“The language-as-fixed-effect fallacy”: Some simple SPSS solutions to a complex problem

Marc Brysbaert

Royal Holloway, University of London

Report written for RTN-LAB [1]
Version 1.0, Wedneday, March 28, 2007

Address: Marc Brysbaert
Royal Holloway, University of London
Department of Psychology
Egham TW20 0EX
United Kingdom
marc.brysbaert@rhul.ac.uk

 If you found the report helpful in the analysis of your data, it would be kind to acknowledge so by citing it using the form:
Brysbaert, M. (2007). *“The language-as-fixed-effect fallacy”: Some simple SPSS solutions to a complex problem (Version 1.0).* Report Royal Holloway, University of London.
The report is available on the internet.

1. Introduction

In recent years psycholinguists have been criticized for using suboptimal statistical tests (Baayen, Davidson, & Bates, 2006; Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999). In particular, the use of F1 and F2 tests "to generalize over participants and items" has been called into question. At the same time, rumors are spreading about a much better type of analysis few people understand. In this paper I try to translate my (limited) knowledge in a form that is easy to master, because it consists of a series of cookbook recipes. It is the form used increasingly in stats courses and can be defended on the basis that there are different levels of understanding (e.g., knowing how to work with a statistical package and how to interpret the results vs. being able to build one). My discussion is limited to SPSS, not because I am particularly happy with this package, but because it is most widely used.

2. Why does one need to bother about variance between items?

For a beginning researcher it is tempting to limit the statistical analysis of psycholinguistic data to an analysis based on the average per condition per participant. For instance, if 10 participants make a lexical decision to 5 low frequency words and 5 high frequency words, we will calculate the mean of the reaction times (RT) to the correctly identified low frequency words and the mean of the RTs to the correctly identified high frequency words (in addition to the percentage of errors, which will be used as a second variable). Table 1 shows some results we may obtain (empty cells are errors made by the participants).

Example data LDT - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : participant 1

	participant	Low1	Low2	Low3	Low4	Low5	High1	High2	High3	High4	High5	var
1	1	655	847	.	687	603	652	.	706	633	593	
2	2	724	954	653	624	613	649	642	505	659	725	
3	3	589	763	.	688	589	639	.	638	596	631	
4	4	647	712	769	594	.	714	566	684	652	545	
5	5	842	.	698	711	657	598	639	652	681	684	
6	6	.	863	647	659	688	655	685	701	706	576	
7	7	711	712	589	624	637	689	625	.	599	703	
8	8	652	914	723	599	725	675	750	692	618	.	
9	9	483	752	642	602	568	497	504	615	587	605	
10	10	756	811	699	705	718	637	649	587	675	636	
11												

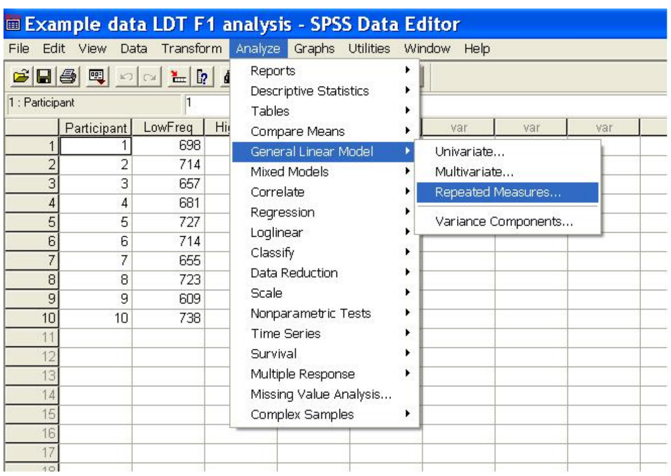
Table 1 : Example data of a lexical decision experiment containing of 5 low frequency and 5 high frequency words. Ten participants in total.

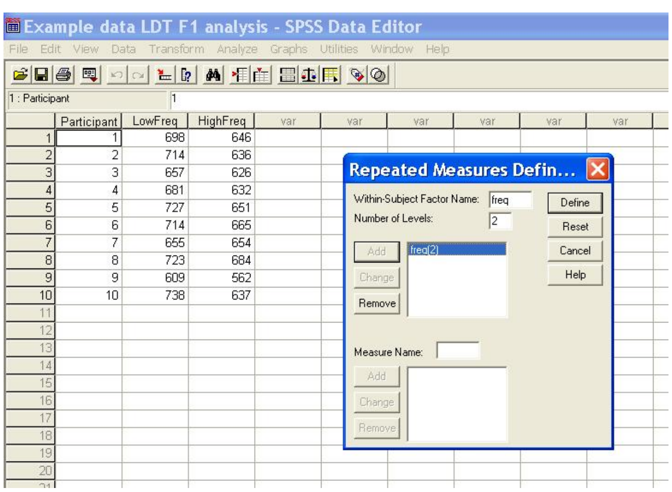
When we calculate the mean RTs of the correct trials for the low and the high frequency words, we get Table 2.

