

# Wide and Long Datasets

Generally, there are two types of datasets: *wide* and *long*. Different types of statistical analyses will require different data formats (e.g., long or wide). And sometimes companies will provide you with data in stupid formats (e.g., a long dataset when a wide one obviously should have been used). Thus, it's crucial that you know how to convert between wide and long datasets.

## Wide Data Formats

Wide data formats are what you're probably traditionally used to seeing. Each variable is a column. Each individual participant is a row.

Each column is a *variable* (e.g., question in your survey)

Each row is an individual participant's data

	male	age	single	married	anxietyg0 1	anxietyg0 2	anxietyg0 3
1	1	26	1	0	2	3	2
2	0	20	1	0	1	1	1
3	0	21	0	0	2	2	2
4	0	30	0	0	2	3	3
5	1	25	0	0	2	2	3
6	0	25	0	1	2	3	4
7	0	20	0	0	3	4	3
8	1	30	0	0	4	4	5
9	0	28	1	0	4	2	4
10	0	26	0	0	5	4	4

Participant #9 is 28 years old

Participant #7 has a score of 3 on the "anxietyg01" question

Here's a different example of a wide data format where employees' work satisfaction has been measured three separate times:

	employee_id	work_satisfaction.1	work_satisfaction.2	work_satisfaction.3
1	30490	5	5	5
2	30491	5	4	4
3	30492	5	4	4
4	30493	5	5	5
5	30495	4	.	.
6	30496	5	5	4
7	30497	4	3	4
8	30498	2	2	2
9	30499	4	4	4
10	30500	4	4	4

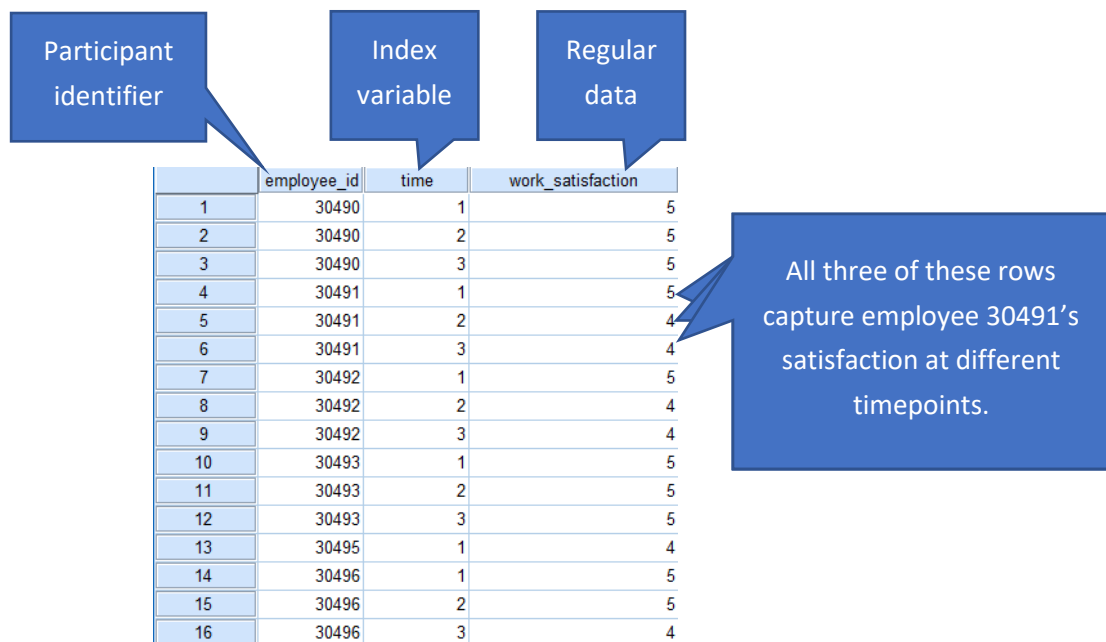
Notice that when the data are in wide format, it's trivially easy to correlate, for example, Time 1 responses with Time 2 responses (e.g., to estimate test-retest reliability):

```
correlations work_satisfaction.1 work_satisfaction.2.
```

However, what if you want to estimate how work satisfaction is changing across time? In order to do so, you'll need to convert your data to *long format*.

## Long Data Formats

In a "long" data format, each participant has multiple rows in the dataset.



	employee_id	time	work_satisfaction
1	30490	1	5
2	30490	2	5
3	30490	3	5
4	30491	1	5
5	30491	2	4
6	30491	3	4
7	30492	1	5
8	30492	2	4
9	30492	3	4
10	30493	1	5
11	30493	2	5
12	30493	3	5
13	30495	1	4
14	30496	1	5
15	30496	2	5
16	30496	3	4

Notice that in this dataset, each participant has one row *per measurement occasion*. The data are equivalent to the above example (e.g., employee 30490 reported work\_satisfaction of 5 on all three occasions, whereas employee 30491 reported work\_satisfaction of 5, 4, and 4 across the three respective timepoints); however, the data format has changed slightly.

