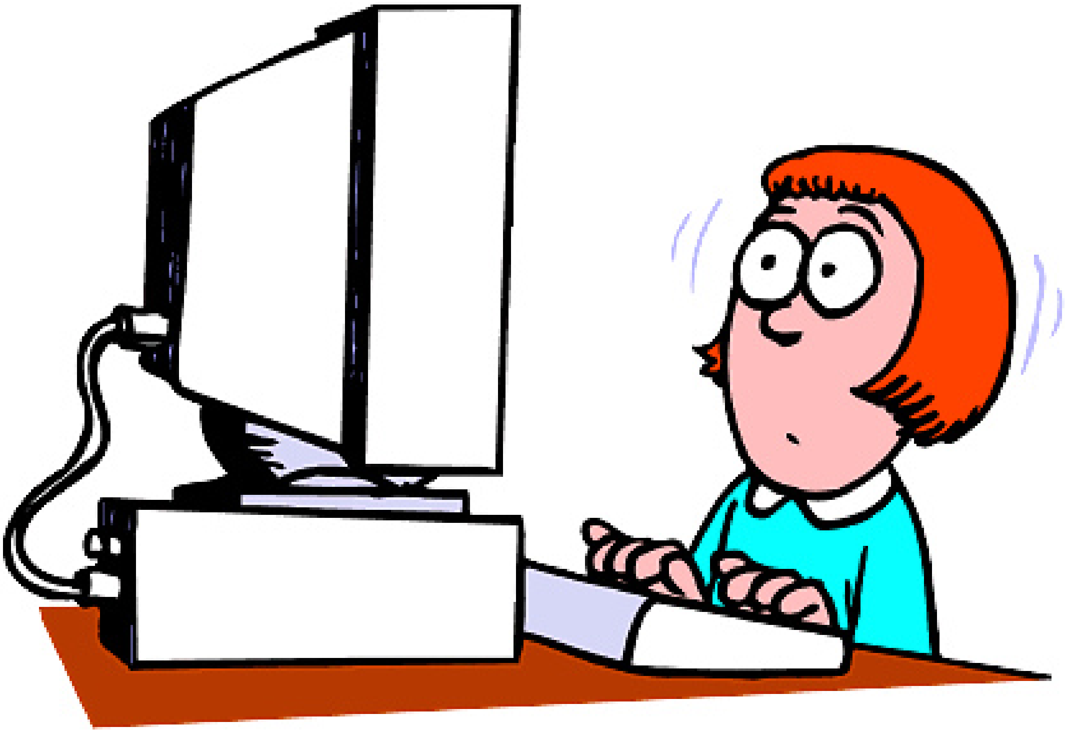
Data entry guidelines using Excel



Dr Dana Hince and Professor Max Bulsara

Institute for Health Research, The University of Notre Dame, Fremantle

Data entry guidelines using Excel

Good data entry is a process that:

- Minimizes data entry error
- Minimizes the risk of misinterpretation
- Minimizes the amount of data manipulation required prior to commencing analysis

Although there are specialised programs available for data entry, Excel is still the most commonly used program we see used for this purpose. Most of the analysis of the data, however, is conducted in specialised statistical programs (e.g. SPSS, Stata or SAS). You will save yourself/your data analyst **A LOT** of time if your Excel data spreadsheet is constructed in such a way that the statistical program can “understand” it.

The more common mistakes we see in spreadsheet construction are a consequence of thinking like a human and not like a computer. For example, 9 the number and 9 the character have the same meaning for people, as do the words male and Male. We can interpret what we see within the context of the presentation – computers aren’t so good at that. So when it comes to setting up a spreadsheet, your life will be easier and you will get your results quicker if you follow the “computer’s rules”.

By the end of this manual you will be able to engage in “good data entry” (as defined at the top of this page) by achieving the following objectives:

1. Structure an Excel spreadsheet that allows easy importation into various statistical packages.
2. Choose appropriate variable names, variable labels, and value codes and labels.
3. Construct a codebook/data dictionary.
4. Format Excel cells.
5. Limit Excel cells to valid ranges or prepare drop down options.
6. Know what ***not*** to include in your spreadsheet.

1. Set up your spreadsheet.

a) Data spreadsheets can be ***wide (Fig1)*** or ***long (Fig2)***.

In both formats

- the columns represent variables
- the first column is a ***unique identifier*** for each participant
- the rows represent cases/events
- the rows must be ***uniquely identifiable***

The difference between the two is how many variables are needed to ***uniquely identify a row.*** Wide format needs only one identifier, whereas long format needs two or more.

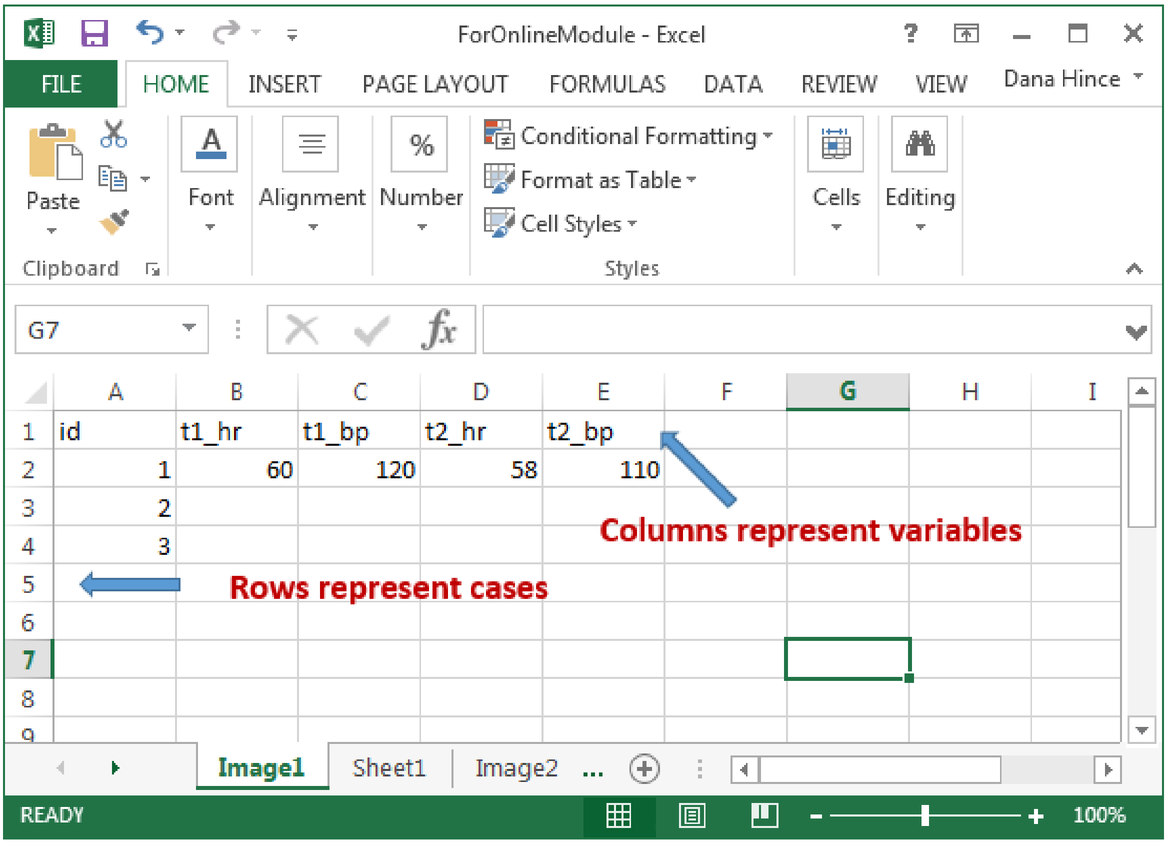
For example, compare Fig1 and Fig2. In wide format (Fig1) knowing the id and variable name (id=1, variable=t1_bp) leaves only one choice of value (120), whereas in long format (Fig2) knowing the id and variable name (id=1, variable=bp) leaves you with two options (120 or 110). In this case, we also need to know the value of time to limit the number of possible rows to one.