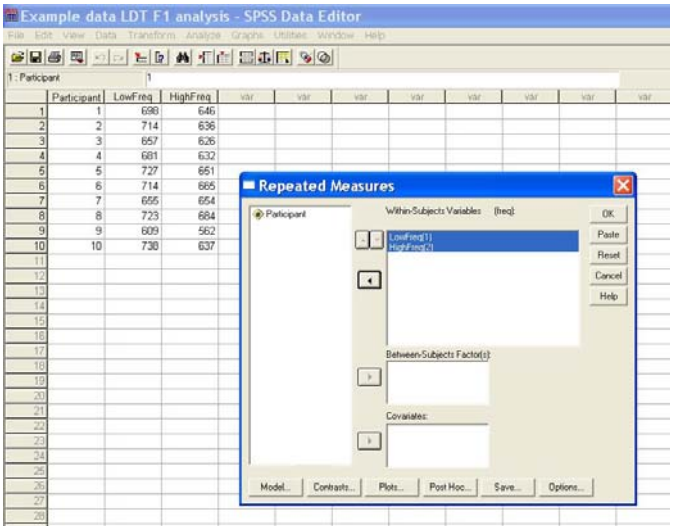
	Participant	LowFreq	HighFreq
1	1	698	646
2	2	714	636
3	3	657	626
4	4	681	632
5	5	727	651
6	6	714	665
7	7	655	654
8	8	723	684
9	9	609	562
10	10	738	637

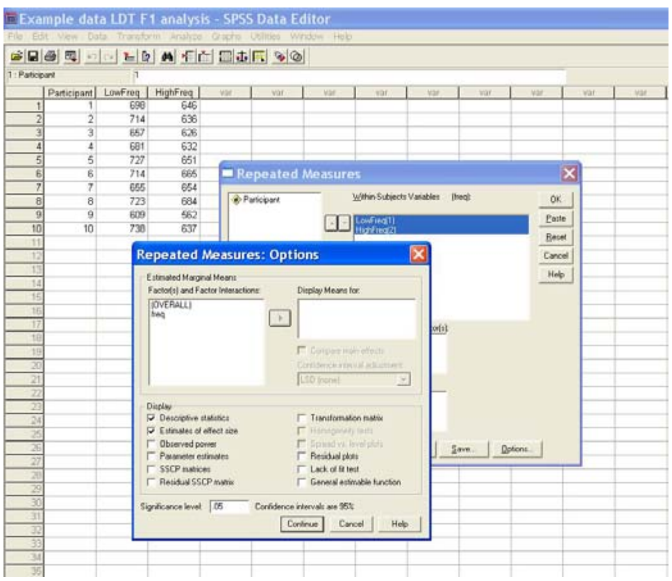
Table 2 : Mean RT of the low frequency and the high frequency words per participant (correct trials only).

To run the analysis, we have to use an ANOVA with a repeated measure. The figures below show how we get there.









Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
freq	Sphericity Assumed	13676.450	1	13676.450	35.646	.000	.798
	Greenhouse-Geisser	13676.450	1.000	13676.450	35.646	.000	.798
	Huynh-Feldt	13676.450	1.000	13676.450	35.646	.000	.798
	Lower-bound	13676.450	1.000	13676.450	35.646	.000	.798
Error(freq)	Sphericity Assumed	3453.050	9	383.672			
	Greenhouse-Geisser	3453.050	9.000	383.672			
	Huynh-Feldt	3453.050	9.000	383.672			
	Lower-bound	3453.050	9.000	383.672			

So, on the basis of our ANOVA with one repeated measure, we get a significant effect: $F(1,9) = 35.646$, $MSe = 383.672$, $p < .001$, Eta Squared = .798 [2]. The effect is extraordinarily strong because no participant has a lower mean RT for the low frequency words than for the high frequency words. This is strong evidence that high frequency words are easier to process than low frequency words, isn't it?

