CHAPTER EIGHT

Testing hypotheses

Key words: causality – dependent *t*-test – *F*-test – hypothesis testing – hypothesis testing using χ^2 – independent *t*-test – tails– testing experimental and non-experimental data – ANOVAs

8.1. Hypotheses, causality and tails

Without explicitly mentioning it, to a certain extent we have already touched on hypothesis testing in Chapter Seven when we discussed the chi-squared test and correlation. In Chapter Two, we defined hypotheses as statements about the potential and/or suggested relationship between at least two variables and said that a hypothesis *must* be phrased in such as way that they can be proven right or wrong – tautological hypotheses are not permissible. Now it is about time that we specify this definition a bit more. In the previous chapter, we have seen that correlation only indicates the strength of *association* between two variables, not *causality*. Indeed, frequently when we carry out a correlation analysis between variables in our data, we will find that several of them correlate, that is, their scores occur in a particular pattern with each other. Using (multiple) regression analysis, we are able to make predictions or forecasts about what may happen given the current pattern of our data, but again, regression does not give us any empirical evidence about causality.

In the example from my own research, we have seen that age of onset and length of exposure correlate; yet, we cannot say that age of onset *causes* length of exposure. Rather, the two variables occur in a distinct pattern. In a fictive example, we may find that there is a positive correlation between the number of children a linguist has and number of pages their books have (the more children, the longer the books), but we would have to go a long

way to argue that there is a causal relationship, that is, that the children are *causing* them to write longer books.

In a typical hypothesis, on the other hand, we do want exactly this: we want a clear constellation of one (or more) independent variables and a dependent variable. In other words, our hypothesis should contain a constellation where there is at least a thorough theoretical basis for assuming causality between the independent and the dependent variable, with the former causing the latter. In an experimental setup, where the researcher has full control over the independent variable(s), any change in the independent variable(s) would cause a change in the dependent variable, given that there is a causal relationship. If there is no relationship, we will not see a change in the dependent variable. We have outlined the three criteria for causality in Section 2.5.

A clear definition of the independent and the dependent variable is also important for the choice of statistical test and their significance level. If we predict causality between variables A and B in such as way that A influences B, we speak of a *1-tailed hypothesis*, and we will eventually use a significance level that reflects this. If you have a look at the tables of critical values in Chapter Ten, you will see that the significance levels for 1-tailed hypothesis are exactly 1/2 of that of 2-tailed tests: since we only consider causality in one direction, we have to ensure that we account for significance in only one direction, that is, we want to be confident that A influences B only, not the other way round. If our hypothesis is *2-tailed*, where *either* A influences B *or* B influences A, we have to check for significance in *both* directions.

Strangely enough, quantitative research has traditionally always assumed that there is no relationship between two variables, and statistical tests are based on this assumption. This is reflected in the way we phrase hypotheses: we generally phrase our hypotheses in such as way that they imply no relationship between A and B. This is also known as the *null hypothesis*, or H_0:

H_0: There is no relationship between social class and use of non-standard forms.

The task of our statistical test, and indeed our entire research, is to prove or disprove H_0. Paradoxically, in reality most of the time we are interested in disproving H_0, that is, to show that in fact there *is* a difference in the use of non-standard forms between socio-economic classes. This is called the *alternative hypothesis* H_1 or H_A. Hence:

H_0: There is no relationship between social class and use of non-standard forms.

H_A: There is a relationship between social class and use of non-standard forms.

The tools discussed in Chapter Seven are all measures of association, that is, they tell us to what extent two variables co-occur. By its very nature, especially the Pearson correlation requires a comparatively large amount of different scores for each variable in order to detect a relationship: we need, for example, several different age of onset scores and several different proficiency scores if we would like to show a relationship between these variables. In other words, we need a variety of scores for the independent variables (here probably age).

However, sometimes this is either not possible, or we deliberately want to avoid this. Traditionally, real experiments are set up in such a way that only one independent variable is deliberately manipulated (with other variables being held constant) and the effect on the dependent variable is measured – see Chapter Three. A typical example is that of a study on the influence of background knowledge on learners' listening comprehension:

> The experimental group received some treatment in the form of topic familiarity, and their background knowledge was activated. Then a 50-item TOEFL test of listening comprehension was administered to both experimental and control groups. A statistical analysis of the results provides some evidence in support of the effect of background knowledge on listening comprehension. (Sadighi and Zare 2006: 110)

In terms of research design, Sadighi and Zare's study is a typical experimental study. The experimental group receives the 'knowledge activation' stimulus; the control group receives just 'normal' input.