Long format is often the format required for the analysis of longitudinal or repeated measures study designs. It is possible to use statistical software to switch between wide and long format (called *restructuring* or *reshaping* the data). If your data requires restructuring, please ***DO NOT CUT AND PASTE IN EXCEL.*** It is too error prone!

CAUTION! Regardless of format, make sure your unique id matches any case report forms or other hard copy or

MAKE A START!

1. What is your spreadsheet format?
2. How is each row uniquely identified?
3. What are your id values?

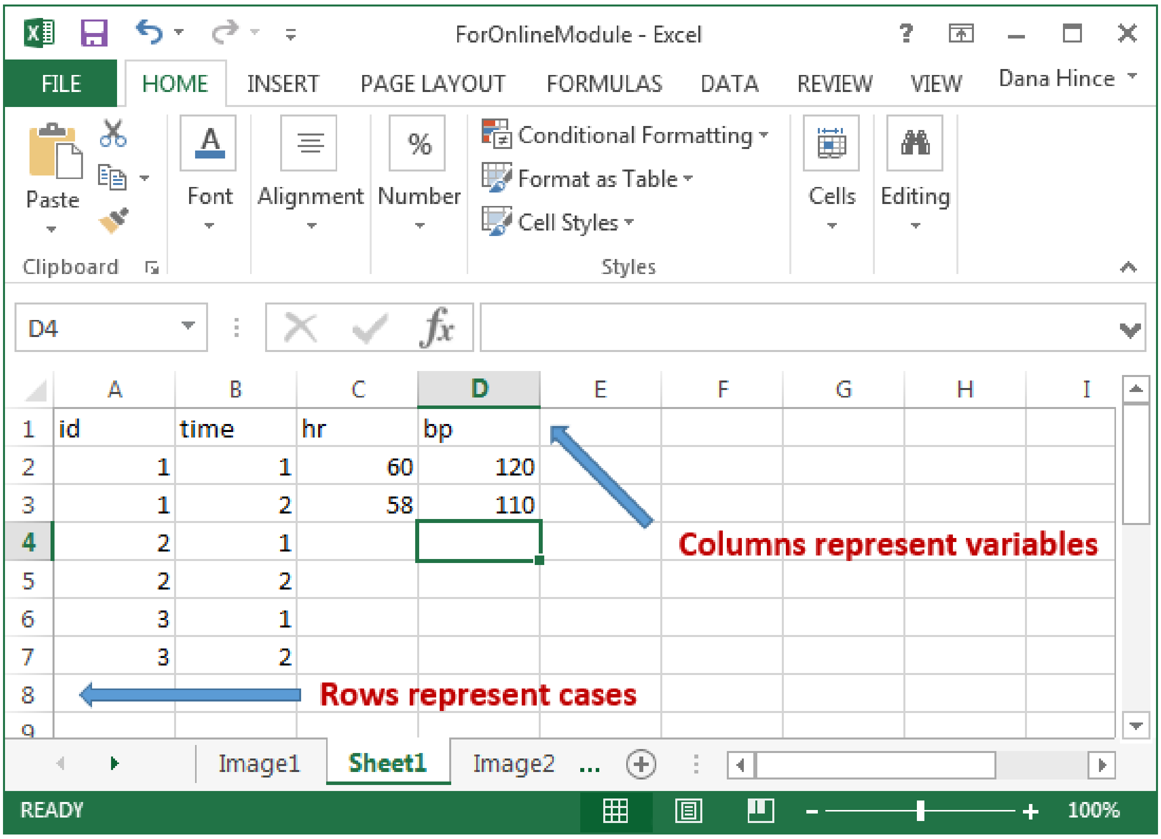


id	t1_hr	t1_bp	t2_hr	t2_bp
1	60	120	58	110
2				
3				

Columns represent variables

Rows represent cases

Figure 1. Basic 'wide' spreadsheet format.



id	time	hr	bp
1	1	60	120
1	2	58	110
2	1		
2	2		
3	1		
3	2		

Columns represent variables

Rows represent cases

Figure 2. Basic 'long' spreadsheet format.

b) The ***first row*** is used for ***variable names***.