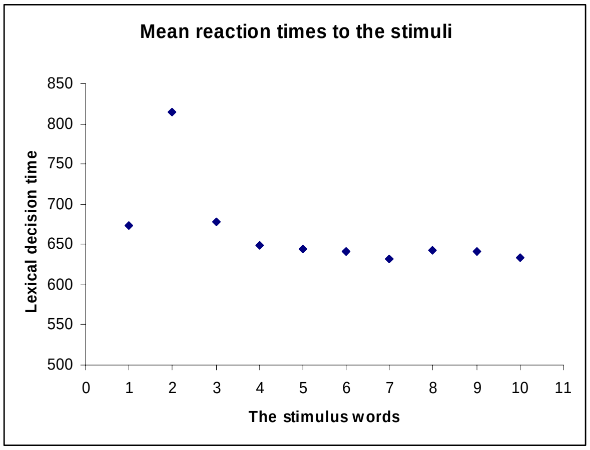


Figure 1 : Mean lexical decision time per word: the first five words are the low frequency words; the final five are the high frequency words.

Figure 1 shows another part of the story, however. This figure displays the mean RT per word stimulus. Now, the evidence suddenly looks less impressive: Nearly all the difference between the high and the low frequency words is due to the long RTs for word Low2 (see also Table 1). If we took another sample of words that does not include word Low2, would we still find a frequency effect?

The discrepancy between Table 2 and Figure 1 is what Clark (1973) called "the language-as-fixed-effect fallacy". If we limit our statistical analysis to the analysis reported above, we assume that there is no variability in the words we have chosen, or that our sample exhausts all possible words we could have selected. Given that this rarely is the case, Clarke argued that in our statistical analyses we have to take into account the variability due to the items in addition to the variability due to the stimulus items. Although his analysis is not that difficult to understand, it requires the reader to know something about the difference between fixed and random effects in ANOVAs and about how to calculate Mean Square terms and F-values. In addition, the analysis Clarke proposed (a quasi-F ratio or F') only works when there are no missing data (i.e., when the participants make no errors or when the missing RTs are estimated).

[2] Eta squared is an index of the effect size. You get it when you click on **Options** and **Estimates of effect size**. The eta squared has a similar meaning as R^2 (how much of the variance is due to the effect). In psychology, most values of eta squared will be around .09 (i.e. $r = .30$, medium effect size). The high value in the present example gives away that it was constructed by hand.

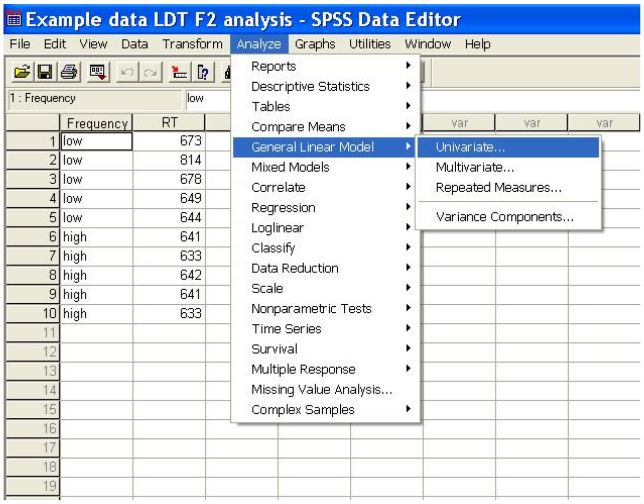
Luckily, Clarke (1973) also included an easier way around the problem (although Raaijmakers claims this has been one of the big mistakes in psycholinguistic research, because psycholinguists used the shortcut in the wrong way).

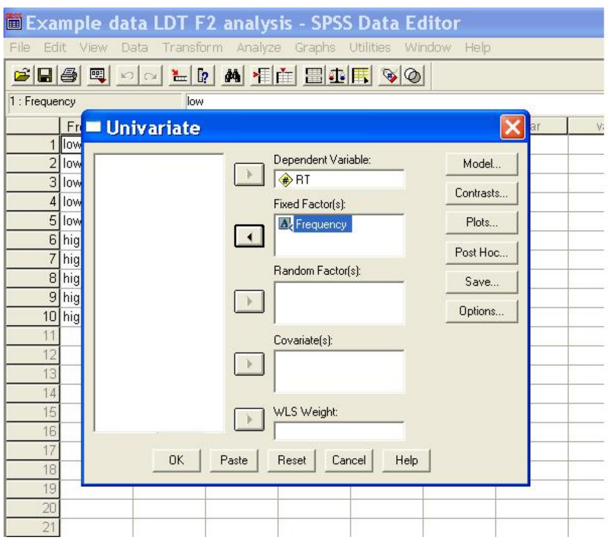
The solution Clarke proposed, was to do an F2 analysis in addition to the F1 analysis and to calculate minF'. F1 is the analysis we have discussed above (Table 2). It consists of an ANOVA on the mean values per participant per condition. There can be as many independent variables (IVs) as one likes (although in reality, it is strongly recommended not to have more than 2; higher-order interactions are a nightmare to interpret and usually are unstable; i.e., the exact same pattern is not obtained in a replication of the study, even when the interaction is significant again; in addition, very few researchers have a priori hypotheses about more than two IVs).