Long datasets allow you to easily correlate variables with the *index variable*. In this case, *measurement occasion*, or *time* is the index variable. Thus, we can easily examine how work satisfaction changes across time using the following syntax:

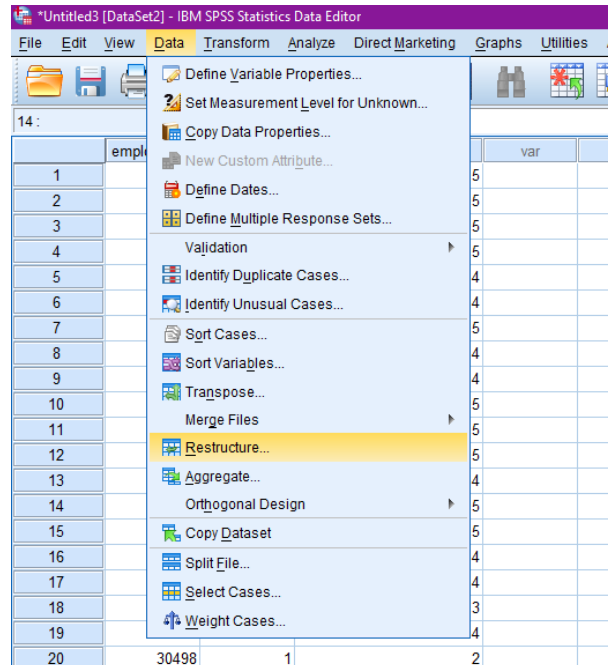
```
mixed work_satisfaction with time  
/fixed=time  
/random=intercept|subject(employee_id)  
/print=solution.
```

We have to use mixed regression to control for within-persons dependencies in the data.

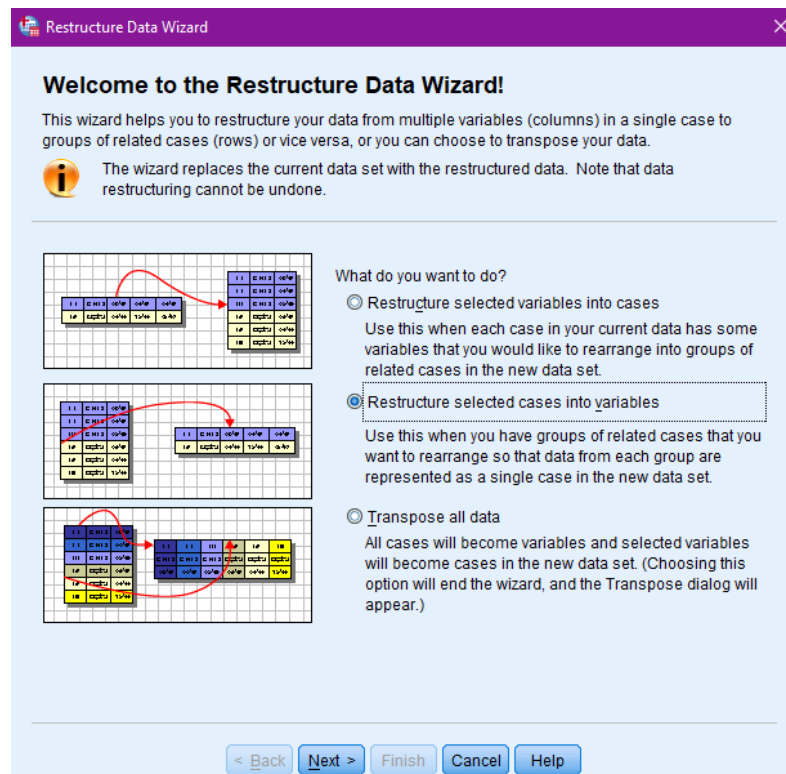
## Converting Long Datasets into Wide Ones

Converting long datasets into wide ones is easy.

