

Prepare your result file for input into SPSS

Isabelle Darcy

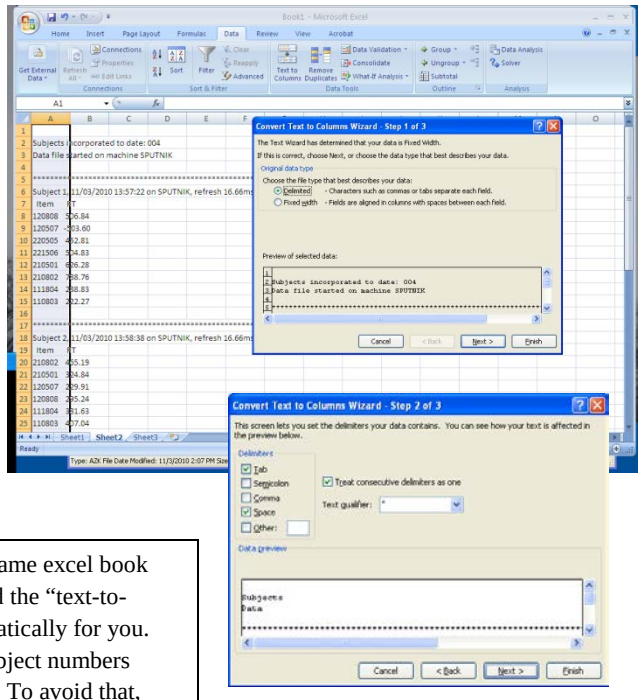
When you use DMDX for your experiment, you get an .azk file, which is a simple text file that collects all the reaction times and accuracy of your subjects. Normally, you have one big .azk file for all your subjects. If you tested on several machines (2 or 3 different computers), you will need to merge the different azk files into one big one. There's an "app" for it: The UNLOAD-AZK util that comes with DMDX. And there's a tutorial for this too: "Step by step use of UnloadAZK and Analyze tools in DMDX".

If you tested different groups with different scripts (beginners, native speakers etc.), it is not necessary to unload these into one AZK, but we will merge these in the Excel file below, for input into SPSS.

Instructions and utensils:

- you need the final results file(s), .azk files
- you need Microsoft Excel
- (you might need Crimson Editor (highly recommended anyway!) Work on your own computer where you have this installed)
- you need to master the key strokes of "Ctrl + C" and "Ctrl + V", which are shortcuts for "copy" and "paste". These are crucial.
- you need about 3-4 hours, depending on how large your .azk files are.

Step 1: Get the information into Excel

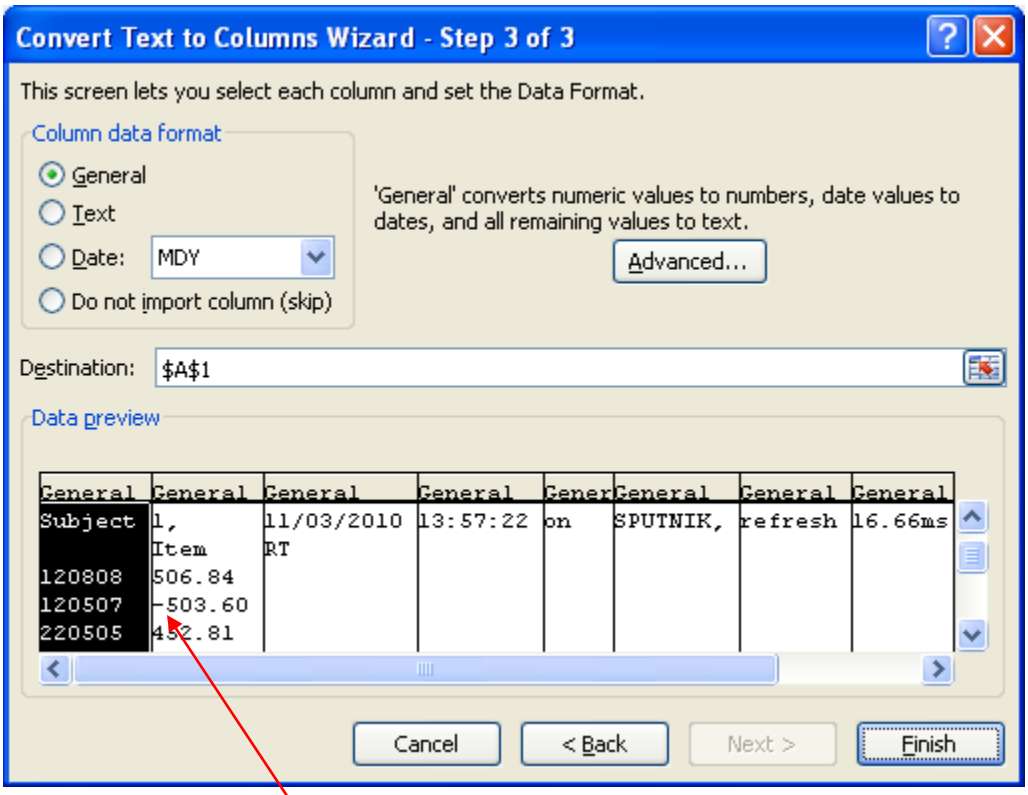


- Take the AZK file (open it with Crimson, for example) and copy and paste it into an empty Excel sheet.

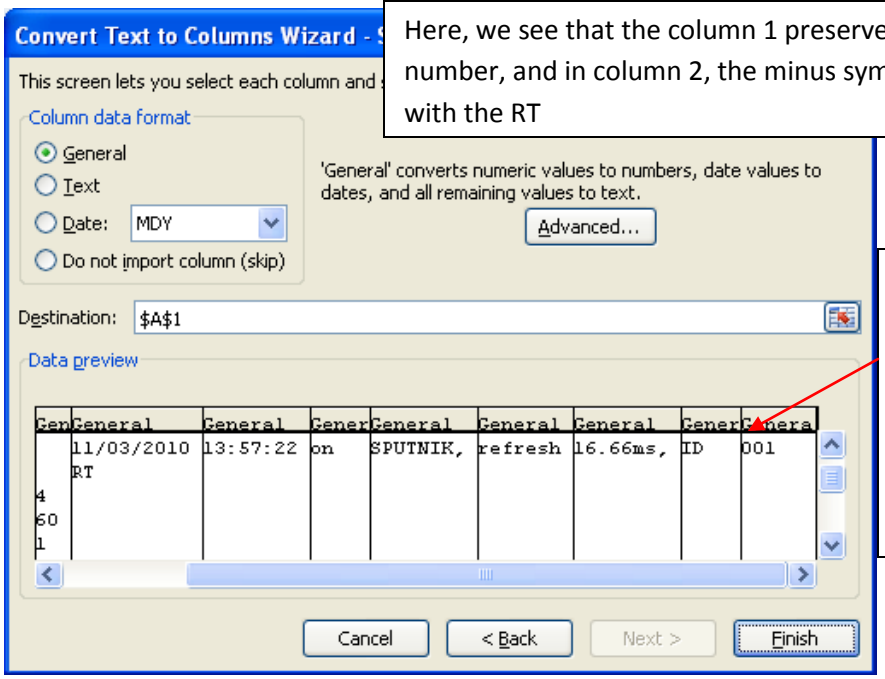
- Select the one column where everything was pasted and click "Data/Text to Columns"; Choose "DELIMITED"
  - Choose "tab" and "space" for delimiters, and check the box "count several consecutive delimiters as one".

Update [Nov. 2012]: If you are doing this several times in the same excel book (on separate sheets, for example) or if you have previously used the "text-to-columns" function, Excel might remember this and do it automatically for you. **This is not a good thing**, because it might also change your subject numbers (e.g. 04-2) into dates (Apr. 2). Which you don't want of course. To avoid that, you can close Excel and open it again. Then, select the entire worksheet (Ctrl+a) and right-click "Cell format". Choose "Text" and click "ok". Be aware of this and verify before you continue…. **This does NOT happen if your subject numbers use underscore (e.g. 04_2)**

- Make sure the "–" symbol (minus symbol) is preserved together with the reaction time when you look at where Excel will split the different columns. This symbol is REALLY important, because it is how you will figure out the accuracy. (a "–" means that it was a wrong answer)

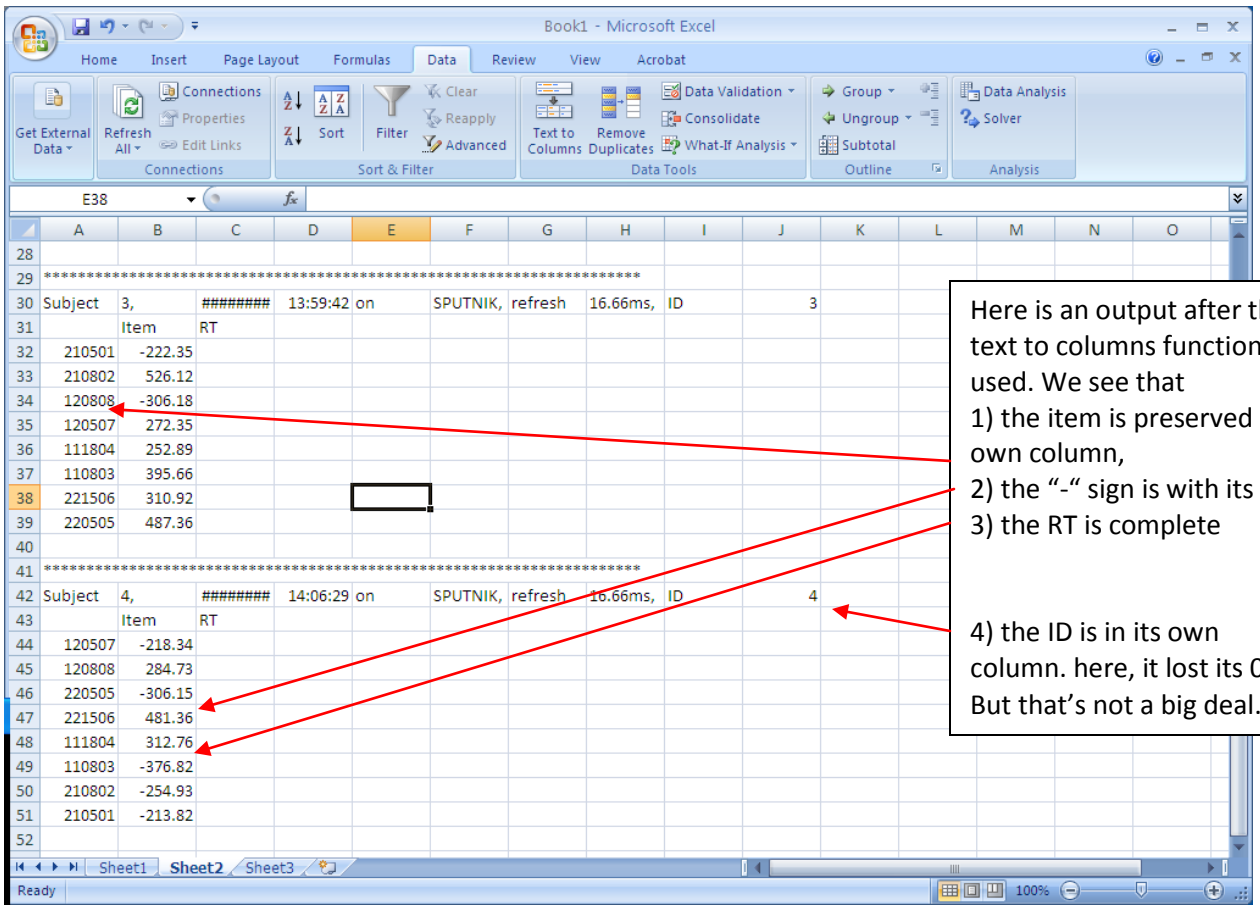


Here, we see that the column 1 preserves the item number, and in column 2, the minus symbol is together with the RT