	Step 1	Step 2	Step 3	Step 4
Experimental group	Obtain pre-test score A	Stimulus: 'knowledge activation'	Obtain post-test score X	Statistical Analysis
Control group	Obtain pre-test score B	No stimulus	Obtain post-test score Y	

Step 4 leaves us with four options for analysis:

1. Compare A with B; that is, compare the two pre-test scores.
2. Compare X with Y; that is, compare the two post-test scores.
3. Compare A with X; that is, compare pre- and post-scores for experimental group.
4. Compare B with Y; that is, compare pre- and post-scores for control group.

We will discuss each step in the course of this chapter. We can obtain a rough idea about the relationship between pre- and post-scores by running a simple Pearson correlation: ignoring the questions of causality and whether pre- and post-test scores a significantly different, the Pearson correlation will tell as whether there is a relationship between the two scores, which direction is takes, and how strong it is. For the experimental group, we would expect to see a positive correlation, as we assume that the stimulus improves respondents' post-test scores, so there should be a pattern whereby post-test scores are higher than pre-test scores. If the stimulus does not have any effect, there should not be a difference in pre- and post-test scores, and the correlation coefficient will be small, indicating a weak relationship. For the control group, we would expect exactly that: there is no stimulus, so improvement should be less pronounced, indicated by a small correlation coefficient.

The question is, how exactly can we compare pre- and post-test scores statistically? How can we show that the post-test scores are statistically significantly higher than the pre-test scores? That is where the *t*-test can help us.

8.2. The *t*-test: Preliminaries

The *t-test* (sometimes also known as the *Student's t-test*) allows us to compare the scores, or more precisely the arithmetic means of either two groups of respondents or two sets of data from the same sample. We will discuss both issues here subsequently. First, let's have a look at some basic underlying issues.

In our discussion of the arithmetic mean in Chapter Six we have said – and shown – that two sets of data can have the same arithmetic mean, even if they are substantially different. Roughly speaking, the t-test is a statistical procedure that compares the arithmetic means of two groups of data while taking their variability (i.e. their standard deviation or variance) into account. Hence, it allows us to draw conclusions whether or not there is a real difference between two groups. To illustrate this problem, Table 8.1 shows two sets of (fictive) scores for ten people as well as the mean score for each set of scores. If we only look at the mean, or if in an article there was only the mean reported, we could assume that the two sets of scores are different: the mean for score B is 3.3 points higher than the mean score for A, so we could easily argue that B scores are higher than A scores. Yet, if we have a closer look at the actual data, we can see that A and B only differ by a single pair of scores: only R10 has a higher B than A score; for all other respondents the two scores are identical. So, the difference in means for the two sets is based on only one differing score in a set of 20 scores overall. One score for one person makes all the difference. Imagine this table were

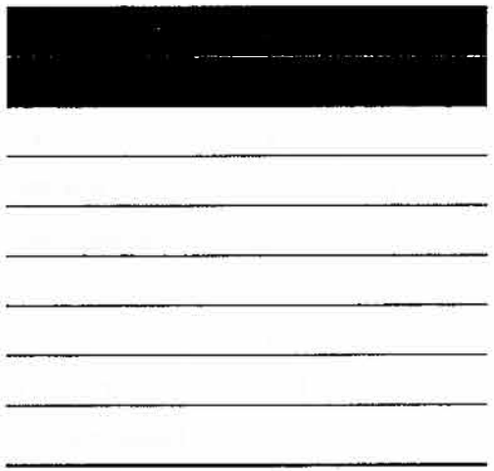
TABLE 8.1 Fictive example

	Score A	Score B
R1	2	2
R2	3	3
R3	4	4
R4	6	6
R5	7	7
R6	8	8
R7	12	12
R8	13	13
R9	16	16
R10	17	50
Mean	8.8	12.1

the results for Sadighi and Zare's experimental group: would it not be odd to argue that the post-test scores (score B) are higher than the pre-test ones, hence the stimulus must have some impact? The answer is, yes, it would be pretty odd. How can we validly argue that the post-test scores are different from the pre-test scores only based on the arithmetic means, when we can see that in fact there is next to no difference, and that the mean by its very nature is notoriously sensitive to any changes in the values that it is based on? We obviously need a solution for this – the *t*-test.