Variables names

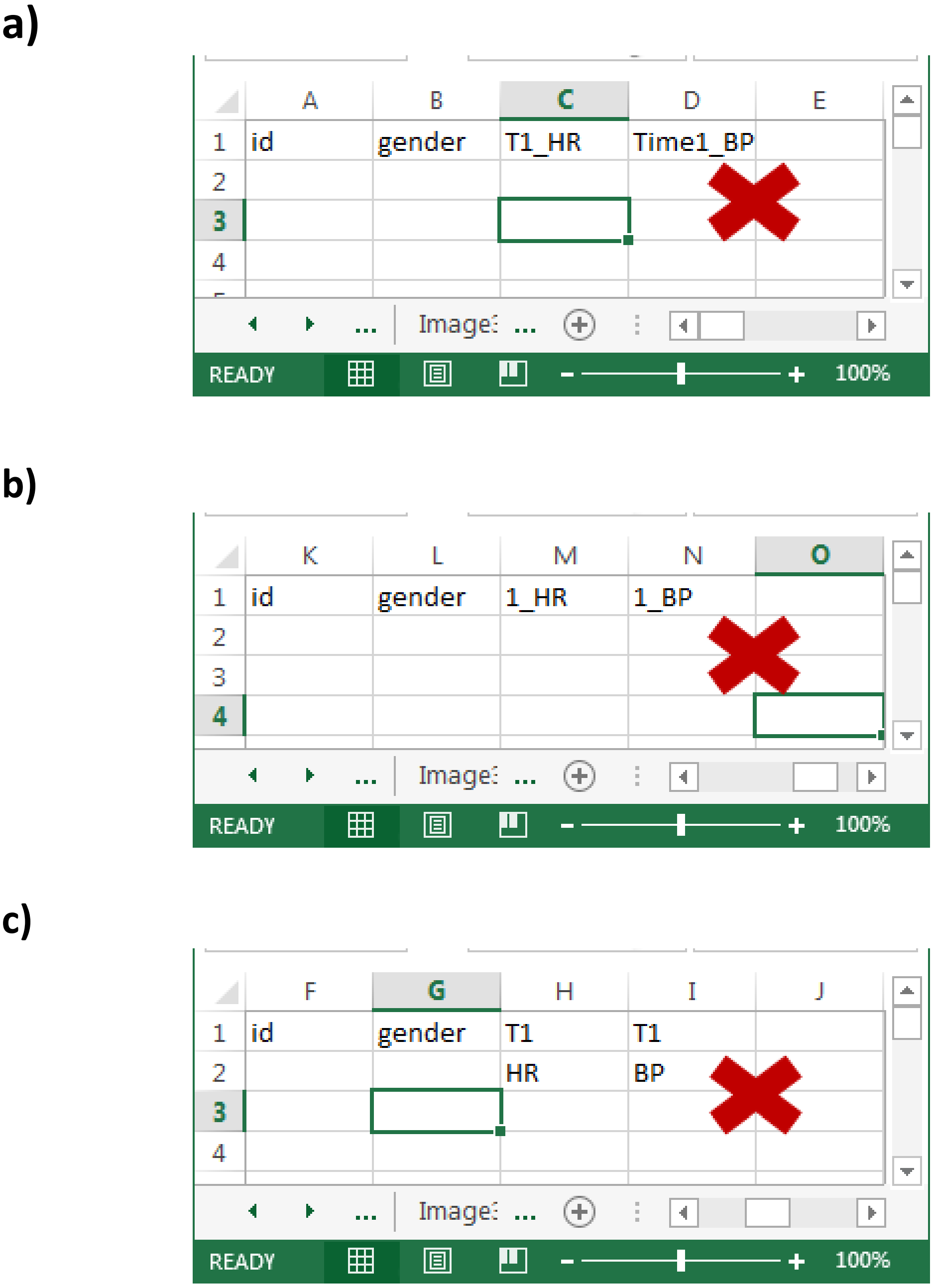
- are meaningful and short
- contain all relevant information
- have no spaces between words
- start with letters only (or _)
- only take one row (the first one!) in the spreadsheet
- are consistently named (e.g. heart_rate1, heart_rate2 *NOT* HR_1, hrate2) **THIS IS PARTICULARLY IMPORTANT IF ANY DATA RESTRUCTURING IS REQUIRED** (see page 3)

*TIP! Using lower case letters **only** for variable names helps reduce typos if you are using command driven software*

MAKE A START!

4. Test your knowledge on the next page then
5. Enter your variable names into a blank Excel

POP QUIZ 1: what is wrong with these variable names?



a)

	A	B	C	D	E
1	id	gender	T1_HR	Time1_BP	
2					
3					
4					

b)

	K	L	M	N	O
1	id	gender	1_HR	1_BP	
2					
3					
4					

c)

	F	G	H	I	J
1	id	gender	T1	T1	
2			HR	BP	
3					
4					

Figure 2a), b) and c). Examples of troublesome variable names.

Answers.

a) Time1 and T1 refer to the same time. Needs to be consistently labelled.

b) Beginning variable names with numbers upsets many statistical programmes.

c) Variable names need to be contained in one cell only.

c) Decide on ***codes*** for

- ***categorical variables*** i.e. use numbers to denote the categories

  So gender could be 0 = female and 1 = male and 99 = missing data
- Although you don't need to code ***continuous variables*** given the measurement is already numeric (e.g. height, weight, blood pressure) you **will** need a missing value code that is
  - i. **NOT** within the range of expected or probable values for that variable (e.g. 99 would not be a good missing code for diastolic blood pressure because it **could** be a true value).
  - ii. **NOT** text! If you use text (e.g. N/A) in a numeric variable, some statistical packages assume that the whole variable is text (string), and won't be able to perform the calculations you need it to do.

d) For ***free text variables*** (e.g. answers to open ended questions or important notes from data collection).

- The number of variables required for each free text question depends on the number of answers given.

  For example:
  - i. If the question is "List 3 breeds of dog you would consider as a pet" then you need to have 3 variables (e.g. breed1, breed2, breed3).
  - ii. If the question is "List all the breeds of dogs you would consider as a pet", you need enough variables so that the maximum number of breeds listed as an answer can be accommodated with one breed per variable.
- Enter free text answers ***exactly*** as they appear on the survey.

e) Be careful with ***date variables.***

To people, 14 May 2009 is the same as 12 May 09 is the same is 14/5/09 is the same as 14/05/09. To your statistical package, they are potentially all different. Always enter your dates into excel in the same format, and **set the date value cells to date format by**

- Right click in the cell (or selected group of cells)
- Choose Format Cells in the drop down option (Fig. 3a)

- In the pop up box that follows, choose Date, and the format you would like and click OK (Fig. 3b: I like the 3rd option down because it includes the 4 digit year).