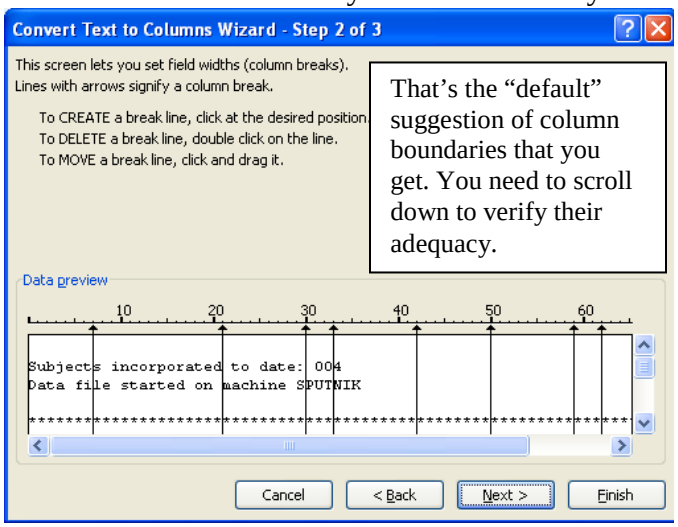


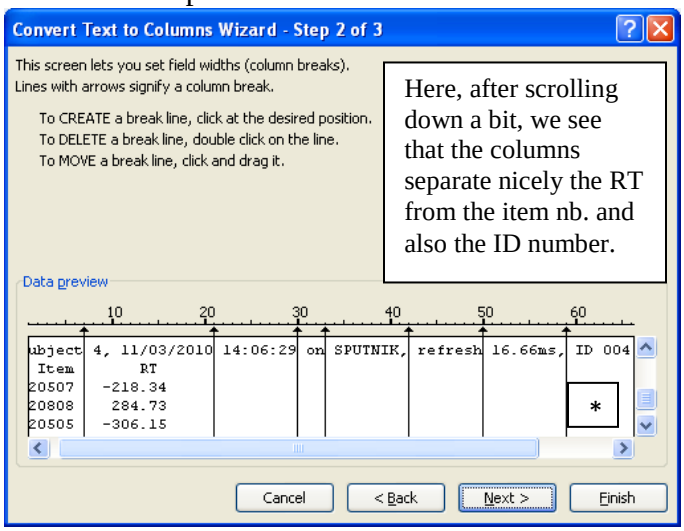
Here, we see that the ID code is clean in one column and not split over 2 columns. This is very important. Make sure (through scrolling down) that your ID numbers are correctly preserved.

Here, click finish and verify again that the column split is preserving 1) entire item numbers, 2) minus symbol together with RT and complete RT numbers, and 3) the ID number is not split.



Potential Problem: The "text-to-column" function using the characters such as space or tabs as delimiters sometimes yields a messy output. It is also a possibility to select "Fixed Width" instead of "Delimited", and to click in the window (see below) to put the boundaries yourself. This is a bit more tricky and you need to scroll down in that mini window to verify that you don't separate the "-" and the RT, and that you delimit a column that does not split the ID number at the end of the ID-information lines. But is is possible to do. Be especially careful if you have training item numbers that are shorter than the test item numbers, for instance. This can be difficult to set one boundary that won't cut any Item number AND preserve the "-" with the RT.





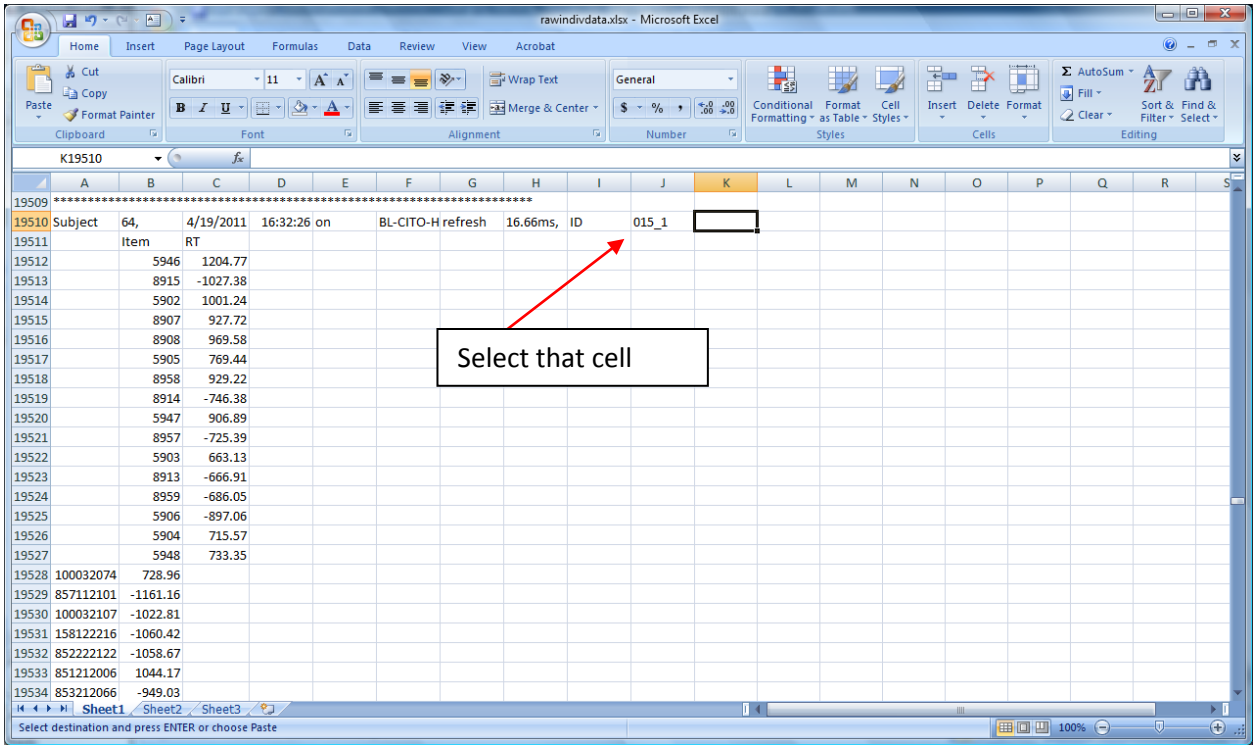
Note: Regarding the preservation of the "00" in the ID number: Excel gets rid of the 00 before a number if the cell is formatted as number (because 001 is the same as 1). If you create a boundary before the "ID", then the cell will be formatted as text and the 00 will be preserved. While it is not very important usually, it might be in some cases. You can see that this is what I did in the second picture above on the right (see the *).

Once this first step is done correctly, you can proceed to the following:

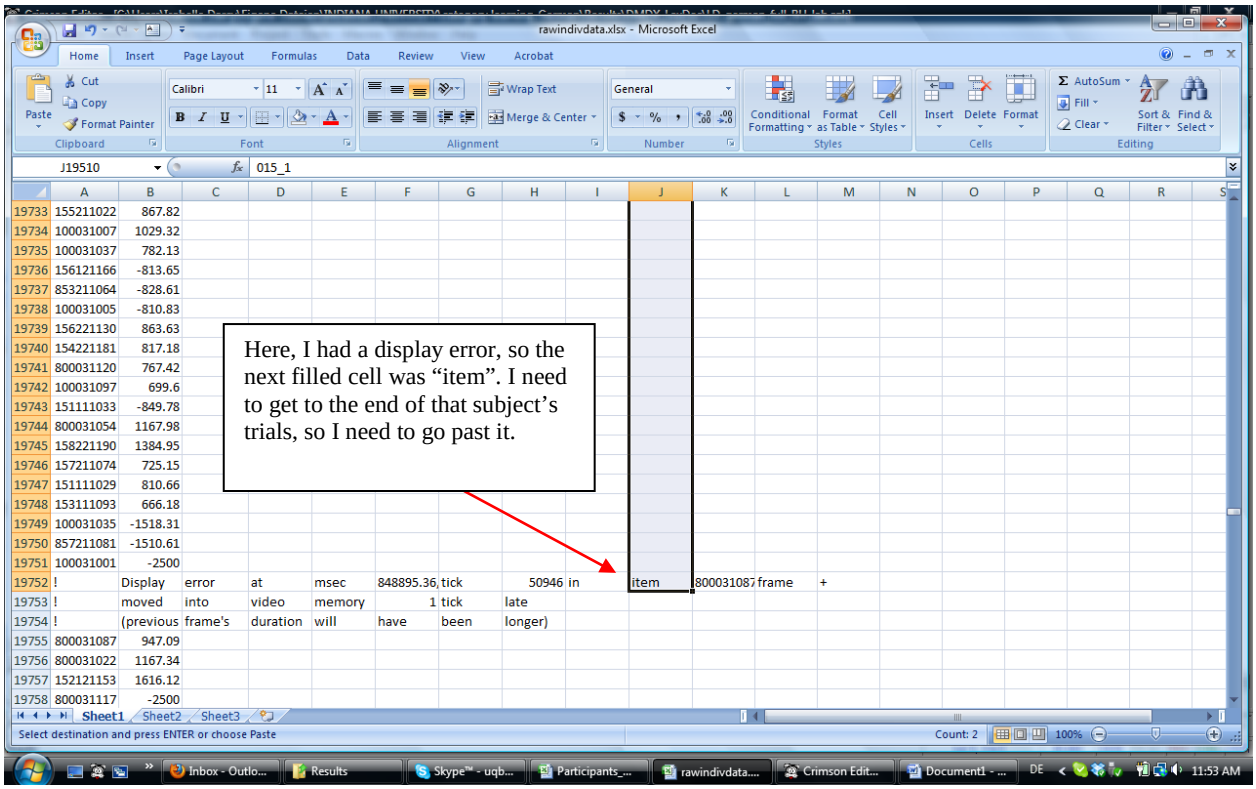
- First column: Select the whole column, and right-click on it. Select "Format Cells", and choose "number" with 0 places after the decimal. This will preserve the item number, and we will need it later.
- The lines where there is text (like the subjects ID codes, or any display errors) will look a bit scrambled, but don't worry about it for now.