Note that as a parametric test, the *t*-test can only be applied to normally distributed data, and should ideally only be applied if we have at least ten items (respondents, scores, etc.) per group; as you can see, sample size and amount of data is a recurring pattern. *t*-tests come in two main shapes: t-test for independent samples, and t-test for dependent sample. We will discuss independent and dependent t-tests subsequently. But before we go over to the *t*-test, we need to look at something else.

One section removed here



8.4. *T*-test for independent samples

As should be obvious from the name, the *independent samples t-test* (sometimes also called *unrelated t-test*) is used to compare the means of two different *groups*. In Sadighi and Zare's study, we have two groups, experimental (EG) and control group (CG), each one with its own set of results for both pre- and post-test. Respondents belong to either EG of CG, so there is no overlap of respondents and/or scores; hence, they are independent from each other and one group cannot influence the scores of the other.

As for hypotheses, we can assume that:

H_0: There is no difference between EG and CG scores after stimulus was introduced.

H_A: There is a difference between EG and CG after the stimulus was introduced.

Note that the H_0 assumes no change; yet, it is change that we are really interested in. We assume that the stimulus influences respondents' test performance in such a way that ultimately the EG scores are higher than the CG scores, so we can use a 1-tailed hypothesis. If we do not want to make any prediction, we use a 2-tailed test (as, in fact, Sadighi and Zare did). Some people use a 2-tailed test by default for everything. This is not wrong but will have an impact on the significance level applicable to interpret our data, so it might make a difference. If in doubt, you may want to run both a 1-tailed and a 2-tailed test – after all, it is fairly easily done with the appropriate software.

Two aspects should strike us straightaway: the EG scores are higher than the CG scores by 9.29 points, so at a first glance, there seems to be a difference and we could reject H_0. However, we also see that CG scores are slightly more dispersed. The t-test will account for exactly this. The *t*-test is rather tedious to calculate, so will restrict our discussion here to how to calculate it with Excel. The programme will ask you about equal and unequal variances. Our calculations of the *F*-test have just shown that the variances between EG and CG are not significantly different, so we can use the 'equal variances' version of the test. As before, if you use SPSS, the software will make the decision for you (in fact, it provides you with results for both). Note that some people will by default run both versions of the test, to be on the save side, mainly because there are reliability issues with the *F*-test calculated by Excel. While there is nothing wrong with being cautious, bear in mind that contradictory results from the equal variances and unequal variances test may be difficult to interpret. Seeing that this is an introductory textbook which cannot cover the intricacies of the various statistical tools, we trust our *F*-test result.

Doing statistics using Excel: *Independent t-test* (detailed display)

As with the *F*-test, there are two ways of doing *t*-tests with Excel. The former is the preferable.

1. Using the Data Analysis Tool:
   - Go to 'Tools', 'Data Analysis'.

- Choose '*T-test*: Two-sample assuming Equal Variances'. Obviously, if you have data where the F-test shows significantly different variances, you choose the 'Unequal variances' version of the test.
- For 'Variable 1 range' highlight the scores for the EG, and for 'Variable 2 range' the scores for CG.
- Leave the 'hypothesised mean difference' box empty.
- We are being picky today and want to make sure we have a sound significance level: choose Alpha 0.05 for a 95% confidence level. If we need to be less strict, we can adjust it to 0.1 (90% confidence) – but we really should not go any lower.
- Click ok.

(→See companion website for a clip on how to conduct a t-test in Excel.←)

You should get the following table (Table 8.4).

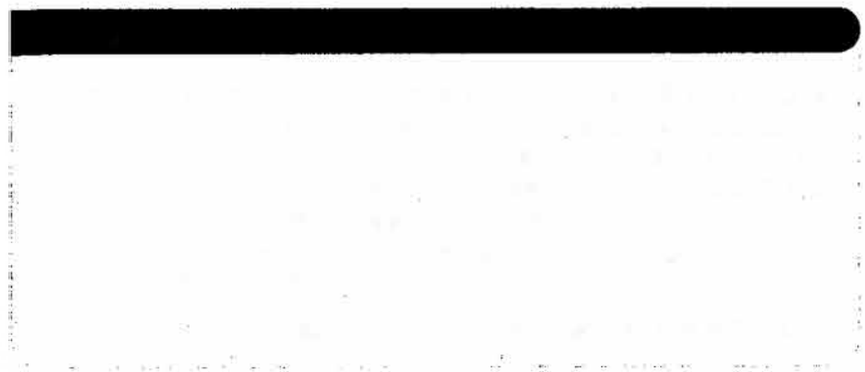
The first three lines give us the arithmetic means, variances and number of observations for each group. 'Pooled variance' is the combined variance of EG and CG, which is used to calculate the *t*-statistic – we ignore it here, and we also ignore the 'Hypothesized Mean Difference'. Df gives us the degrees of freedom, which for the independent *t*-test is the sum of observations of both samples minus 2 (i.e. 12+12–2=22). The '*t* Stat' row gives us the

