**PCO ‘22**

**Data workshop 1: Data transformation with the WVS time series data.**

Before quantitative social scientists can answer research questions and test hypotheses, they need to prepare data in such a way that they can be answered. And for that they need to truly grasp the concepts they want to measure. This phase, data manipulation, will be central in the first data workshop. The book we read in class is based on the WVS data and the results of the post-materialist module which includes postmaterialist indices (4 items and 12 items). The immediate goal of this workshop is to recreate the four item index and to compute post-materialism scores for each region and country in the WVS data. The more general goals is to understand what postmaterialism is about and how it differs across and between countries.

**Contents**

[**Section 1: Getting started** 2](#_Toc103166899)

[**Section 2: Getting familiar with the data.** 3](#_Toc103166900)

[**Section 3: Four item index** 4](#_Toc103166901)

[**Section 4: Creating Postmaterialism Scores** 6](#_Toc103166902)

# **Section 1: Getting started**

First of all, be sure to have SPSS installed on your computer and download the data from Brightspace (WVS\_TimeSeries\_spss\_v1\_6.sav). If you do not have SPSS installed please find a partner to work with who has SPSS installed. For those who want to know where the data comes from. Please check this website: <https://www.worldvaluessurvey.org/wvs.jsp>

Before we start with a project in SPSS, we need to set a working directory. This means that SPSS will use a specific directory on your computer as default for importing and exporting datafiles. Once you set this you do not have include the whole pathname in the get file or save file commands. Preferably a separate directory in your PCO directory on your laptop/pc in which you store the data for this data workshop. For this procedure we can use the following code.[[1]](#footnote-1) Open the syntax window in SPSS and copy paste the ‘blue’ spss code. You need to add the path to your directory.

*\*\*\*\*PCO DW1\*\*\*\*.*

*\*set working directory.*

cd ’your custom path’.

*\*I always include this line to clean my global environment when I work.*

*\*This makes it possible to do a “ctrl-a, ctrl-r” procedure.*

*\*This means rerunning the script to make sure you use clean data.*

*\*A lot of problems in data manipulation occur by reusing ‘saved’ data.*

*\*This is bad practice. If you make tidy and clean code, you normally do not need to save data ever.*

*\*It also makes your work reproducible. Unless your data is huge and computations take up a lot of time.*

dataset close all.

*\*import the WVS longitudinal file and rename it to wvs\_long.*

get file 'WVS\_TimeSeries\_spss\_v1\_6.sav'.

dataset name wvs\_long.

dataset activate wvs\_long.

**Task 1:** Set working directory, import the data, and create a dataset name now.

# **Section 2: Getting familiar with the data.**

A second step in data analysis, especially if you just started with a new dataset, is to look at the data and see how it is structured. The data we use is the WVS time series file. This is nothing more than all seven rounds of WVS data stacked into one harmonized file. If you look more closely at the data you can find some evidence for this, which you are going to do.

(hint: use the [*descriptives*](https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=reference-descriptives) and [*frequencies*](https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=frequencies-overview-command) commands).

**Task 2:** How many respondents does the WVS data include in total?

**Task 3:** What is the number of respondents for each WVS round?

**Task 4:** What is the number of respondents for each country? (hint: use the ISO code variable). Does this look unusual to you? If so, can you clarify what is unusual.

In the answer of T4, we can clearly see that the observations for each country differ greatly. The most logical clarification is that not every country participates in every round. With longitudinal data this is something you always have to check, which you can do with a simple table. The [*crosstable*](https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=reference-crosstabs) and frequency table are your biggest friends when checking data (for checking a recode/compute for example).

**Task 5:** Perform a cross table command. Compare this table to the table of participating countries in the appendix of the Inglehart book (p.393). What is the core difference?

# **Section 3: Four item index**

One of the main workhorses of research into postmateralism is the four item index. It is probably less accurate than the 12-item, but it contains fewer items so most surveys contain the four item index. Inglehart is so kind to share his code with us and the original item specification (they can both be found in appendix 4 of the book).