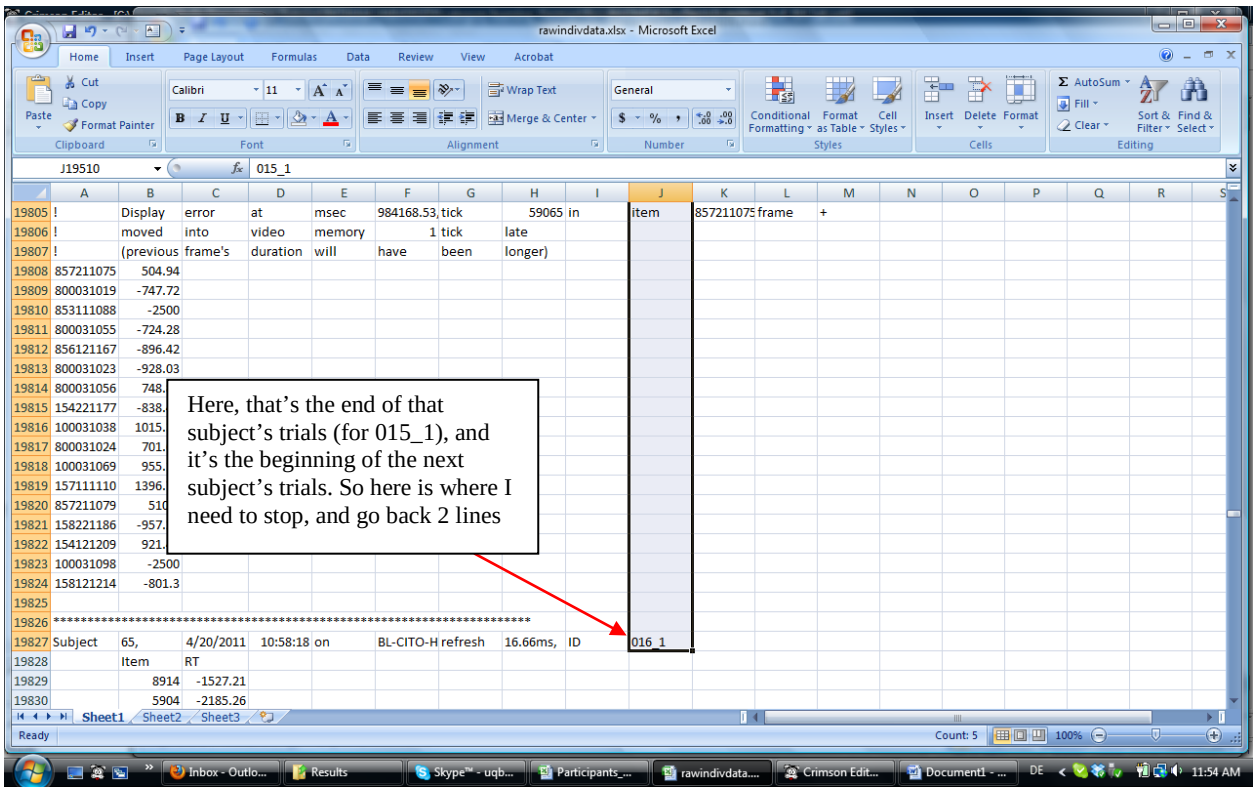
Before reordering anything, we need to make sure each trial is associated with a given learner. DMDX can't report the subjects' ID for each trial: instead, this is saved in the first line of the whole trials for that person.
So we will have to make sure each trial gets the participants' ID code.
A couple of key-strokes can greatly speed up the process. The following screen shots demonstrate how that works:

1. Select the ID code (e.g. 015_1) by highlighting the cell (click on the cell only once, not twice).
2. Now copy the cell with "Ctrl + C"
3. The cell starts "blinking"

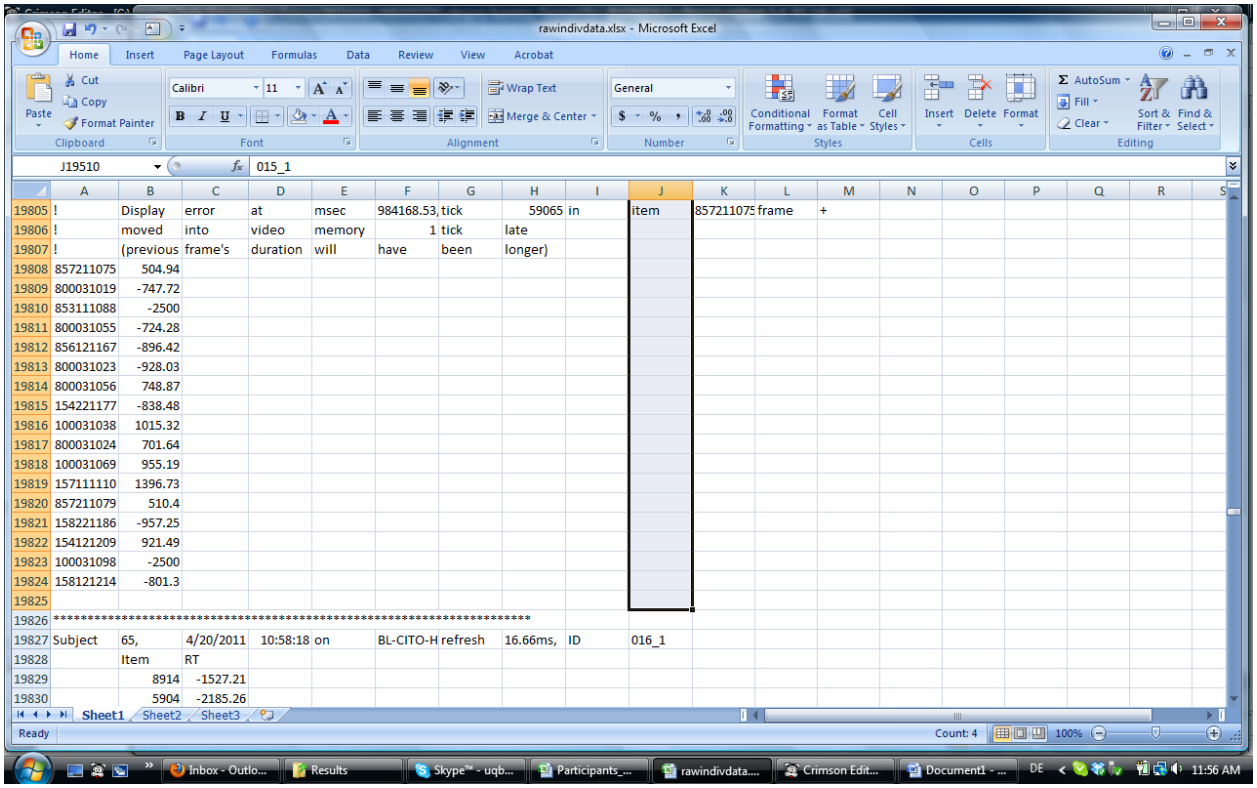


4. Now, select the entire group of cells you want to copy that into. The point is to make sure that this code will be copied in the same column, on the same line as each single trial for that one person. If you have 260 trials in your script, this will mean copy this cell 260 times until you reach the data for the next subject.
    a. First screenshot: Ctrl + Shift + down-arrow: This will extend the selection all the way to the next filled cell. That can be one of the words "item", or anything else that was written in the AZK file, e.g. by a display error.
    b. Press down-arrow again to go past it (but keep Ctrl + Shift pressed the whole time). Until you reach the next ID number. This can mean pressing down-arrow again a few times or just one time, depending on how many display errors there were for any given subject.
    c. Second screenshot: Here, we reach 16_1.

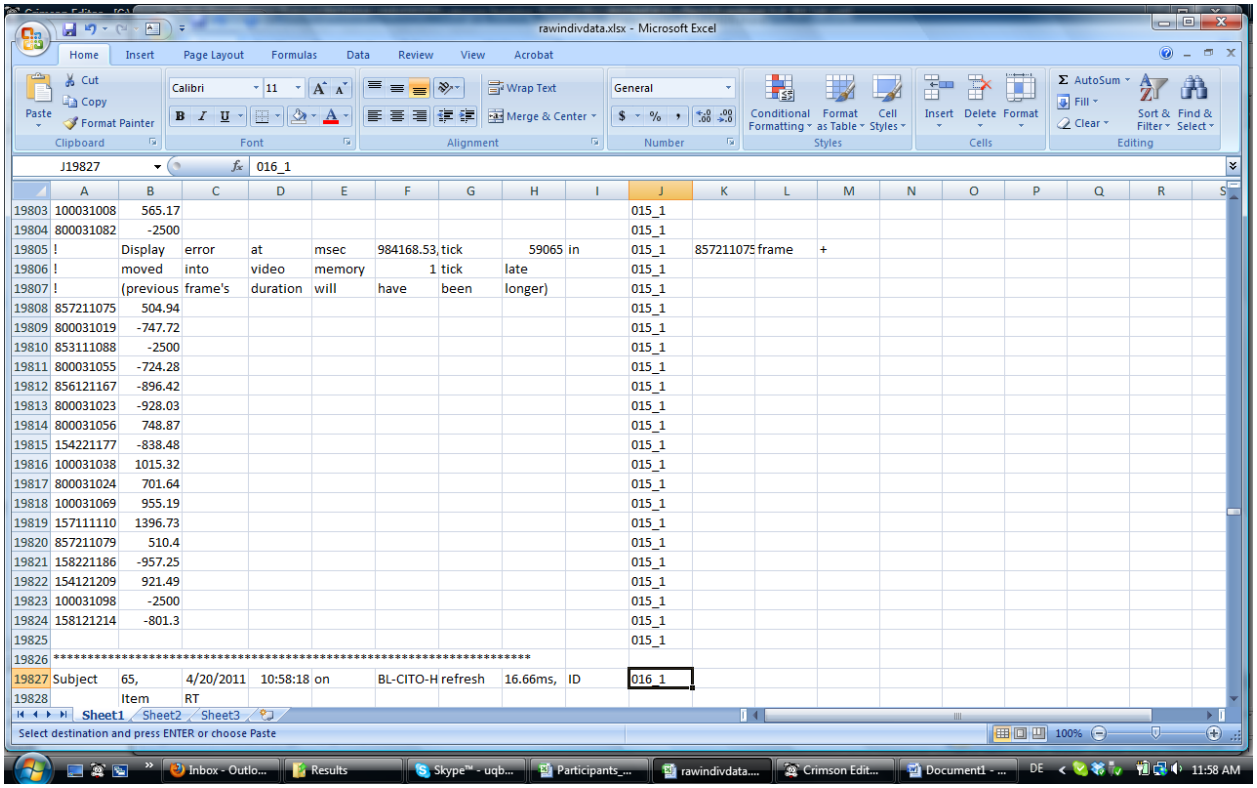




5. Now, to go back a few lines, unpress **Ctrl** while keeping **Shift** down (that's a bit like piano). So only lift up your finger from the **Ctrl** key, keeping the other on **Shift** down, and hit the **arrow-up** key twice.