TABLE 8.4 *T*-test between EG and CG. Excel output

	EG	CG
Mean	43.67	34.38
Variance	12.97	20.60
Observations	12.00	12.00
Pooled Variance	16.78	
Hypothesized Mean Difference	0.00	
df	22.00	
t Stat	5.56	
P(T≤t) one-tail	0.00	
t Critical one-tail	1.72	
P(T≤t) two-tail	0.00	
t Critical two-tail	2.07	

It's not necessary to know how to do the calculations or use the Excel function, but you should be able to understand the logic of the t-test, and the difference between independent sample and paired sample t-tests.

result of the actual t-test. Here t=5.56. The most important rows, though, are P(T≤t) one-tail and P(T≤t) two-tail, as these give us the significance levels. We said at the beginning that we predict that the scores for the EG are higher than those of the CG after the stimulus, and we decided to use a 1-tailed test. We see that for the 1-tailed test, p=0.00, which is smaller than 0.05. This means that the difference in means between EG and CG is statistically highly significant. In other words, the stimulus seems to have made a real difference in performance between EG and CG. Had we not predicted which group would score higher, the 2-tailed test, too, is highly significant, again with p=0.00. In a research paper, we report the result somewhere along the lines of 'after the introduction of the stimulus, respondents in the EG group performed statistically significantly higher than those in the CG, with $t(22)=5.56$, p=0.00.'



If you used the Sadighi and Zare's (2006) data for the graph, you should get a graph similar to Figure 8.1. We can see that the means for EG and CG are different, as the bars are of unequal height. We also see that the two error bars – the vertical lines in the middle of each bar – do not overlap. This indicates that the means are statistically significantly different. In fact, the ends of both error bars are quite far apart, which strengthens the argument even further. Error bars that overlap indicate that the two means are potentially not significantly different; and the more the overlap of the error bars, the less likely a significant difference is. Note, though, that in any report, a bar chart with error bars alone will not do the job – you need to provide the numerical values, too!

Is this really sufficient information to argue that the stimulus 'knowledge activation' results in better performance? If you spend a minute or so thinking about it, you will realise that we have only compared the two post-test scores. It does not actually tell us how the stimulus influenced the result. Sadighi and Zare have also compared the pre-test scores of both

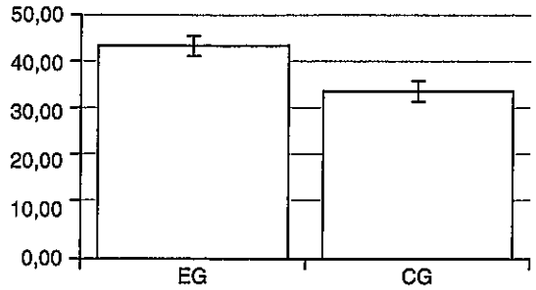


FIGURE 8.1 *EG CG bar chart.*

groups and found no significant difference. That is, both groups were at the same proficiency level before the stimulus was introduced. Yet, we still need a decent measure that tells us the actual change that occurs between EG and CG, o rather: within EG and CG. And this is where the *dependent t-test* comes into play.

8.5. Dependent *t-test*

The dependent *t-test* is also known as *paired t-test, related sample t-test* or *related measures t-test.* I am sure I could find even more terms if I kept trawling through literature. They all refer to the same test though. The main difference between the independent t-test discussed above and the dependent t-test is that in the latter, we compare two sets of data from the *same group* (or sample). In other words, rather than comparing two groups we compare scores from one group under two conditions.