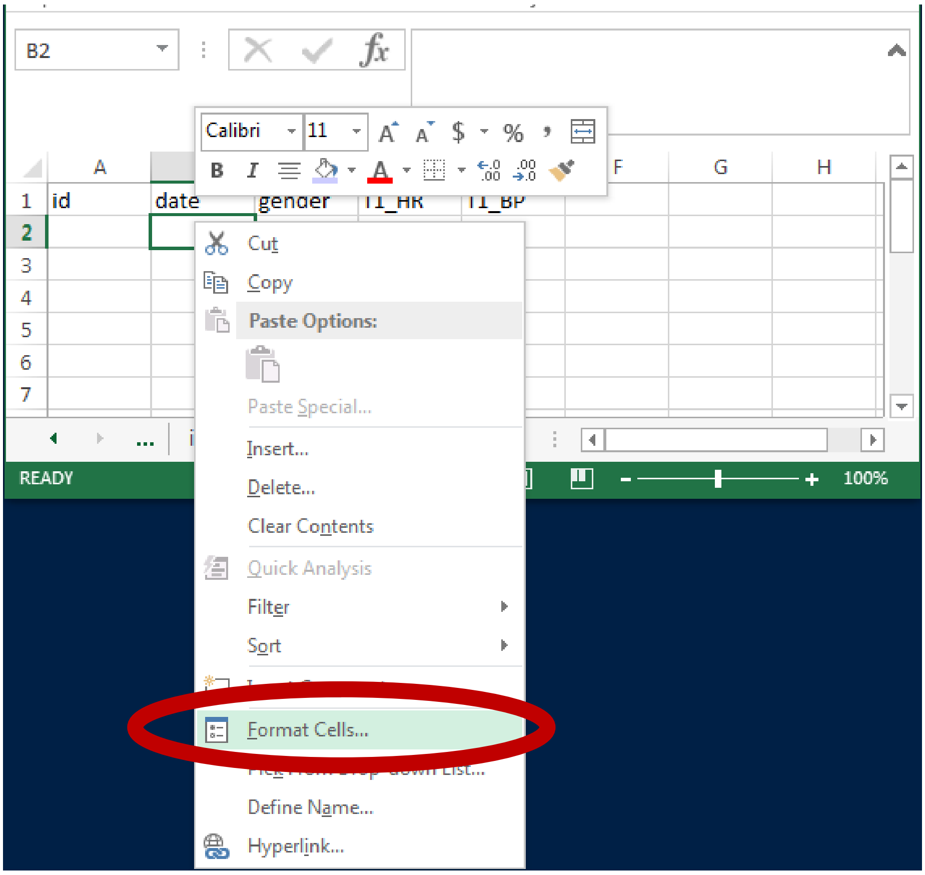


Figure 3a. Formatting cells

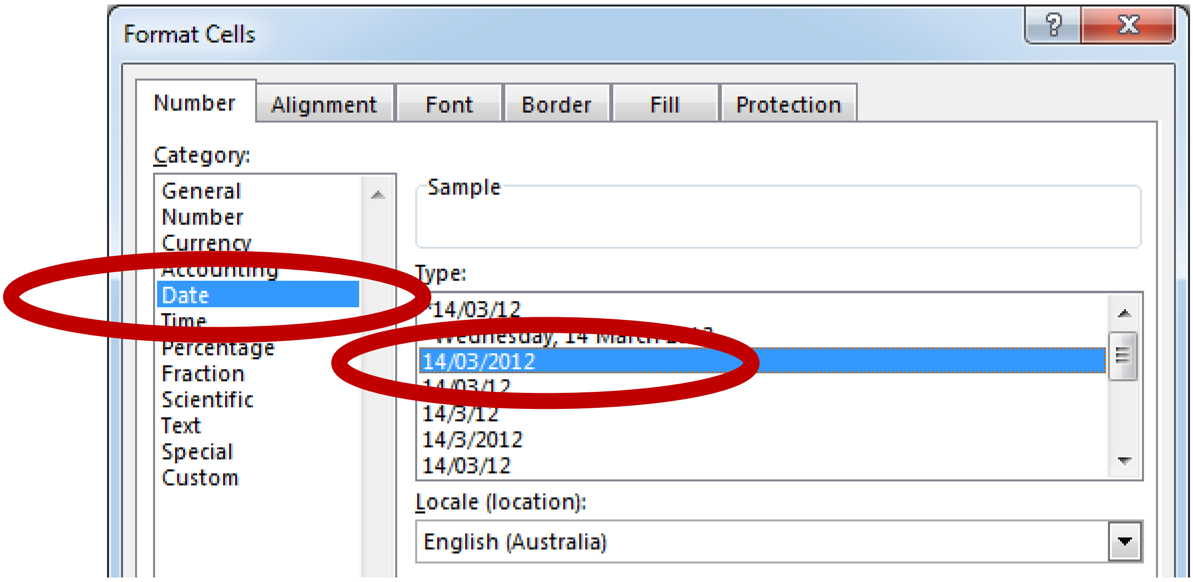


Figure 3b. Formatting cells containing date values

Make sure dates are entered in a consistent format, with 4 digit years, because inconsistent/incomplete dates can equal hours of data manipulation

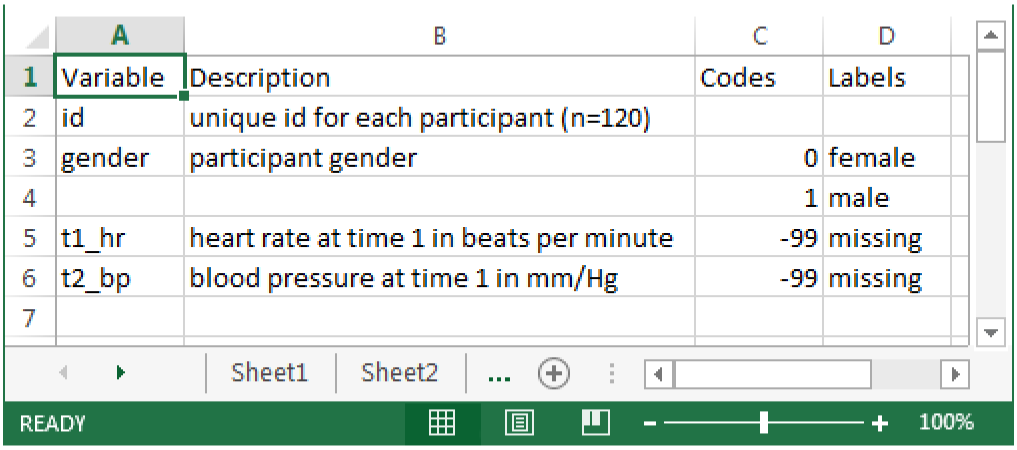
MAKE A START!

6. Format all the cells that will contain date values now

2. Reduce the risk of data entry error or misinterpretation

a) Set up a codebook or data dictionary

The codebook is how you communicate what each variable represents, and what to expect to find in it, to other members of your research team.