6. And now, while the selection is still highlighted, lift your hand, and press **Ctrl + V** to copy what you had copied into all these selected cells.
7. All the cells are now filled with the 015_1 ID code.

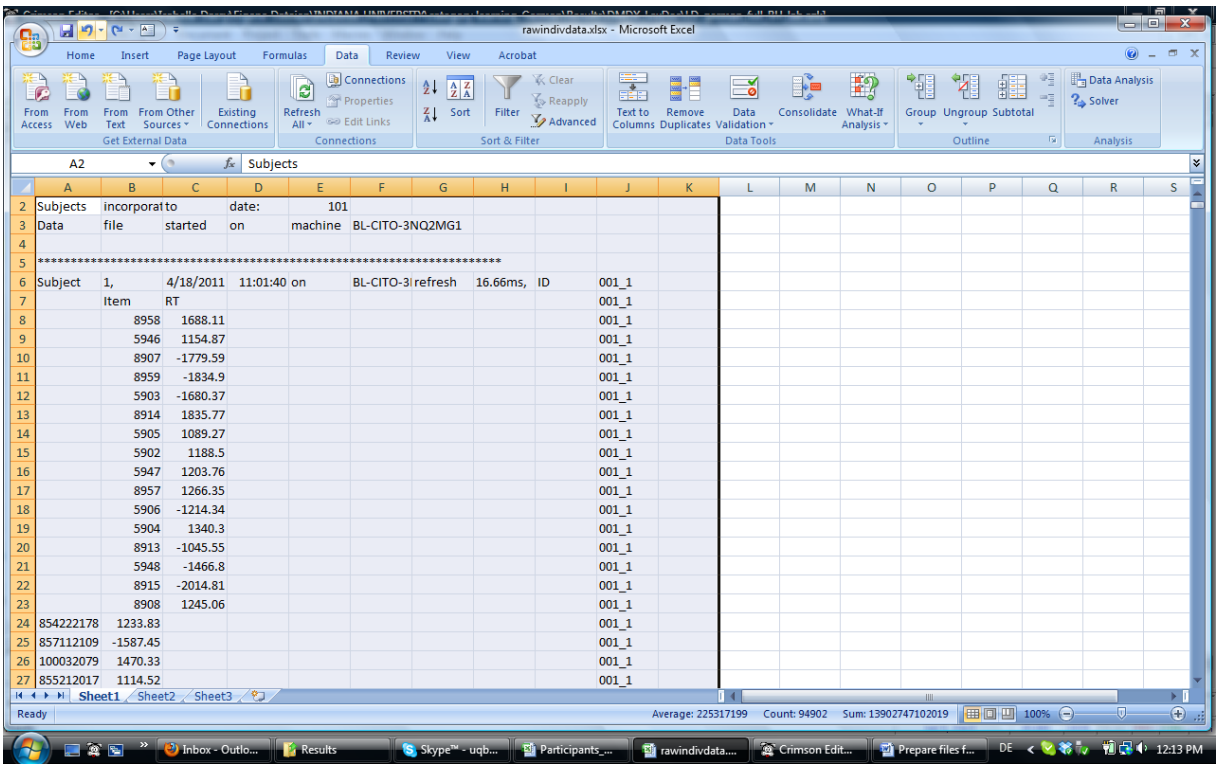


	A	B	C	D	E	F	G	H	I	J	K	L	M
19803	100031008	565.17								015_1			
19804	800031082	-2500								015_1			
19805	!	Display	error	at	msec	984168.53	tick	59065	in	015_1	857211075	frame	+
19806	!	moved	into	video	memory	1	tick	late		015_1			
19807	!	(previous	frame's	duration	will	have	been	longer)		015_1			
19808	857211075	504.94								015_1			
19809	800031019	-747.72								015_1			
19810	853111088	-2500								015_1			
19811	800031055	-724.28								015_1			
19812	856121167	-896.42								015_1			
19813	800031023	-928.03								015_1			
19814	800031056	748.87								015_1			
19815	154221177	-838.48								015_1			
19816	100031038	1015.32								015_1			
19817	800031024	701.64								015_1			
19818	100031069	955.19								015_1			
19819	157111110	1396.73								015_1			
19820	857211079	510.4								015_1			
19821	158221186	-957.25								015_1			
19822	154121209	921.49								015_1			
19823	100031098	-2500								015_1			
19824	158121214	-801.3								015_1			
19825										015_1			
19826	**********												
19827	Subject	65,	4/20/2011	10:58:18	on	BL-CITO-H	refresh	16.66ms,	ID	016_1			
19828		Item	RT										

8. Once you have copied, your cursor will go back to the first line (so where you started with 015_1). That's annoying, because you need to quickly move down to the next cell 016_1.
    a. To do that: hit Ctrl + down-arrow (without Shift, which is to select). The cursor will jump to the last cell filled, which is just 2 cells above the one you need to be, the 16_1. So now you just need to go down 2 cells.

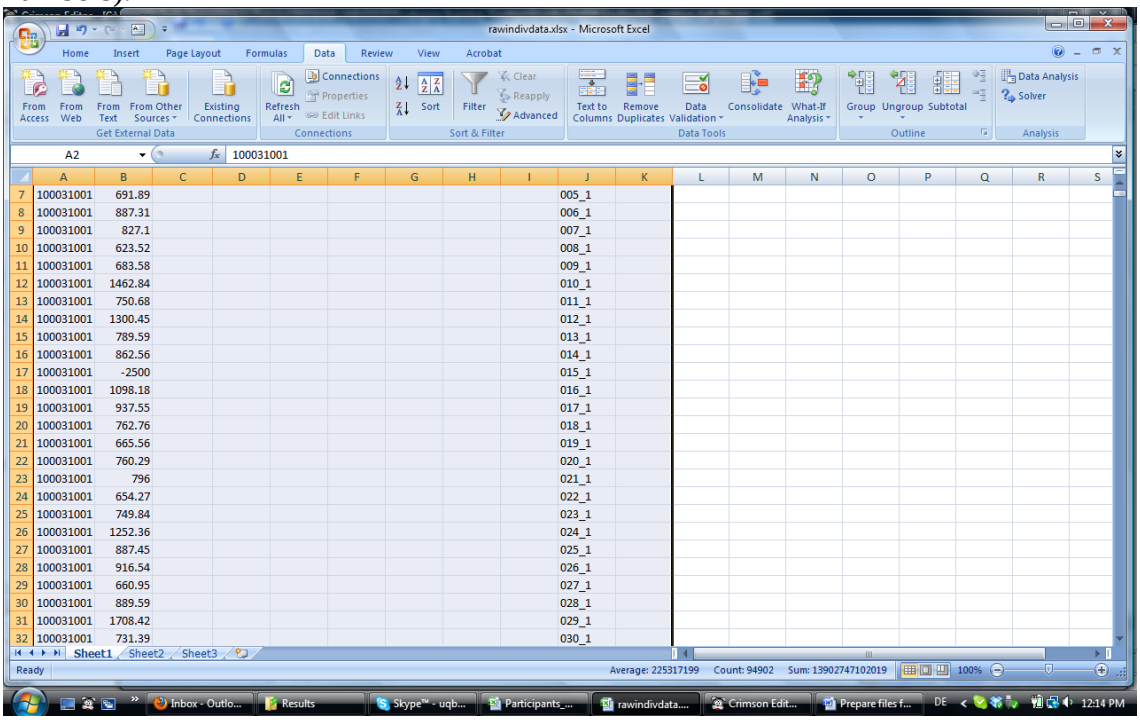
9. **Now, do steps 1-8 again with the next one, until you are done with all subjects. This is really important to do without mistakes. Keep your concentration. If you mess up the subjects, your results will be wrong.**

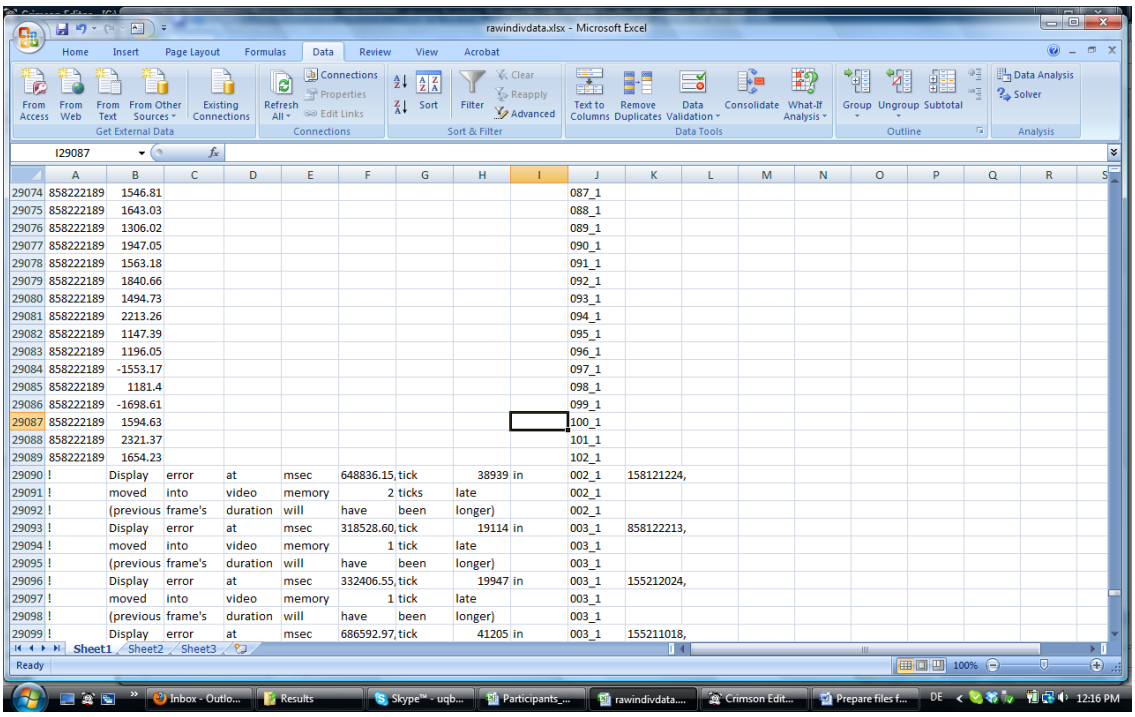
Now, you will start "tuning up" the file, so that you can input it into SPSS.