Earlier on in this chapter, we have identified four options with regard to analysing Sadighi and Zare's data. With the independent t-test, we have chosen option two:

2. Compare X with Y; that is, compare the two post-test scores.

We have shown, and indeed Sadighi and Zare have shown that there is a significant different in performance between the two groups in the post-test, and we have seen that the pre-test scores were similar. Now, we need to compare the pre- and post-test scores for each group. Unfortunately, Sadighi and Zare do not give us the pre-test results, so we have to make them up (Table 8.5).

Task 8.2

Run a *t*-test for all *independent* constellations of the data in Table 8.5.

For the dependent t-test, we are interested in the differences between the scores *within* each group, that is, the differences between EG_pre and

TABLE 8.5 Pre- and post-test results

Respondent	EG_pre	CG_pre	EG_post	CG_post
A	22	22	40	27.5
B	24	24	40	30
C	24	25	40	30
D	26	26	40	30
E	27	27	42	32.5
F	30	31	42	35
G	32	32	44	35
H	34	34	44	35
I	35	34	46	37.5
J	35	35	48	37.5
K	36	37	48	40
L	37	37	50	42.5
Mean	30.17	30.33	43.67	34.38
STD	5.36	5.30	3.6	4.54

Source: Adapted from Sadighi and Zare (2006). Fictive values for EG_pre and CG_pre

EG_post, and CG_pre and CG_post, respectively. As both data sets for the dependent *t*-test come from the sample, we do not need to worry about variances and hence, we do not need to conduct an *F-test* first.

Task 8.3

Create a bar chart with error bars for the dependent t-tests EG_pre/post and CG_pre/post. Interpret the result.

Doing statistics using Excel: *Dependent t-test*

As with the independent *t*-test, go to 'Tools', 'Data Analysis'. This time, we choose '*t*-test: Paired Two Sample for Means'. For the dependent *t*-test, variable range 1 and 2 refer to the EG_pre and EG_post scores, respectively. Leave Alpha at the default of 0.05 for 95% confidence. Click ok.

You results for EG pre/post should look like Table 8.6.

TABLE 8.6 Dependent *t*-test for experimental group

	EG_pre	EG_post
Mean	30.17	43.67
Variance	28.70	12.97
Observations	12	12
Pearson correlation	0.94	
Hypothesized mean difference	0	
df	11	
t Stat	−19.86	
$P(T \leq t)$ one-tail	2.88E−10	
t Critical one-tail	1.80	
$P(T \leq t)$ two-tail	5.77E−10	
t Critical two-tail	2.20	

The interpretation of the results are almost identical to the interpretation of the independent samples test. Instead of a pooled variance, we now find the Pearson correlation coefficient, which indicates a strong positive correlation between the pre- and the post-scores for the experimental group: the higher respondents scored in the pre-test, the higher they scored in the post-test. You may want to plot a scatter graph to get visual evidence.

In the bottom four rows, we see that the difference between the two means is statistically significant, with $p<0.05$ in both 1- and 2-tailed test. That is, respondents in the experimental group scored significantly higher after the stimulus had been introduced. To complete our discussion, answer task 8.4.

Task 8.4

Run a *t*-test for the control group.

I am sure you have found in task 8.4 that the difference in means between the pre- and the post-test for the control group is statistically significant for both 1- and 2-tailed test with $p=0.000$ and $p=0.000$, respectively. It might be a small improvement, but a significant one. As before, if you draw a bar chart with error bars, you get some visual information in this, too.

Time to summarize all our results for the Sadighi and Zare (2006) study – including the fictive data. We have found that:

- There was no significant difference in performance between the two groups at the beginning of the experiment
- Performance for the experimental group increased significantly by 13.5 points after the stimulus was introduced.

- Performance for the control group increased significantly by four points when no stimulus was introduced.
- Performance of the EG is significantly higher than that of CG in the post-test, that is, after introduction of the stimulus in EG.

If we assume that all other factors were constant, we now have very strong evidence that the stimulus, that it, 'knowledge activation' significantly improves respondents' performance. Even though the control group's performance also increased, this increase was only marginal and can probably be explained as normal development. As regards causality, we shall refer back to our three criteria in Chapter Two:

a *X* and *Y* must correlate. We have seen above that EG_pre and EG_post scores correlate strongly and positively. So, this criterion is met.

b There must be a temporal relationship between *X* and *Y*. The post-scores, by their very nature, occur after the pre-scores, hence this criterion is met.

c The relationship between *X* and *Y* must not disappear when controlled for latent variables. This is the most difficult one to prove. We can only hope that external circumstances do not change over the period of the experiment. However, we have no reliable evidence that this is the case.