Variable	Description	Codes	Labels
id	unique id for each participant (n=120)		
gender	participant gender	0	female
		1	male
t1_hr	heart rate at time 1 in beats per minute	-99	missing
t2_bp	blood pressure at time 1 in mm/Hg	-99	missing

Figure 4. An example of a codebook/data dictionary.

- List your variable names with labels/descriptions in sufficient detail so that someone else can understand what they are.

- Include the codes and value labels for any categorical variables.

- Include ranges and/or units of measurement for continuous variables (e.g mmHg, ng/L) as this may not be obvious to others.

- Include codes for missing values.

MAKE A START!

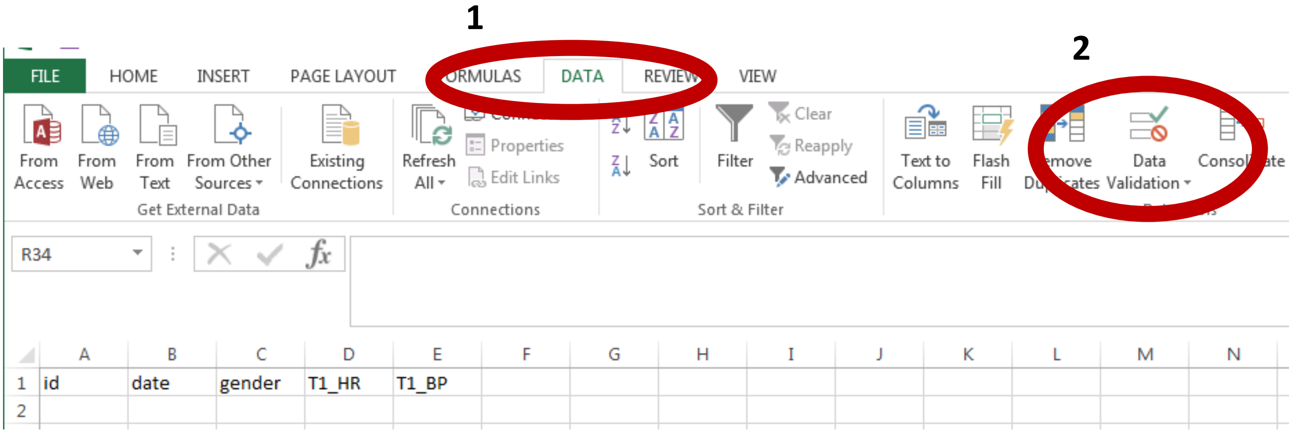
7. Set up your data dictionary on the next sheet in your Excel file, following Fig4 as an example

b) Excel Data Validation function

Now you have your data dictionary sorted, you can use this information to reduce the risk of data entry error using ***Excel's data validation*** tool.

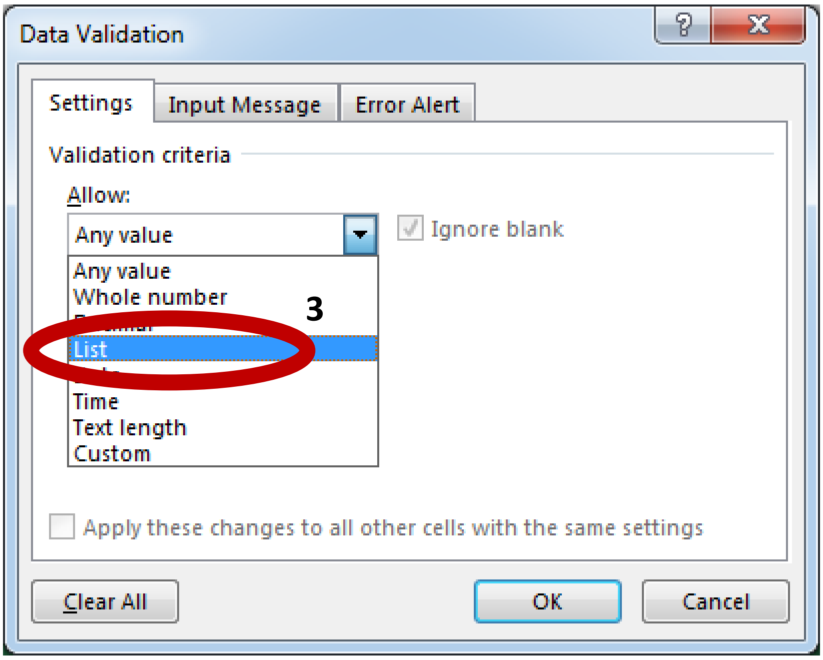
This function allows you to specify allowable ranges/values for variables, provide data entry prompts and/or data entry error messages. There are many options available – let's take two very useful as an example.

Firstly, let's set up a drop down list with an error message for the variable gender. Go to the Data ribbon on the top of the Excel window **(1)**, and click Data Validation **(2)**.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	date	gender	T1_HR	T1_BP									
2														

The window below will pop up. Choose List from the Allow: drop down menu **(3)**.



Data Validation

Settings | Input Message | Error Alert

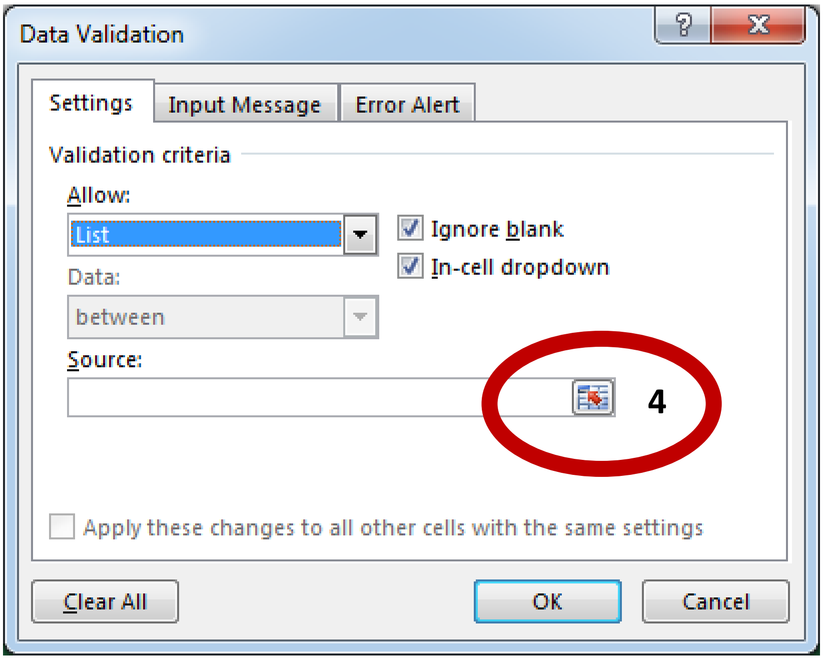
Validation criteria

Allow: Any value

☑ Ignore blank

□ Apply these changes to all other cells with the same settings

Click on the button next to the Source: field **(4)** this will open another pop up box.