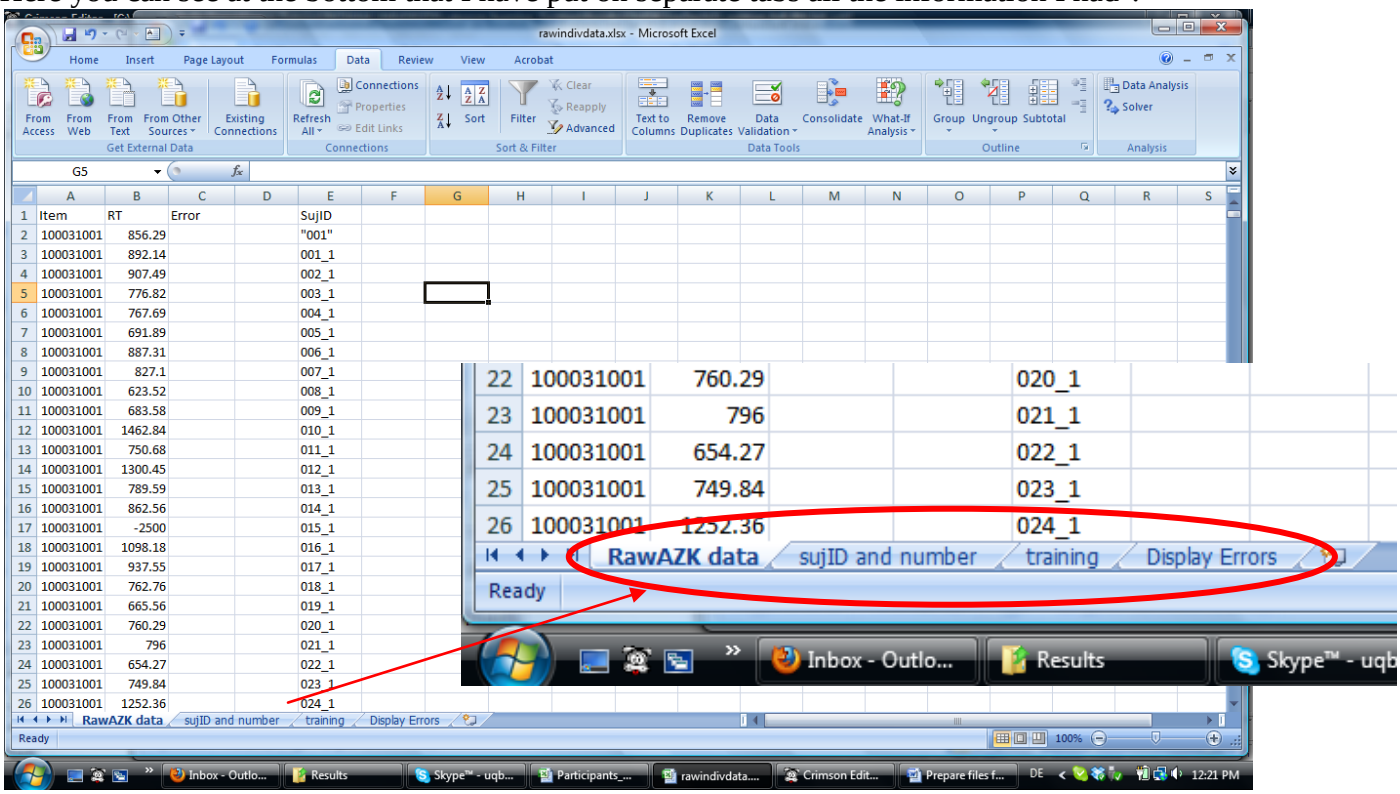
That‘s how your file now looks like.

Here: Sort everything according to column A (the item numbers) AND Column J (The ID numbers).





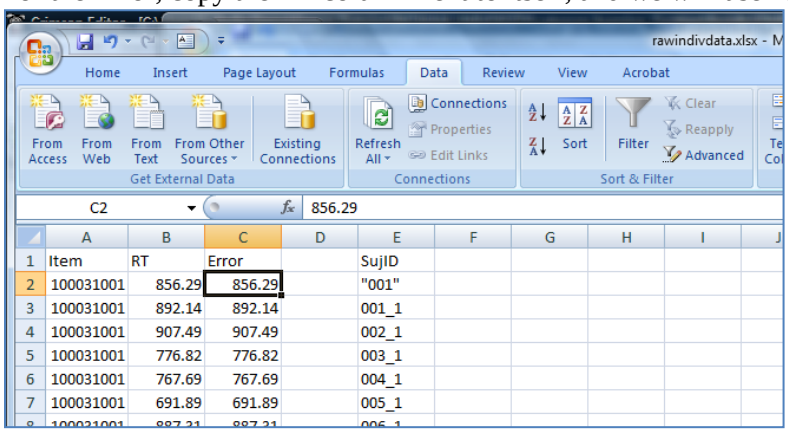
Now start to tune-up your raw data file: "clean out" all lines of the file that you won't use for analysis, that is all the display errors, the lines of subject ID, etc. I usually don't DELETE that information, I save it on separate tabs.
Here you can see at the bottom that I have put on separate tabs all the information I had[1].



[1] The "001" subject number is just another regular subject number that I used for the debug

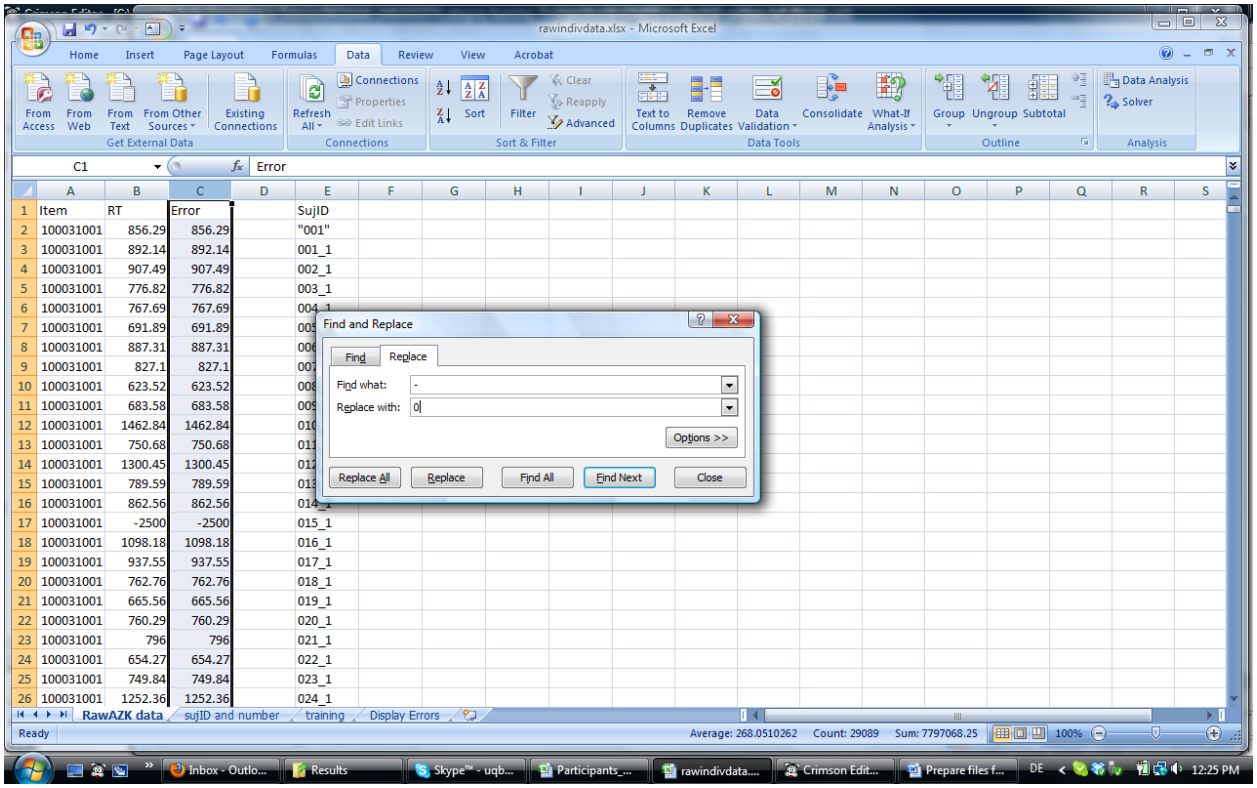
Copy all the subject info on one, the training items on another, and if you have display errors, also copy these in a separate tab. Your "rawAZKdata" tab should only contain the items from the main test (without the familiarization or training items).