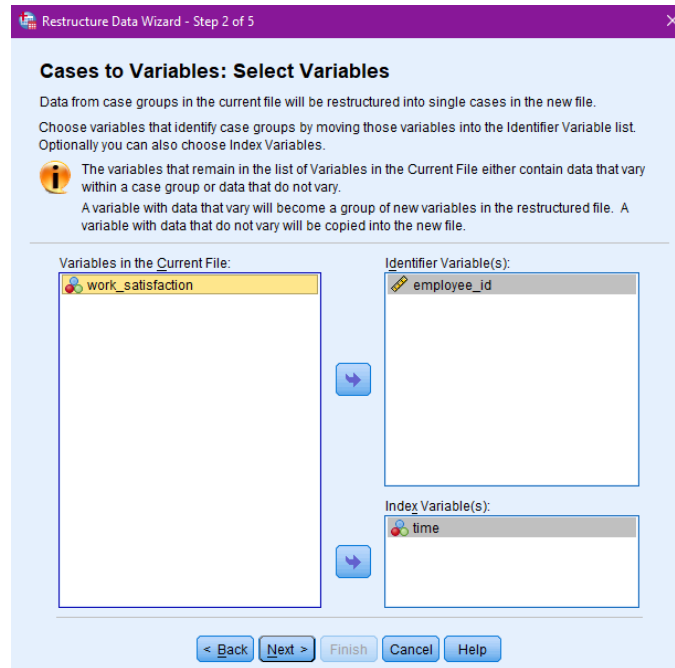
1. In SPSS's main menu, select Data > Restructure



2. Select "Restructure selected cases into variables"



- Put your participant identifier variable into the “Identifier Variable(s)” box, and put your index variable (“time” in this case) into the “Index Variable(s)” box, and click “Next”



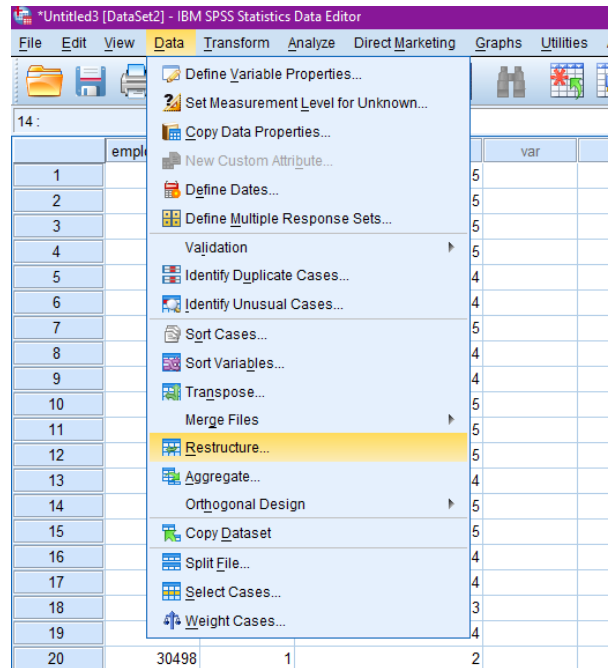
- Click “Finish.” Your data have been successfully converted into “wide” format. You’re ready to correlate “work\_satisfaction.1” with “work\_satisfaction.2” to estimate test-retest reliability!

	employee_id	work_satisfaction.1	work_satisfaction.2	work_satisfaction.3
1	30490	5	5	5
2	30491	5	4	4
3	30492	5	4	4
4	30493	5	5	5
5	30495	4	.	.
6	30496	5	5	4
7	30497	4	3	4
8	30498	2	2	2
9	30499	4	4	4
10	30500	4	4	4
11	30501	4	4	4
12	30502	5	5	5
13	30503	3	3	4
14	30504	4	4	4
15	30505	2	4	4
16	30506	4	4	4

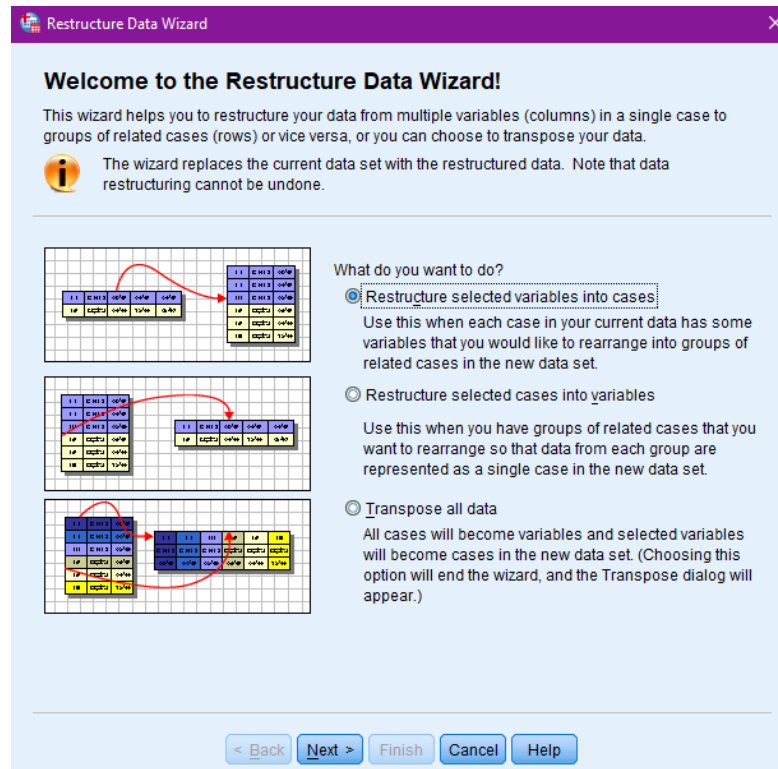
## Converting Wide Datasets into Long Ones