Data Validation

Settings | Input Message | Error Alert

Validation criteria

Allow: List

☑ Ignore blank

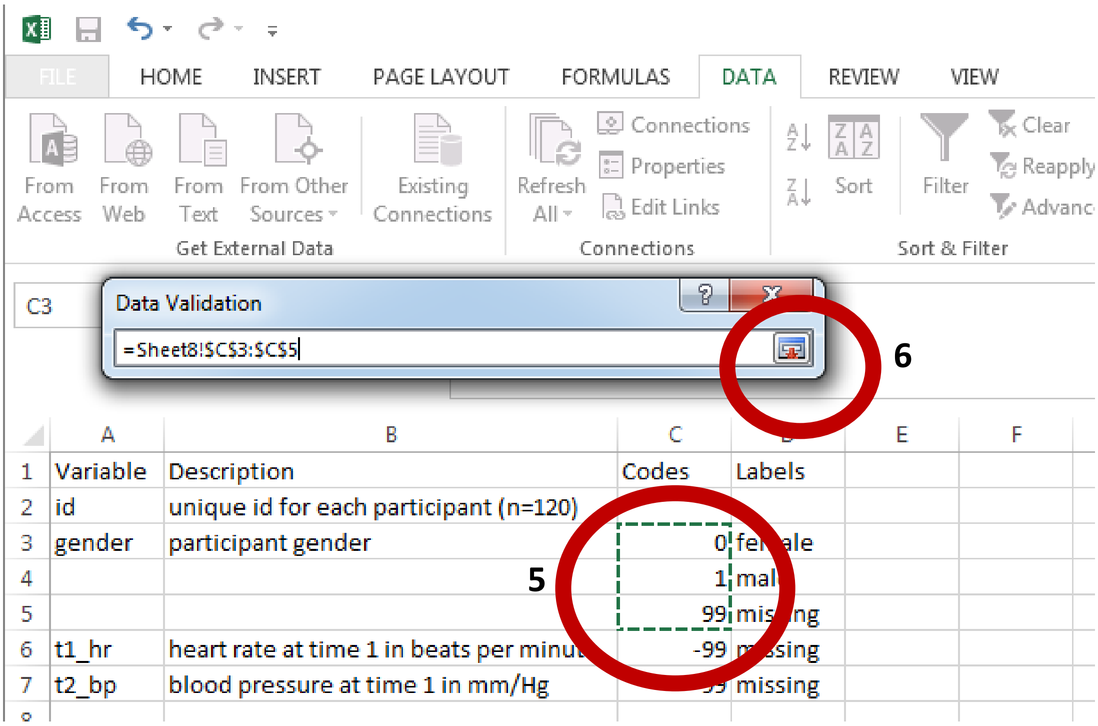
☑ In-cell dropdown

Data: between

Source:

□ Apply these changes to all other cells with the same settings

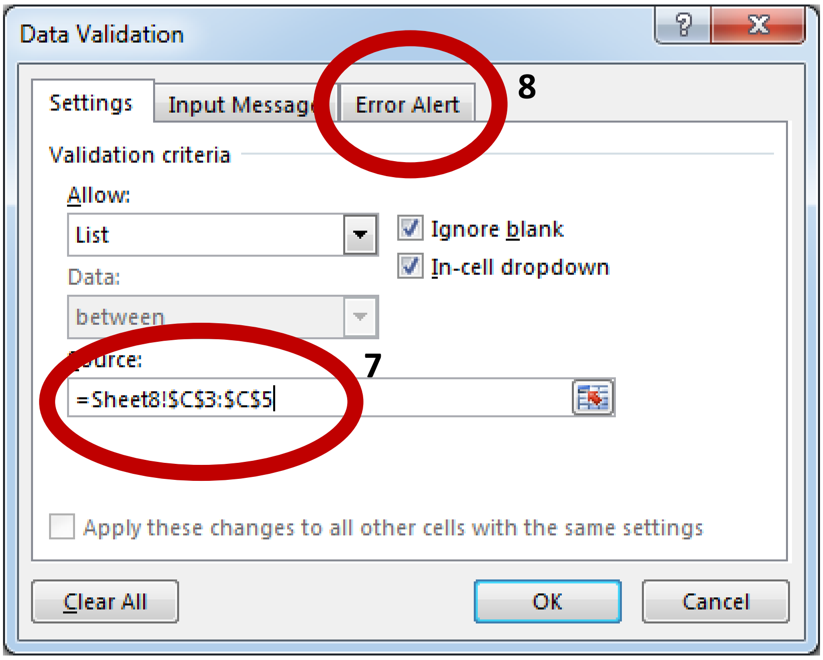
Now choose the cells that have the values you want to allow as valid input from your data dictionary. In this case, they are the codes for gender (**5**: see also Fig 4). Now click the button marked **6**.



=Sheet8!$C$3:$C$5

	A	B	C	D	E	F
1	Variable	Description	Codes	Labels		
2	id	unique id for each participant (n=120)				
3	gender	participant gender	0	female		
4			1	male		
5			99	missing		
6	t1_hr	heart rate at time 1 in beats per minu	-99	missing		
7	t2_bp	blood pressure at time 1 in mm/Hg	-99	missing		

You should be back at the Data Validation pop up, with the cell reference now in the Source: field **(7)**. Now click the Error Alert tab **(8)**.