Now, we will start filling in all the different columns we need for the stats analysis. The first one is to code "ERROR RATE". Create a column, with a header called "Error" next to the RT column
For the Error, copy the RT column next to itself, and we will use "copy and replace" to "clean it up"



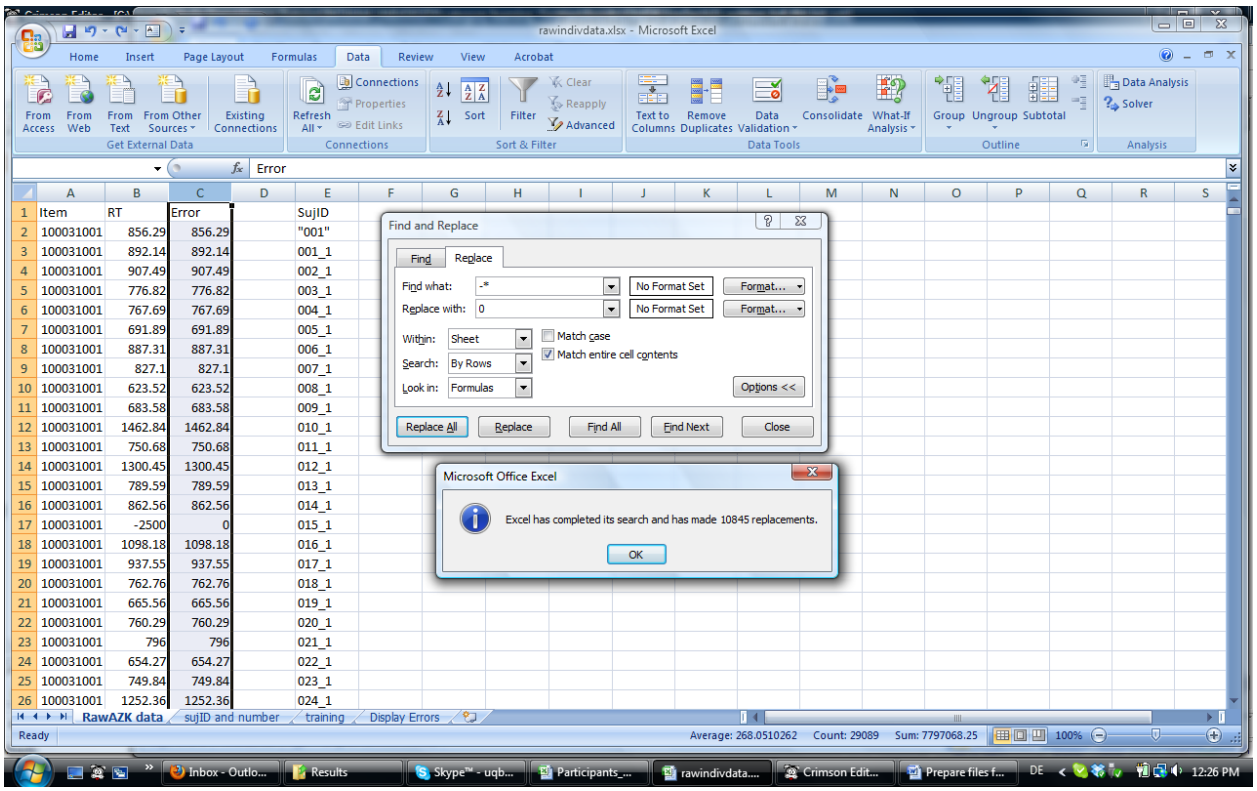
	A	B	C	D	E
1	Item	RT	Error		SujID
2	100031001	856.29	856.29		"001"
3	100031001	892.14	892.14		001_1
4	100031001	907.49	907.49		002_1
5	100031001	776.82	776.82		003_1
6	100031001	767.69	767.69		004_1
7	100031001	691.89	691.89		005_1

Right now, error is a copy of RT.
Remember that an error is always coded with a "-" in front of the RT. So to code errors out of the RT should be straightforward. We can replace all the negative RT with "0" and all the others with "1"



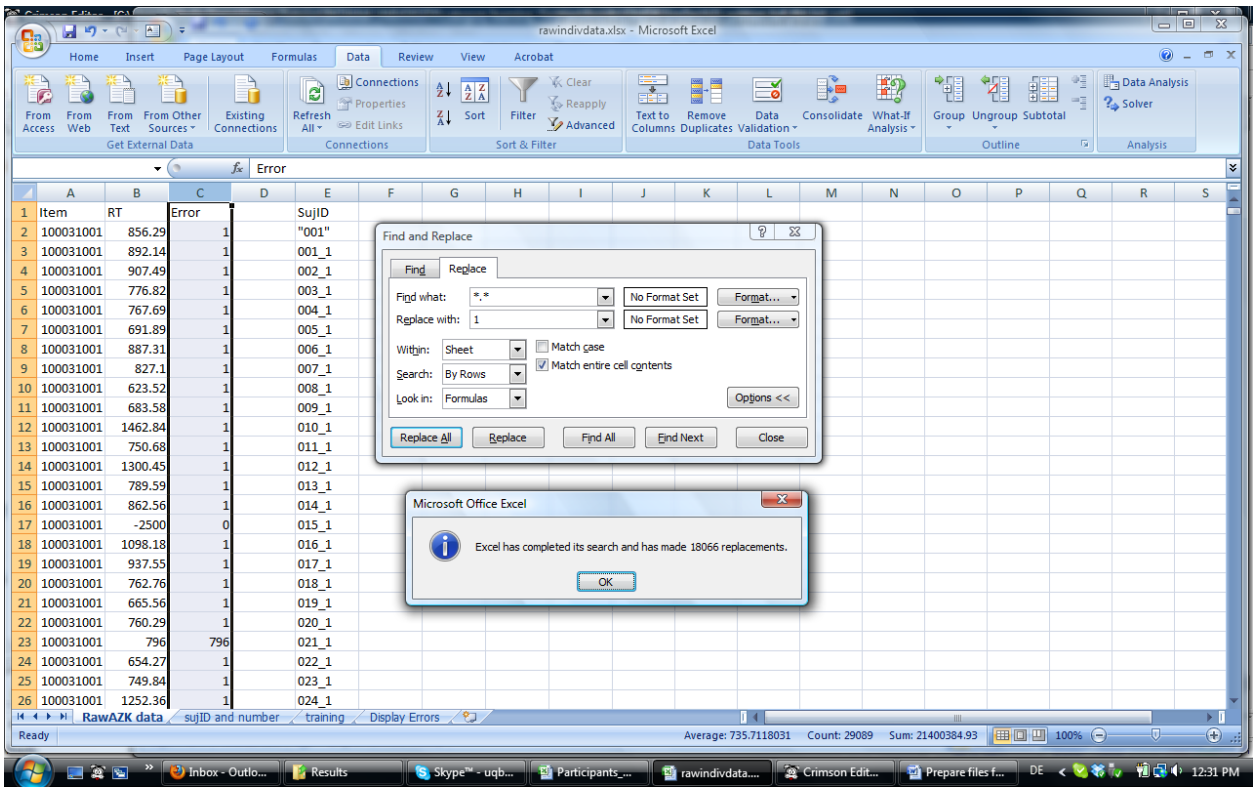
	A	B	C	D	E
1	Item	RT	Error		SujID
2	100031001	856.29	856.29		"001"
3	100031001	892.14	892.14		001_1
4	100031001	907.49	907.49		002_1
5	100031001	776.82	776.82		003_1
6	100031001	767.69	767.69		004_1
7	100031001	691.89	691.89		005_1
8	100031001	887.31	887.31		006_1
9	100031001	827.1	827.1		007_1
10	100031001	623.52	623.52		008_1
11	100031001	683.58	683.58		009_1
12	100031001	1462.84	1462.84		010_1
13	100031001	750.68	750.68		011_1
14	100031001	1300.45	1300.45		012_1
15	100031001	789.59	789.59		013_1
16	100031001	862.56	862.56		014_1
17	100031001	-2500	-2500		015_1
18	100031001	1098.18	1098.18		016_1
19	100031001	937.55	937.55		017_1
20	100031001	762.76	762.76		018_1
21	100031001	665.56	665.56		019_1
22	100031001	760.29	760.29		020_1
23	100031001	796	796		021_1
24	100031001	654.27	654.27		022_1
25	100031001	749.84	749.84		023_1
26	100031001	1252.36	1252.36		024_1

One possibility to recode the RT into an error rate is to do a search looking for the "–" symbol and replace the whole cell with "0". **HOWEVER, This won't work**, because excel will only look for the specific – sign, and replace only that sign with 0, so that it won't even show (-2500 will turn into 02500, and therefore into 2500)… Exactly what we don't want.



You have to search for "-*" (minus star). This works, because we look for "- followed by anything" , anything being represented by the *. Then, clicking the box "match entire cell content" will replace the whole cell with a 0, and you see, the -2500 has been replaced by a 0.
Now the trick question: how do we replace the other numbers with 1?
well, we could go step by step, asking it to look for certain numbers, such as "1", or "6" and then continue until there's nothing left but 1s. Or…. we can try to use the dot, which is for most RTs the case.

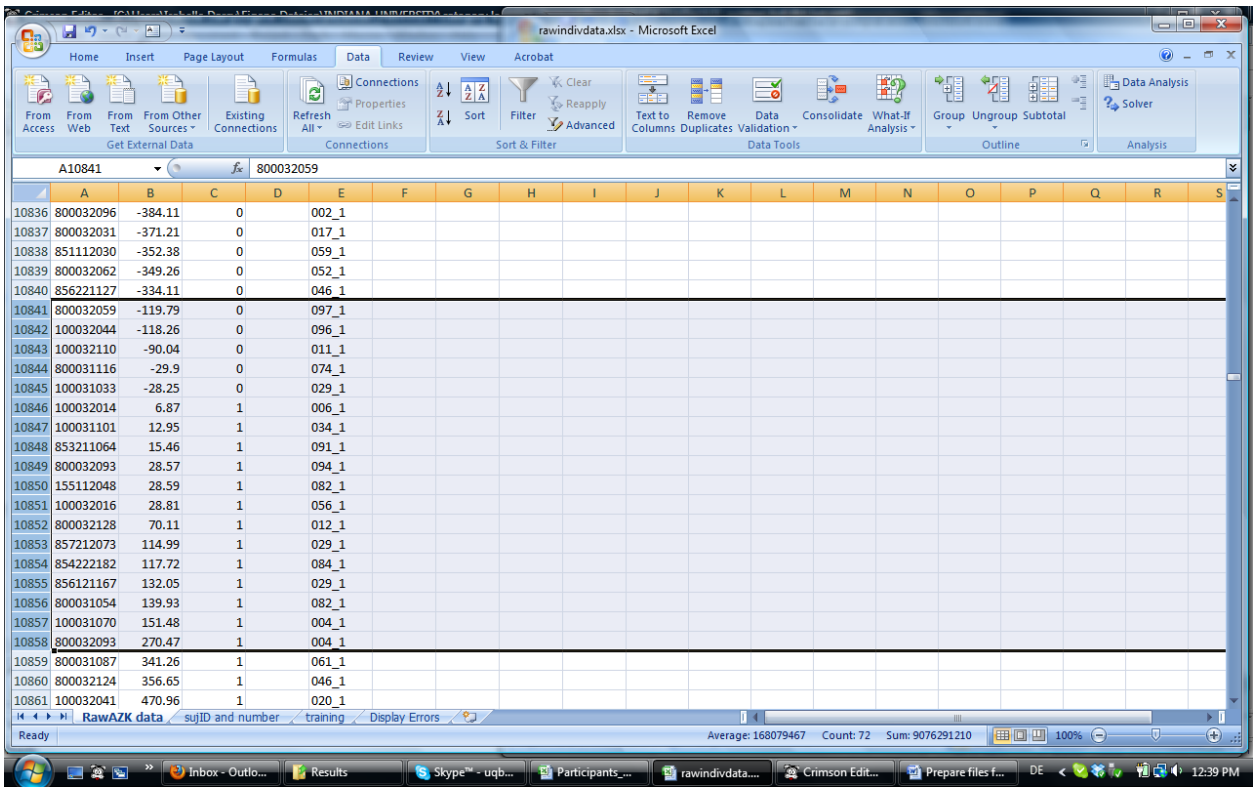
If we put the dot between two ** (*.*) , it works.
A few RTs were "round", so they didn't have a dot. We can either replace those now by going with a couple numbers (they probably share some numbers),



ANOTHER OPTION ALTOGETHER: which works well, too, instead of using "find and replace": Sort the whole data sheet according to the column "Error" (while the Reaction times are still in there), from smallest to largest and replace all the negative ones with 0, and all the positive ones with 1 (only in the "error" column, of course!)
To select the whole thing quickly: Ctrl+Shift+downArrow (selects the whole thing in one column vertically) and then, Ctrl+Shift+rightArrow (selects the whole thing over several columns).
To sort: Data-Sort-By column B in my case. All the RT are now sorted from smallest to largest, and we can replace everything in the column next to it (C) with 0 and 1, depending whether it is a hit or a miss. (any negative RT means that there was an error)

In the screenshot below, you can see how all the negative RT (preserved in the next column) are accompanied in our "error" column with a "0" and all the positive RT are accompanied with a "1"

Next we need to eliminate all the items with an RT below 300 ms.



