coded as 0.

In some questions there is a combination of both closed and open-ended questions (a number of defined responses and an additional category *Other* that the respondents can tick if the response they want to give is not listed). In this case the respondent can write the answer he or she wants to give on the provided line. Please try to code the given open answers in one of the provided answer categories. If a given answer cannot be categorized into one of the given categories they can stay blank. Coding for partial open questions is not provided by WP7.

There are also some open-ended questions. Responses to open-ended questions have to be converted into the given coding schemes: Question B2a (*title of programme*) and D4 (*expected job at the age of 30*) into ISCO 88 (two-level) and question G3c (*citizenship*) into the provided list of nationalities. You find both coding schemes at the end of the codebook and in the data entry mask (variables: B2a_ISCO, D4_ISCO, G3c_nation). For some background information about ISCO 88 classification, please check:

- http://www2.warwick.ac.uk/fac/soc/ier/research/isco88/english/
- http://www2.warwick.ac.uk/fac/soc/ier/research/isco88/english/groups/

When entering the data for each student, you compare his or her response with those listed in the codebook and enter the appropriate number into the data set under the variable B2a_ISCO, D4_ISCO or G3c_nation. You can keep the original (uncoded) programme names, expected job titles and citizenships in your national dataset. In the international data set there will be only the coded categories. Therefore, job titles don't need to be translated in English.

Open numeric questions like the duration of the programme or grades (e.g. B2b, B3_1) accept exclusively numerical answers. But students often give additional information if they are free to do so, e.g. about 3 years, circa 3 years, between 3 and 4 years. Please round the information up or down and only enter numbers in the entry mask and do not fill in the comment text. If the students answer with a time range (e.g. between three to four years), please fill in the arithmetic mean. Coding rules for this type of variables:

- *age* of the students must be coded as a whole number (e.g. 17) (G2_age, G3b)
- *grades* could be coded with one decimal number (C1a_1a, C1a_2a, C1a_3a, C1b)

- *duration* of the programme (B2b) should be coded in half-year intervals (2, 2.5 , 3, 3.5)
- *school hours* spend at school (B3_1)/ *hours worked* per week (C6b_1) should be coded in half-hour intervals (e.g. 3.5, 40)
- *years* learning the language (E2c) should be coded in half-year intervals
- *siblings* (G5) must be coded as a whole number
- *household size* (G6) must be coded as a whole number

In general, you can correct the spelling of an open question when entering it into the dataset.

You don't need to code or translate the comment box at the end of the questionnaire for international dataset because the comment box will be skipped.

Missing values:

A missing value is a number that indicates to SPSS that the response is not valid and should not be included in the analysis. We differentiate between three types of missing values: 1) missing values where we have skip or filter questions that result in some questions being not applicable to all respondents; 2) questions where a respondent simply did not give an answer; 3) questions that allow for '*Don't know'.* In all of these cases we need to enter a missing value into the relevant cell.

- -77 = Not asked / Filter (Not applicable) → Values that are missing due to filter questions will be coded as -77
- -88 = Not answered → Values that are missing due to refusal will be coded as -88
- -99 = Don't know → If a respondent ticks the answer option '*Don't know*', this will be coded as -99

If respondents do not give an answer to open ended or partial open ended questions you don't need to enter a missing value in the relevant cell. Just leave the cell blank.

Entering data

To enter data you need to have the Data View active (to activate the Data View you have to click on the Data View tab at the bottom left-hand side of the screen of the Data Editor window). A spreadsheet appears with the already defined variable names listed across the top. Down the side you see the numbers 1, 2, 3 and so on. These are the case numbers that SPSS assigns to each of your lines of data.

Click on the first cell of the data set (first column, first row). Type in your country number. Press the right arrow key on your keyboard, this will move the cursor into the second cell. Move across the row, entering all the information for case 1. Please make sure that the values are entered in the correct columns. Afterwards, move to the second row and enter the data for case 2. Remember to save your data file regularly.

Quality control

Around 1000 questionnaires should be coded by 3 to 4 coders. It is of essential importance to train the coders adequately. For quality control 10% of the questionnaire should be entered twice. Therefore you receive an additional data entry mask. Please send us both data entry masks so we can check for reliability. If there is not enough consistency we will come back to you. The questionnaires which are entered twice should be randomly chosen. We also suggest to pay coders by hour and not by number of questionnaires.

Cleaning the data

Before you send the data set to us or start to analyse your data, it is essential that you check your data set for errors.

Step 1: Checking for errors

You need to check each of the variables for scores that are out of range (values that fall outside the range of possible values for a variable). For example, gender is coded 1=male, 2=female, there should not be any scores other than 1 or 2 for this variable.

To check for errors you will need to inspect the frequencies for each variables.

Procedure for checking categorical variables:

1) Check your Minimum and Maximum values: Do they make sense? Are they within the range of possible scores on that variable?

2) Check the number of valid and missing cases. If there are a lot of missing cases, you have to look why. Have you made errors in entering the data (for example put the data in the wrong columns)?

Syntax for this procedure:

```
FREQUENCIES

VARIABLES= G1_gender G4 G7.
```

Procedure for checking continuous variables:

1) Check your Minimum and Maximum values: Do they make sense? Are they within the range of possible scores on that variable?

2) Does the Mean score makes sense? If there is an out-of-range value in the data file, this will distort the mean value.

Syntax for this procedure:

```
FREQUENCIES

VARIABLES= G2_age.
```

Step 2: Finding and correcting the error in the Data file

If you find some out-of-range responses in the dataset you need to find where in the data file this error occured or which cases are involved. Thereafter you have to correct or delete the value. Ways to find an error in the dataset:

For example, gender should have no other scores than 1 and 2. If you find that the maximum score is three you have to check. Therefore you can sort cases by variables.

Syntax for this procedure:

```
SORT CASES BY G1_gender (D).
```

With (A = ascending) or (D = descending) you decide if you want the higher values at the top or the bottom. For gender, we want to find the person with the value 3, so we choose descending. The case with the error should afterwards be located at the top of your data file.

Look across to the gender variable column. Check the studentID from the case listed with the error and try to find out in the original questionnaires what the correct score is.

The general procedure to select specific cases. Syntax for this procedure:

```
TEMPORARY.

SELECT if variable=value.

FRE StudentID.
```

All alterations to the data set should carefully and clearly be documented, for instance in a syntax or a log. This procedures are only univariate. Further bivariate analysis is needed.