As such, with two out of three criteria met, we have a fairly good indicator that performance and stimulus are causally related; yet, the evidence is not conclusive.

Doing statistics using Excel: *t-test (significance only)*

Sometimes we may only want to quickly check whether differences between two groups are statistically significant or not. For this, we can use the =t-test function. As usual, either type it in directly, or use the function wizard.

=t-test(array1, array2, tails, type). array1/2 refers to the data of each group, tails to the number of tails (1 or 2). There are three types:

- Paired (=dependent sample) – 1
- 2-sample equal variance – 2
- 2-sample unequal variance – 3

For the calculation of the *t*-test in Section 8.2 – the independent sample – the function is:

=t-test(dataEG, dataCG, 1, 2)

The result we obtain only gives us the significance level.
(→See the companion website for an Excel template for t-tests.←)

8.6. Hypothesis testing, *t*-tests and non-experimental data

The Sadighi and Zare (2006) study is a convenient example to illustrate hypothesis testing with the help of the t-test: in a nicely controlled environment we manipulate one particular variable and measure the outcome. However, in linguistics we are frequently confronted with non-experimental data. And where there is non-experimental data, there is always plenty of scope for latent variables to influence any result. Trouble looms!

Another example from my own work: in my research, I looked at target-likeness in spoken English as a second language among Bangladeshi migrants in London. Data was collected by means of interviews and the elicitation of personal narratives – in other words, we have spontaneous speech in an uncontrolled setting. Table 8.7 summarizes the results for 12 respondents, including age of onset, length of residence in England ('residency') and the performance score GPI – the higher the score, the more target like the language is (see Table 8.7).

TABLE 8.7 Performance of 12 speakers

Respondent	Sex	Age of onset	Residency in years	GPI
HF	male	6	24	26.35
SA	male	8	22	23.28
NA	male	26	3	20.49
MA	male	15	12	17.35
MMA	male	30	4	16.46
MFA	male	23	2	15.47
MT	male	26	1.5	15.37
SNF	female	23	9	23.17
SB	female	20	5	22.27
HB	female	13	12	21.67
AZ	female	15	15	20.7
FB	female	19	3	19.34

Source: Rasinger 2007.

Task 8.5

Using a *t*-test assuming *unequal* (!) variance, analyse the difference between male and female performance. Is the answer as straightforward as it seems? What are potential pitfalls?

The full table for the *t*-test is in Chapter Eleven. If you have not made any mistakes – and I guess you have not – then you will have probably found that the mean performance of men is slightly lower than that of women, with $\bar{x}=19.25$ and $\bar{x}=21.43$, respectively. However, when we look at the significance values, we see that $p=0.25$ for the 2-tailed test. We have no predictions as to who scores better, men or women, so the 2-tailed test is the one to choose. That is, the difference is statistically not significant: there is a 25% chance that the difference in means is just a fluke. If you look closely at Table 8.7, you will see that the two male respondents HF and SA are quite different from the rest of the sample, as they have a much lower age of acquisition onset and have lived in England for much longer. From what we know about age of onset in second language acquisition, this should immediately ring alarm bells: if we have two data points which are that far off the rest of the sample, and also cross a theoretical 'magic line', then we should carefully reconsider our result.

Task 8.6

Remove HF and SA from the sample and re-calculate the *t*-statistic, this time assuming equal variance (as we have an equal number of men and women). Interpret the result with reference to your results from task 8.5.

If we remove HF and SA from the sample, the mean for male respondents decreases slightly to 17 points. More importantly, now $p=0.005$ (2-tailed), that is, the difference is statistically significant. It is almost impossible to say which of the two results is 'correct' – it all depends on your methodological and theoretical viewpoint and line of argument. Outlying values are always potentially problematic. But then, you may find ways to argue around them.

Furthermore, neither result tells us anything about causality: even though women score significantly better in the second test (without HF and SA), there is a myriad of variables we cannot account for and hence cannot control. In fact, when probing a bit further, it became obvious that it was in particular women who had school-aged children *and* had certain interaction patterns who spoke more target like than men without children. And to make it more complex, women with school-aged children do better than those without. So, in a nutshell, even the most sophisticated statistical method with even the fanciest of figures and scores should not ever be accepted unconditionally. Statistics on their own are usually not the answer – always use your critical judgement.