So, it should not be too hard to replicate them right?

**Task 6:** Identify the items in the WVS that measure post-materialist attitudes and the preprogrammed four item index. Run frequency tables. (hint: you can even use ctrl-f in SPSS).

Now we have identified the correct variables, we can create the four item index with the code from the book. Below we provided the code (but without the variables, this is something you need to do).

\**use the syntax in the appendix to create the four item index.*

*\* Use Y002 to check your coding. (preprogrammed postmat variable (four item)).*

compute V1000 = 2.

if ((? = 1 AND ? = 3) or (? = 3 AND ? = 1)) V1000 = 1.

if ((? = 2 AND ? = 4) or (? = 4 AND ? = 2)) V1000 = 3.

execute.

*\*add value labels.*

value labels V1000

1 'Materialist'

2 'Mixed'

3 'Postmaterialist'.

*\*check coding.*

frequency V1000 Y002.

**Task 8:** Fill in the right variables and create the four item index.

Uh oh, there seems to be a problem with just applying the code from Inglehart. The two variables should be exactly the same, but they are not. Probably the code that is provided does not work that well.

**Task 9:** What do you think is the problem. How would you solve it?

**Task 10:** Solve the issue. Check if the variables are correct.

# **Section 4: Creating Postmaterialism Scores**

In this section we will use the four item index to create post-materialism scores, which you just created (of found in the data set some other way).

One way to create post-materialism scores for each country/year or country year combination is by running the compare means function in SPSS. This has one major drawback, namely: The result is just a table. So if we want to have the mean scores per country/year or the combo thereof (like what Inglehart often presents) we need to copy it into excel or some other program. These are very tedious processes which you always should try to avoid. So, since our friends at IBM also do not like to copy-paste things, they have made a function for it:

[*Aggregate*](https://www.ibm.com/support/pages/how-do-i-aggregate-spss)

What this function does is fairly simple and straightforward, but oh so powerful. It groups the data into subsets based on the values of a given variable. Then it performs calculations on those subsets and stores the results.

For our purpose we use the variable country-year combination: S025. Aggregate can store the newly created variables as variables in the dataset.

\*use aggregate to calculate mean scores for each country-year combo. save it as a variable.

aggregate outfile \* mode addvariables

/break S025

/mean\_pm4\_cy = mean(V1000).

**Task 9**: run the code and command SPSS to make a frequency table of the new variable.

**Task 10:** Find the country year combination with the highest mean. Does this make sense to you?

However, what we did above is calculating a mean score per country or country year. That is not what Inglehart did in the book. What we want is to create a score of post-materialism, as provided by Inglehart, which is %Post-Materialist - %Materialists. So, how do we go about this? First we apply the procedure for one country-year combination: the Netherlands. After that we can try to scale this to all the countryyear combinations. The code for doing this:

\*simple way to create postmataterlist score for a countryear combination. Just only do this for country year combination: 5282012 \*Netherlands (2012).

\*first create a filter variable.

compute filter\_dutch = 0.

if (S025 = 5282012) filter\_dutch = 1.

execute.

\*use filter.

filter by filter\_dutch.

frequencies V1000.

filter off.

**Task 11:** calculate the postmateralist score for the Netherlands in 2021. Use the four-item variable you have created.

There are different ways of scaling this method to the whole dataset. A simple (but cumbersome) way is to [split](https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=reference-split-file) the file on country-year combinations and run frequency tables. Subsequently we can calculate, by hand, for each of the country-year combinations the score. This quickly becomes a lengthy endeavor. Nevertheless, this is the code:

\*you can scale this method by using split file.

\*for split file you need to sort cases.

sort cases by S025.

\*use split file.

split file by S025.

\*frequency command. Split file will now perform this on all subsets.

frequencies V1000.

\*turn off split file.

split file off.

**Task 12**: Run the code above.

I advise you not to perform the previous procedure in full, it is tedious and computers are invented to automate the boring stuff, and in doing so save precious time. Fortunately you have just learned about the aggregate command. This provides an easy way of automating this process, which saves a lot of your precious time. The code below will do just that. The aggregate command has two really nice functions.