Validation criteria

Allow: List

☑ Ignore blank

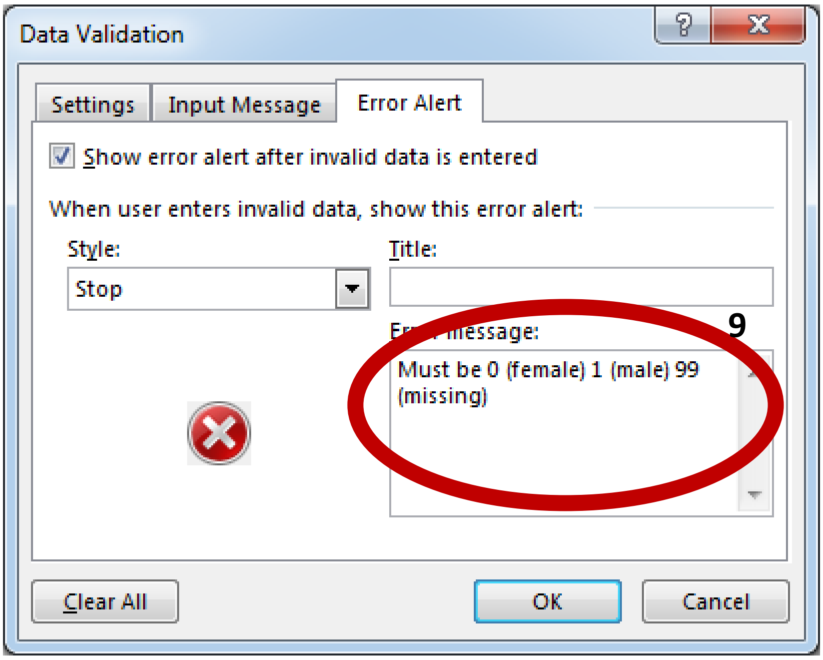
☑ In-cell dropdown

Data: between

Source: =Sheet8!$C$3:$C$5

☐ Apply these changes to all other cells with the same settings

Enter in the error message that you would like to appear if someone makes a mistake in the error message box **(9)**, and click OK.



Data Validation

Settings | Input Message | Error Alert

☑ Show error alert after invalid data is entered

When user enters invalid data, show this error alert:

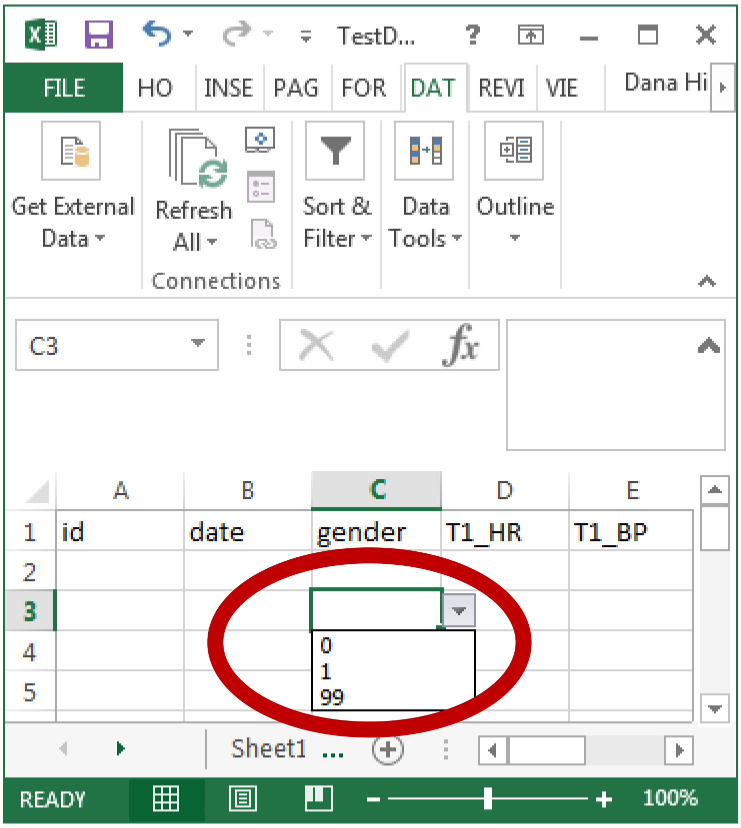
Style: Stop

Title:

Error message:

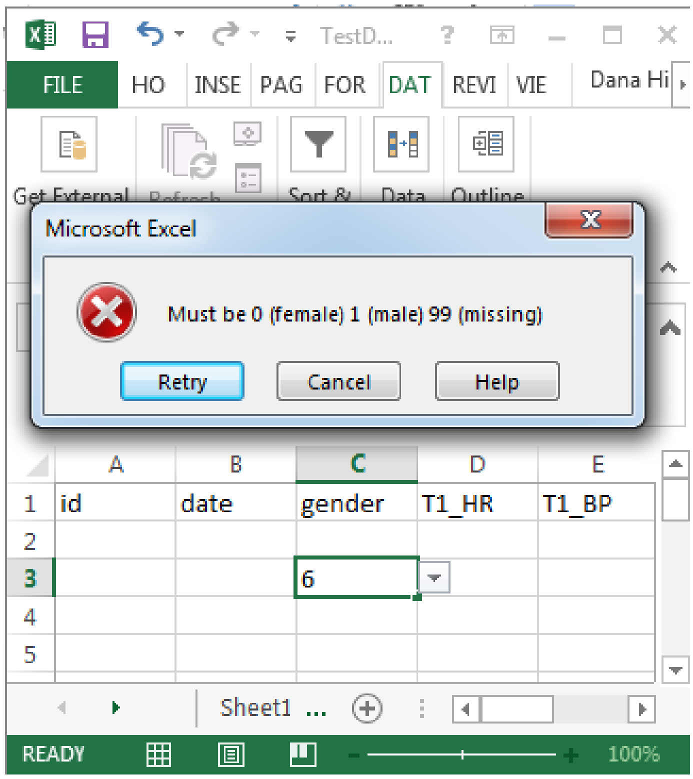
Must be 0 (female) 1 (male) 99 (missing)

When you place the cursor in the cell/s you just set up, a little arrow will appear next to the cell. Click this to access the allowable values.