	A	B	C	D	E
10836	800032096	-384.11	0		002_1
10837	800032031	-371.21	0		017_1
10838	851112030	-352.38	0		059_1
10839	800032062	-349.26	0		052_1
10840	856221127	-334.11	0		046_1
10841	800032059	-119.79	0		097_1
10842	100032044	-118.26	0		096_1
10843	100032110	-90.04	0		011_1
10844	800031116	-29.9	0		074_1
10845	100031033	-28.25	0		029_1
10846	100032014	6.87	1		006_1
10847	100031101	12.95	1		034_1
10848	853211064	15.46	1		091_1
10849	800032093	28.57	1		094_1
10850	155112048	28.59	1		082_1
10851	100032016	28.81	1		056_1
10852	800032128	70.11	1		012_1
10853	857212073	114.99	1		029_1
10854	854222182	117.72	1		084_1
10855	856121167	132.05	1		029_1
10856	800031054	139.93	1		082_1
10857	100031070	151.48	1		004_1
10858	800032093	270.47	1		004_1
10859	800031087	341.26	1		061_1
10860	800032124	356.65	1		046_1
10861	100032041	470.96	1		020_1

Examine the RT that are around 0. Any RT that is below 300 ms should be trimmed or the item discarded, because that is too fast a RT for being a true response. It is more likely a delayed response to a previous item (which is then most likely an error).

These highlighted items in the above screenshot would need to be deleted. I would not delete them all, I would copy and paste them in a new tab. But I cut them from the main result tab.

The same goes for high RTs if you don't have a cut-off. In our case, we do, so that no RT can be really longer than 2500 ms. (and we don't have anything really longer than that)

Get the number of such items with a fast RT (in my case, 18 out of xxxx) and indicate the percentage of these items. In my case, it is 18 out of 29088 datapoints = 0.000619%, not bad, really.

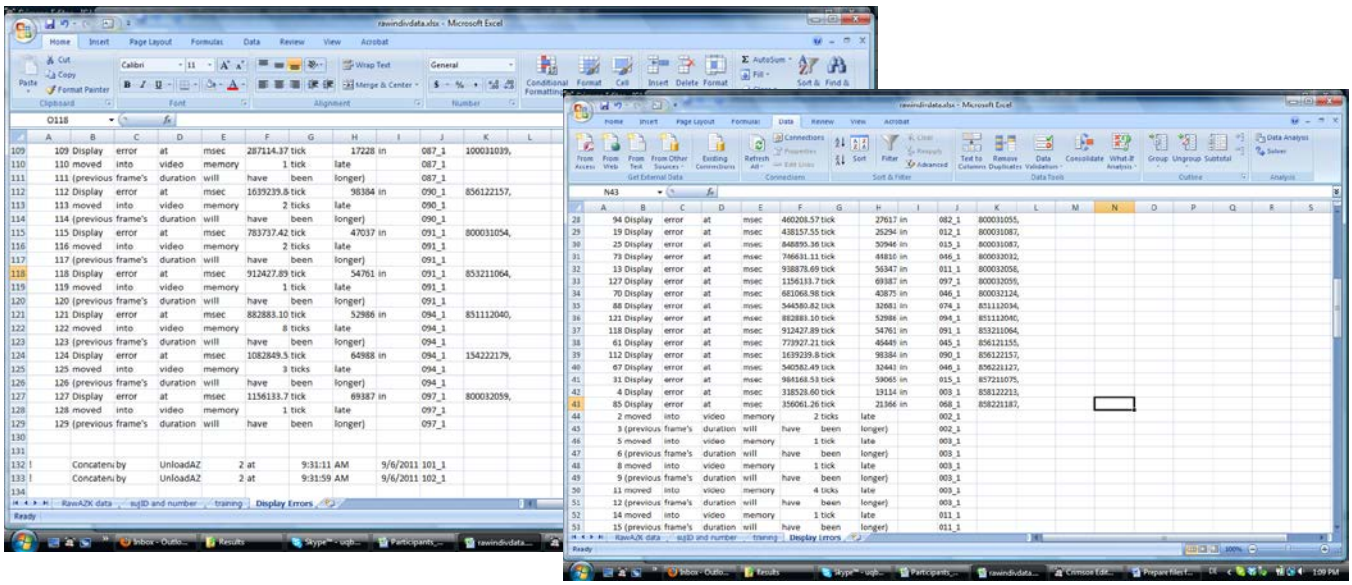
Similarly, a few other trimmings may be performed. (compare with the DMDX spc files where you select options. You should do the same options in your excel file)

EXKURS: THE DISPLAY ERRORS

In particular, the "display errors" items (that you saved extra) may be consulted to see whether these items RTs need to be either trimmed (set to the mean) or deleted. In our case, I prep that data also using "sort" functions. First, I add a column with running numbers from 1 through the last line of this data, so that I will later be able to restore the original order it came in.

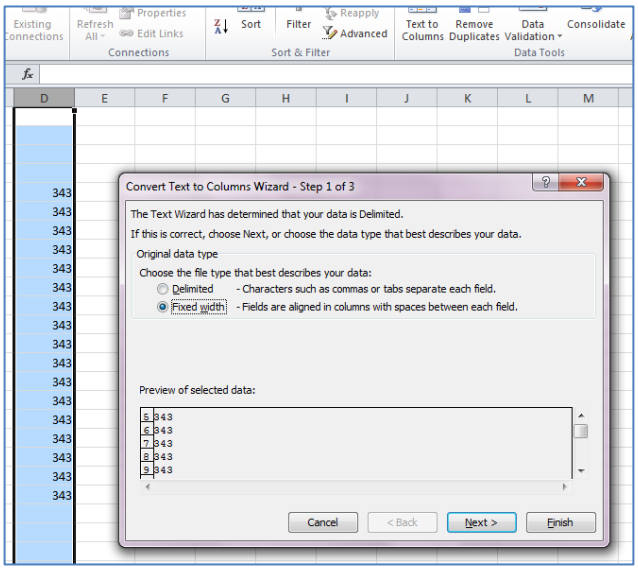
Then, I sort the dataset according to the column "K" which is the item number.

The layout before and after is presented below:

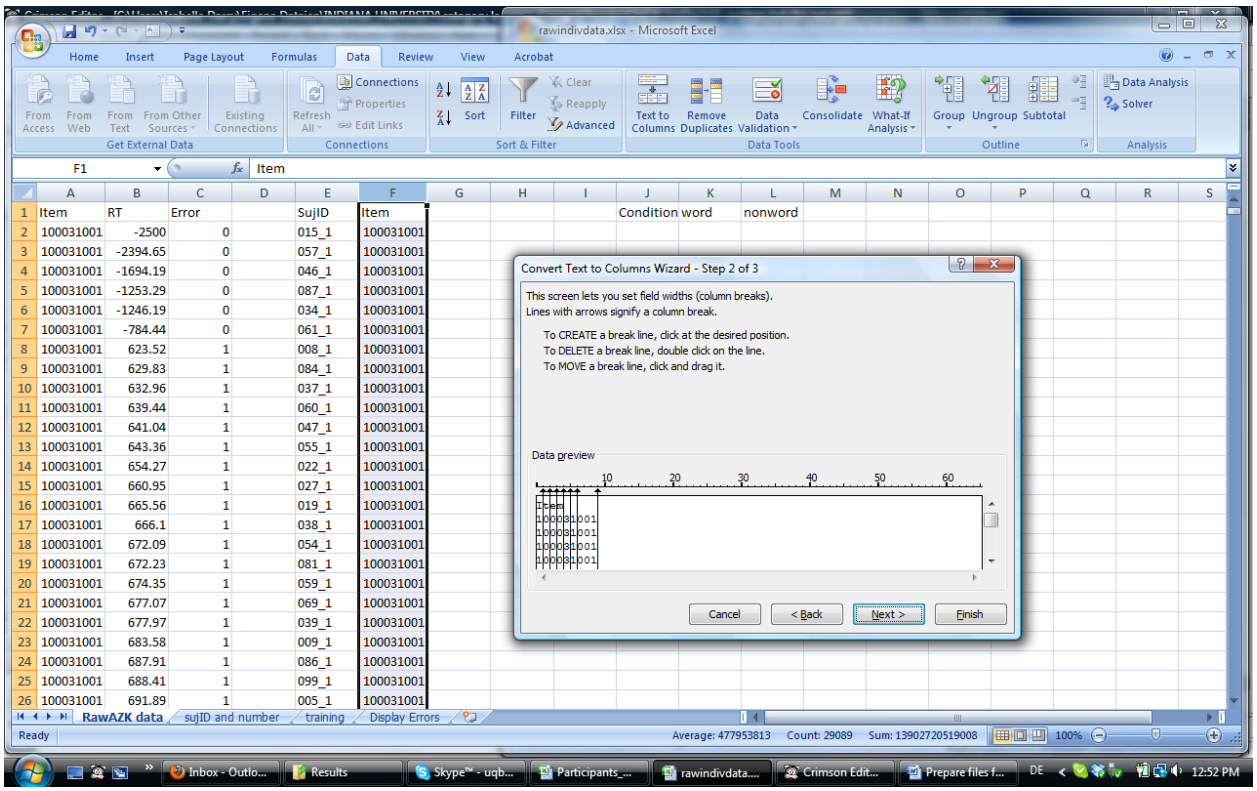


This is why the first column is important… because the actual duration of the delay (in ticks) is presented on a separate line. So if you sort it without having a way to restore the original display, you may lose that information of which item had a delay of how many ticks. Anyway, it may be important later, even if it's not right now.
What we see is : We had 43 items late for a total of 29088 (0.001%), again, pretty good! So even if we actually delete these items, it is not too big a deal. For now, I leave them in.