PGT(varlist,value) **P**ercentage of cases **g**reater **t**han the specified value.

PLT(varlist,value) **P**ercentage of cases **l**ess **t**han the specified value.

So we can let SPSS do the calculations for us in an aggregate command. Unfortunately, we cannot mix these in one formula, so we need to create two separate variables %post-materialists and %materialist. After creating the aggregated data we can use these two variables to create a post-materialist value for every country-year combination.

Now that is very effective!

\*use aggregate to create new aggregated data.

aggregate outfile \* mode addvariables

/break S025

/p\_postmat\_fouritem\_cy = PGT(V1000, 2)

/p\_mat\_fouritem\_cy = PLT(V1000, 2).

\*now we can actually create the postmat variable for every country year combination.

compute postmat\_cy = p\_postmat\_fouritem\_cy - p\_mat\_fouritem\_cy.

exe.

\*frequency table of postmat\_cy.

frequencies postmat\_cy.

**Task 13:** Run the code above.

An important part of the intergenerational theory of value change is cohort replacement. Older cohorts are replaced by newer cohorts which eventually leads to societal level change. We are going to apply the procedure you have just learned to create post materialism scores of each cohort in the data. You can create these for each and every country-year combination, by adding two variables in the /break section of the aggregate function.

An important part of the intergenerational theory of value change is cohort replacement. Older cohorts are replaced by newer cohorts which eventually leads to societal level change. We are going to apply the procedure you have just learned to create post materialism scores for each cohort in the data. You can create these for each and every country-year combination, by adding two variables in the /break section of the aggregate function. Below the code for creating a cohort variable is given.

\*compute cohort variable.

compute cohort = -99.

if (X002 >= 1886 and X002 <= 1905) cohort = 1.

if (X002 >= 1906 and X002 <= 1915) cohort = 2.

if (X002 >= 1916 and X002 <= 1925) cohort = 3.

if (X002 >= 1926 and X002 <= 1935) cohort = 4.

if (X002 >= 1936 and X002 <= 1945) cohort = 5.

if (X002 >= 1946 and X002 <= 1955) cohort = 6.

if (X002 >= 1956 and X002 <= 1965) cohort = 7.

if (X002 >= 1966 and X002 <= 1975) cohort = 8.

if (X002 >= 1976 and X002 <= 1985) cohort = 9.

if (X002 >= 1986 and X002 <= 1995) cohort = 10.

if (X002 >= 1996 and X002 <= 2005) cohort = 11.

execute.

\*set labels.

value labels cohort

1 '1886 - 1905'

2 '1906 - 1915'

3 '1916 - 1925'

4 '1926 - 1935'

5 '1936 - 1945'

6 '1946 - 1955'

7 '1956 - 1965'

8 '1966 - 1975'

9 '1976 - 1985'

10 '1986 - 1995'

11 '1996 - 2005'.

\*set missing variables

missing values cohort (-99).

\*check the new variable.

frequencies cohort.

\*crosstab check.

crosstabs cohort by X002.

\*list check. Another nice way to check the compute command.

list

/variables = cohort X002

/cases = from 1 to 100.

**Task 14:** Use the code above to create a cohort variable.

**Task 15:** Use the aggregate command to create postmaterialism scores for each cohort-year combination and for each cohort-year-country combination.

Finally we want to export the data. Save a subset of the data as a separate dataset called 'wvs\_aggregated\_postmat.sav'. Also save the complete dataset with a different name than the original. Else you cannot reproduce the results!

Save outfile = 'wvs\_aggregated\_postmat.sav'

/keep S020 S002 S003 S025 postmat\_cy postmat\_cohort\_cy postmat\_cohort\_y cohort.

Save outfile = 'wvs\_dataprepped\_v1.sav'

1. If, after 5 minutes you have not mastered it, just open the data file by double clicking it. [↑](#footnote-ref-1)