	A	B	C	D	E
1	id	date	gender	T1_HR	T1_BP
2					
3					
4					
5					

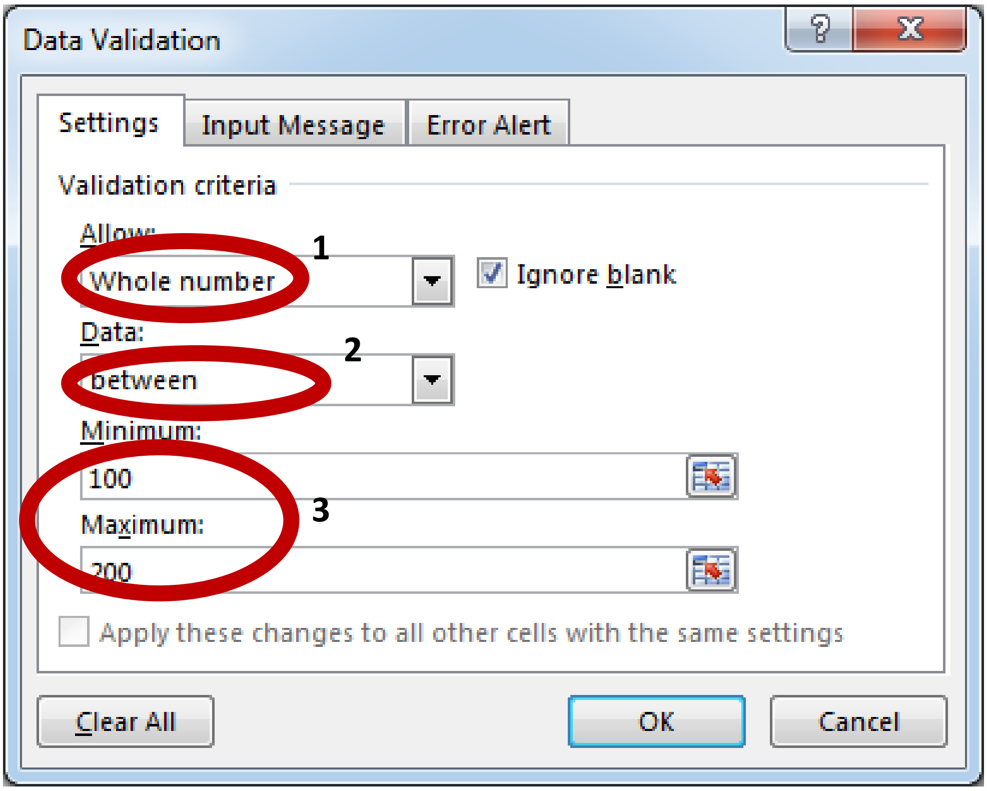
Now try putting in an invalid number. Your error message will appear when you move to the next cell.



Microsoft Excel: Must be 0 (female) 1 (male) 99 (missing)

	A	B	C	D	E
1	id	date	gender	T1_HR	T1_BP
2					
3			6		
4					
5					

There are other options in the Allow: drop down list in the Data Validation window. Let's suppose that our id values can only be between 100 and 200. Choose Whole Number **(1)**, and in the Data: drop down menu that appears **(2)** choose between and then enter the maximum and minimum of the allowable range for that variable **(3)**.



Validation criteria

Allow: Whole number

☑ Ignore blank

Data: between

Minimum: 100

Maximum: 200

□ Apply these changes to all other cells with the same settings

Again, you could set up an error alert that will appear if someone tries to enter an invalid id value.

A bit of extra time at the beginning setting up these Data Validation rules will result in reduced risk of data entry error, and therefore a lot less double checking later on when you come to "clean" the data.

For further "how to" see:

https://support.office.com/en-us/article/Apply-data-validation-to-cells-29fecbcc-d1b9-42c1-9d76-eff3ce5f7249

https://support.office.com/en-us/article/Create-a-drop-down-list-7693307a-59ef-400a-b769-c5402dce407b?ui=en-US&rs=en-

MAKE A START!

8. Consider your data and how you might use Excel data validation to reduce data entry error
9. Set up your variables accordingly!

3. Enter away!

- Take lots of breaks if entering "in bulk". Fatigue is your enemy when it comes to accurate entry.

- Try and enter data in real time if at all possible as this might give you the chance to collect missing data values before it is too late….

- If you have multiple people entering data, make sure that they have access to the data dictionary and please use the data validation tool!

4. What NOT to include in a spreadsheet.

- Column or row means, standard deviations or any other summary statistic calculated in Excel (if you have used them to do preliminary data checks, just remove them before importing into statistical package/sending to the statistician).

- Empty/blank rows or columns.

- Text and numeric values in the same variable. If you find yourself doing this you need two separate variables.

- Colour coding – this is fine if it helps you with ensuring accurate data entry, but all the relevant information ***needs*** to be included in the variables.

- Identifying information (e.g. names address etc). Keep a separate, password protected spreadsheet which links the ID number with the participant identifying information.

POP QUIZ 2: What is wrong with this spreadsheet? Can you find the errors? How would you fix them?

Hypothetical data comparing meditation vs no meditation treatment effects on blood pressure, body temperature and 3 anxiety questionnaires, pre (time 1) and post (time 2) intervention.

id	test 1											
	sbp	temp	anx_q1	anx_q2	anx_q3	2_SBP	t2_temp	t2_anx_1	t2_anx_2	t2_anx_3	gender	
1	120	37	3	3	4	121	37	3	6	5	M	
2	124	37	5	2	6	124	36.9	5	6	6	m	
3	125	37.1	6	0	5	123	37	6	6	9	f	
4	126	36.9	9		2	135	36.9	8	8	9	fem	
5	135	36.8	10	9	9	110	36.8	5	9	3	male	
6	101	37	2	7	8	112	37	8	6	6	male	
7	26	36.6	4	6	3	120	37	5	6	8	female	
8	180	37.2	6	6	8	121		9	3	7	f	
9	125	37.5	8	3	5	117	37.1	6	6	8	f	
10	115	37	7	5	6	160	36.1	5	9	4	F	
11	116	37.4	6	8	8	138	37.4	3	9	5	f	
12	138	36.8	3	4	9	124	36.9	1	5	6	Male	
13	117	36.9	9	9	2	120	36.8	2	8	9	Male	
14	106	37	4	3	9	115	37	5	4	1	F	
mean	118.1429	61.12143				124.2857	61.90769					
Meditation												
No meditation												

Answers.

1. Two rows used for the variable names; these should be in one row.
2. Variable name starts with a number; instead put the number at the end.
3. Variable names are inconsistent; choose one format and stick to it
4. Gender is not numerically coded; assign numbers to male, female and missing and add codes into the data dictionary. NB it is possible to 'encode' text values in statistical packages, but it requires extra steps and it isn't easy if the text coding is not consistent as is the case here!
5. Row means are included; delete them.
6. Colour coding used to define treatment groups; create another variable that assigns numbers to the groups and add into the data dictionary.
7. Missing values are not coded (in anx_q2 and T2_temp); decide on an appropriate code, enter into spreadsheet and into data dictionary.
8. Where is the dictionary?

For further information please contact:

Dana Hince PhD
Research Biostatistician
dana.hince1@nd.edu.au

or

Professor Max Bulsara
Chair of Biostatistics
max.bulsara@nd.edu.au

Institute for Health Research,
University of Notre Dame, Fremantle
19 Mouat Street,
P.O Box 1225, Fremantle, WA 6959