Converting wide datasets into long ones is slightly more complex.

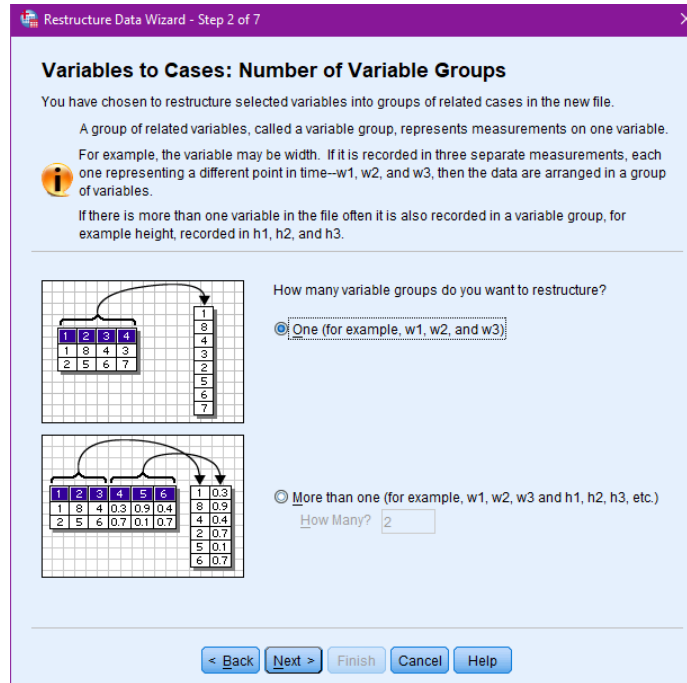
1. In SPSS's main menu, select Data > Restructure



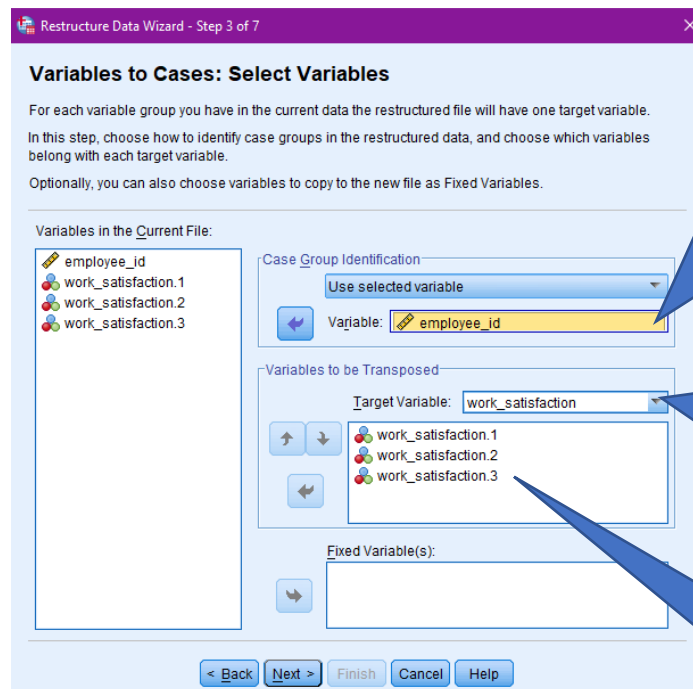
2. Select "Restructure selected variables into cases"



- When asked how many variable groups you want to restructure, you'll have to identify *how many variables* you want to appear in *each row* of your long dataset. In this case, we're only restructuring one variable ("work\_satisfaction"). However, if we had three variables (e.g., "work\_satisfaction," "extraversion," and "happiness"), we'd have to tell SPSS that we want to restructure *three* variables.