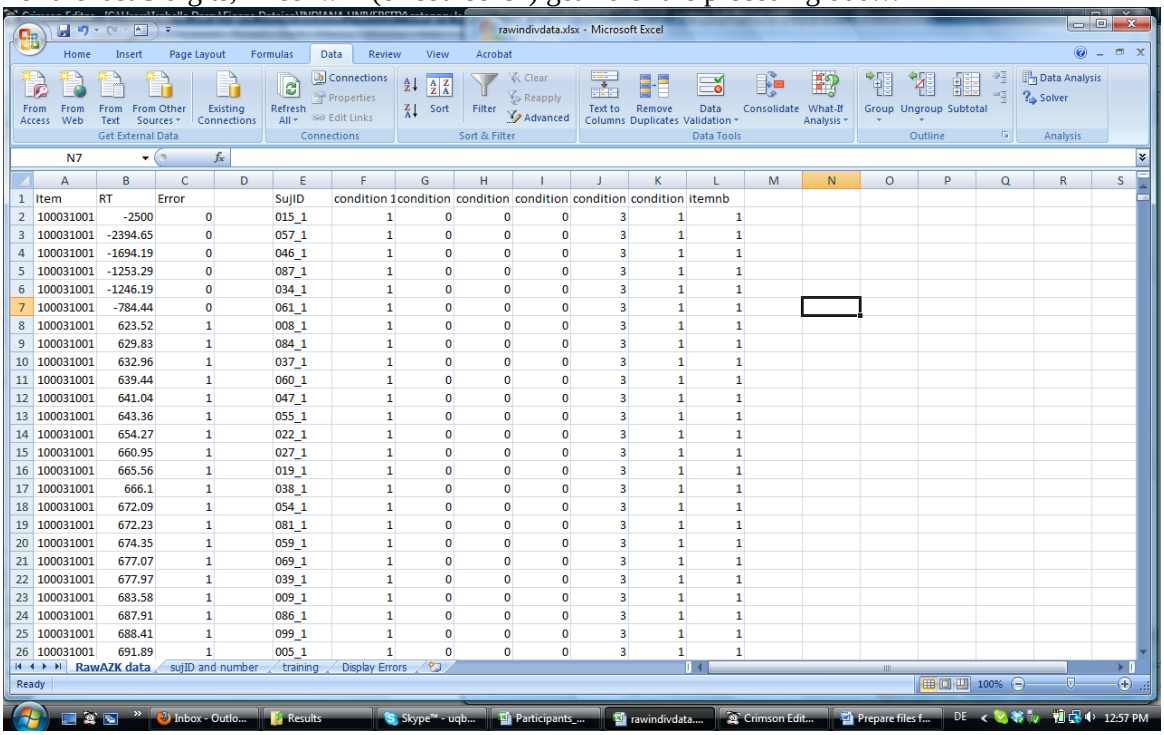
Now, the longer part will start. But it will be greatly helped if you have stored somewhere what the coding for the item number is. E.g. if your first digit is "vowel" vs. "consonant", and you gave it a code of 1 vs. 3, then, if you sort the item number from smallest to largest, you should be able to quickly add the condition for each of the lines.
Another option is to first "spread out" your item numbers onto separate columns, and use the "search and replace" afterwards. First, make a copy of the whole "item" column and click "text to columns" (under Data)



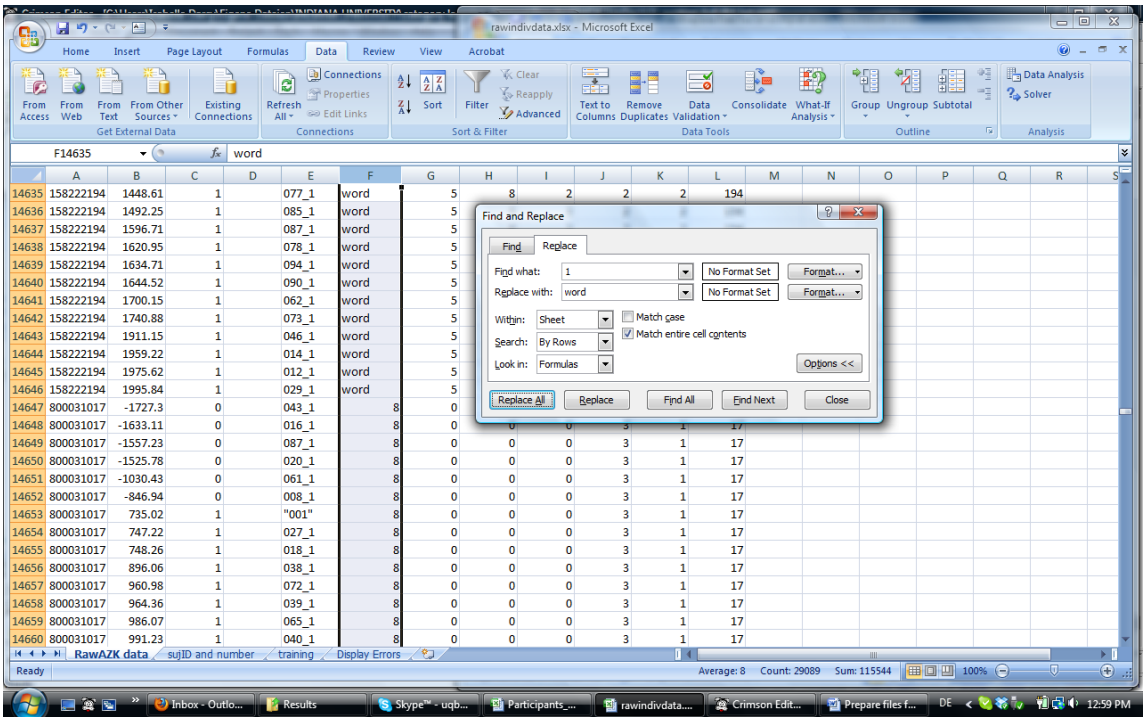
Choose "FIXED WIDTH" so you can select the boundaries yourself.

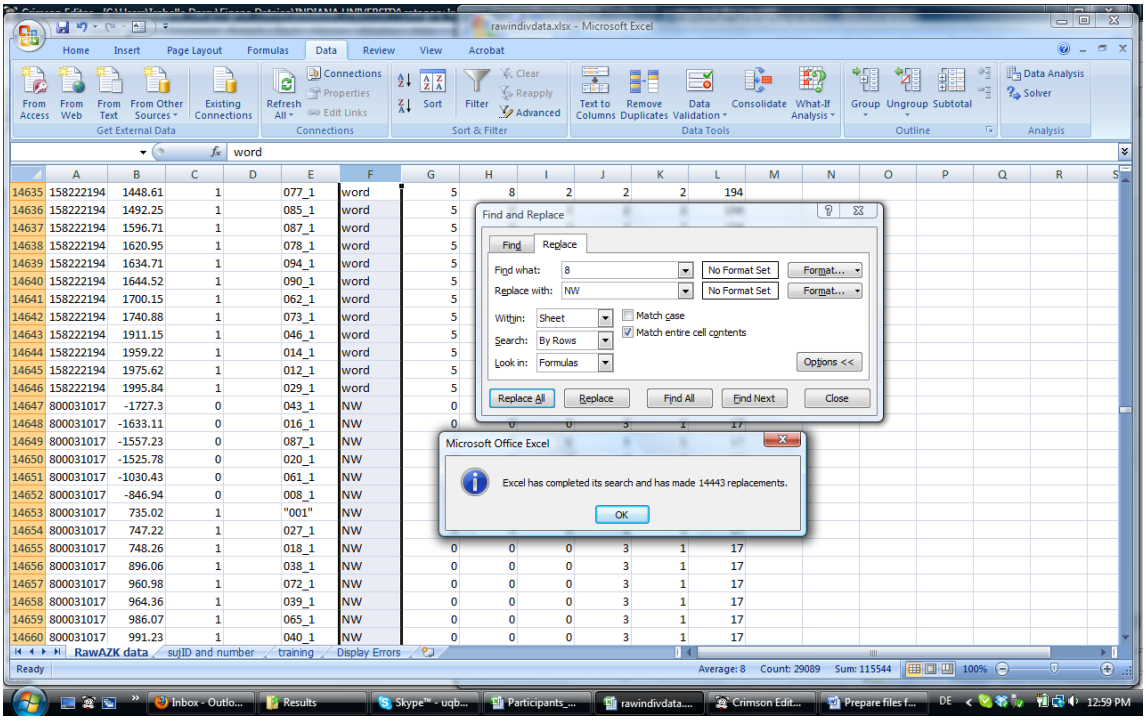


Click in between each item code (one or several digits, depending on your coding). In my case, each digit is a different condition, except for the last 3, which represent the actual running item number (from 001 through 260 for example).
Then, all these will be spread out on separate columns, ready to be "replaced" by condition names !!
For the last 3 digits, Excel will (of course ☹) get rid of the preceding 000…



Now, if your first digit is “word” vs. “nonword”, replace it accordingly. IN my case, “1” = word, and “8” = nonword. Easy. Find and Replace is opened with Ctrl+H or by going to Home – Find and select (far right), and then selecting “replace”.





Etcetera.

One thing you could do now also before moving on is create a column called “Group”, if you have several groups in your datafile. If your subject ID does not code the group, you’ll have to do that manually. But if it does (like in my case, “_1” is Group 1, for instance), you can use the

same technique as above for separating data from one column onto several, and so create a column “Group” next to the subject column. I copy the entire column “subject ID” at the end of all the columns I have, and separate the subject nb (001) from the group number (1) using the “Delimited” function of “text to columns”. I use the “_” as delimiter, and that’s it! The subject number (001) is in one column, and the group (1) in the other. Then I can just delete the repeated subject number column, and type “group” in the first line of the new column. If your subject ID does *not* encode group, I recommend doing it manually at this point. You’ll have to order the file by column “subjectID” and then manually, add a code for the group (such as “ADV” “BEG” and “NS”, or “1”, “2”, “3”, etc….). That’s longer, and make sure you don’t mix up participants in the different groups.

When all is done, you are absolutely ready to bring this into SPSS and get your ANOVAs done!