- Tell SPSS how to restructure the dataset:

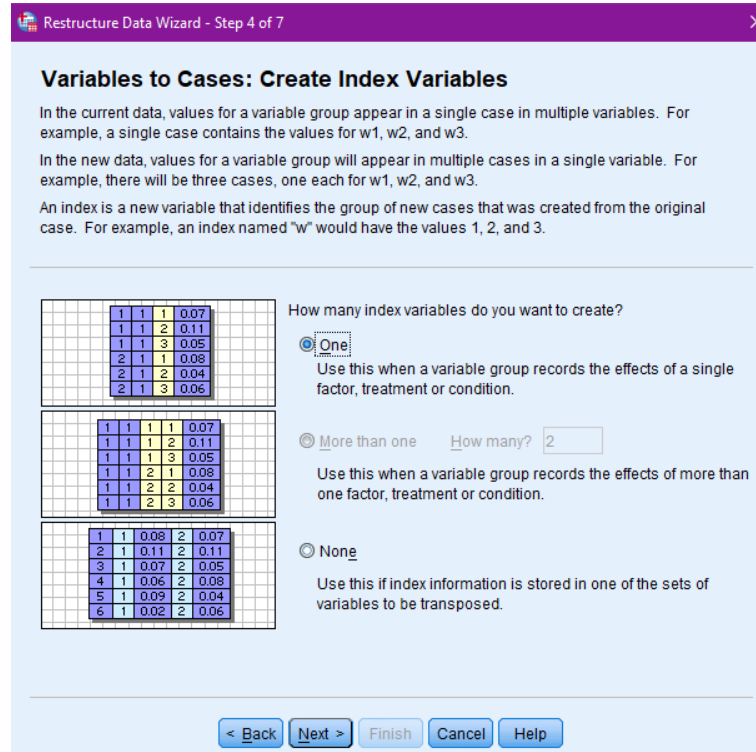


We have to tell SPSS that "employee\_id" is the grouping variable. When the data from a single row is split into multiple rows, this variable links all the new rows together

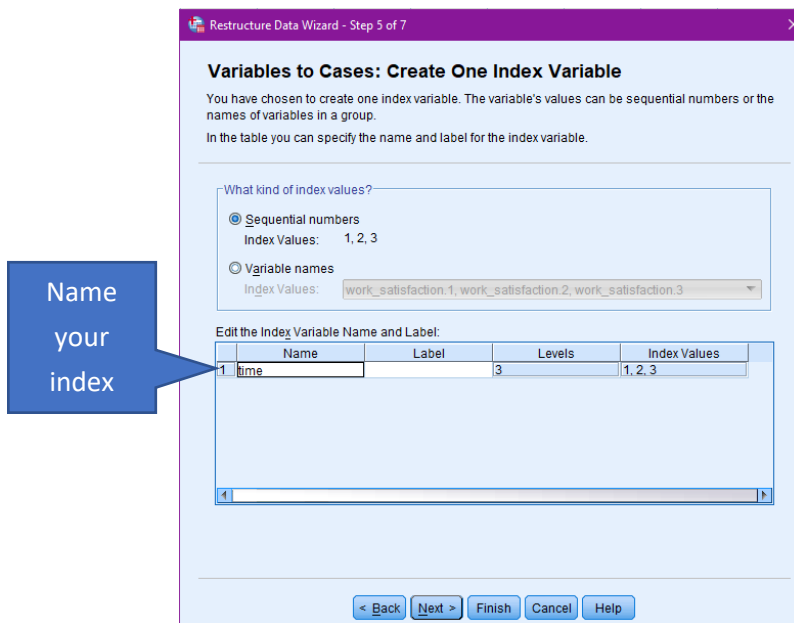
We're going to create a single new variable called "work\_satisfaction"

These three variables should appear IN THIS ORDER.

- Next, we'll need to create an *index variable* (such as measurement occasions). Generally, you'll only have one index variable, unless you're working with a very complex long dataset:



- In this case, we can use sequential numbers as our index variable (make sure your variables are in time-order in step #4 above!), and we will call our index “time.” In other cases, it might be more appropriate to use the original variable names as the index variable.



7. Click "Finish." Your dataset is now in long format! You're ready to correlate "time" with work satisfaction to see how work satisfaction changes across time!

	employee_id	time	work_satisfaction
1	30490	1	5
2	30490	2	5
3	30490	3	5
4	30491	1	5
5	30491	2	4
6	30491	3	4
7	30492	1	5
8	30492	2	4
9	30492	3	4
10	30493	1	5
11	30493	2	5
12	30493	3	5
13	30495	1	4
14	30495	2	.
15	30495	3	.
16	30496	1	5
17	30496	2	5
18	30496	3	4
19	30497	1	4
20	30497	2	3
21	30497	3	4
22	30498	1	2