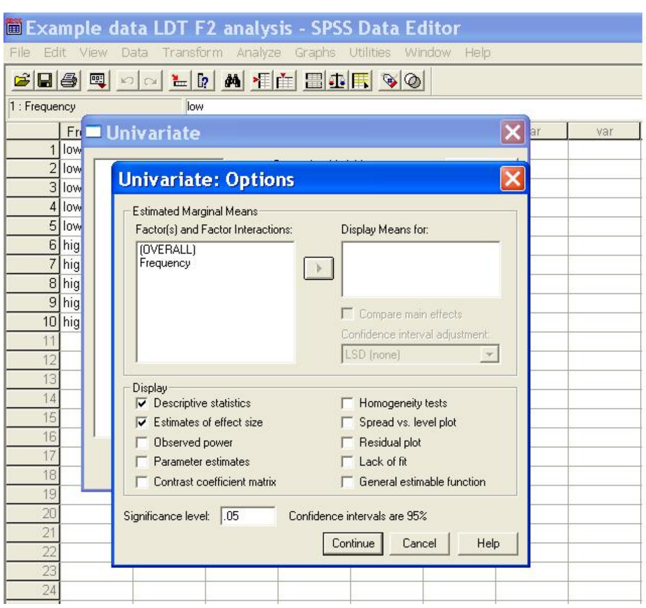
	Frequency	RT
1	low	673
2	low	814
3	low	678
4	low	649
5	low	644
6	high	641
7	high	633
8	high	642
9	high	641
10	high	633

Table 3: Mean RT of the 5 low frequency words and the 5 high frequency words.

Table 3 shows the starting point of the F2 analysis, the analysis over items. For this analysis, the researcher calculates the mean RT per word. Because in the present example the words belonging to the high frequency condition and the words belonging to the low frequency condition are different words, the IV will be a between-items variable. These are the steps of the analysis:


Example data LDT F2 analysis - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Frequency
low
Frequency
RT
var
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Univariate...
Multivariate...
Repeated Measures...
Variance Components...


Example data LDT F2 analysis - SPSS Data Editor
Univariate
Dependent Variable:
RT
Fixed Factor(s):
Frequency
Random Factor(s):
Covariate(s):
WLS Weight:
Model...
Contrasts...
Plots...
Post Hoc...
Save...
Options...
OK
Paste
Reset
Cancel
Help


Example data LDT F2 analysis - SPSS Data Editor
Univariate
Univariate: Options
Estimated Marginal Means
Factor(s) and Factor Interactions:
(OVERALL)
Frequency
Display Means for:
Compare main effects
Confidence interval adjustment:
LSD (none)
Display
Descriptive statistics
Estimates of effect size
Observed power
Parameter estimates
Contrast coefficient matrix
Homogeneity tests
Spread vs. level plot
Residual plot
Lack of fit
General estimable function
Significance level:
.05
Confidence intervals are 95%
Continue
Cancel
Help

Tests of Between-Subjects Effects

Dependent Variable: RT

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7182.400(a)	1	7182.400	2.920	.126	.267
Intercept	4419590.400	1	4419590.400	1796.837	.000	.996
Frequency	7182.400	1	7182.400	2.920	.126	.267
Error	19677.200	8	2459.650			
Total	4446450.000	10				
Corrected Total	26859.600	9				

a R Squared = .267 (Adjusted R Squared = .176)

In the F2 analysis we see that the effect of word frequency is not significant (F2(1,8) = 2.92, MSe = 2460, p = .126, eta squared = .267). In the psycholinguistic community, this "means" that on the basis of the present data we cannot assume that the finding generalizes to other stimuli (notice that it is a null-effect, so the researcher is not allowed to conclude that the effect is 'absent'; the power of the experiment is way too low for that).

Clarke (1973) himself did not pay too much attention to the particular value of F2 (rightly so) and only calculated it because it allowed him to obtain a reasonably good estimate of an F value that would ***generalize at the same time across participants and items***, which he called the **minF'**. The minF' value is calculated as follows:

$$\min F'(i,j) = \frac{F1 * F2}{F1 + F2}$$

$i = df_1\_of\_F1 = df_1\_of\_F2$ (df_1 of F1 has to be the same as df_1 of F2)

$$j = \frac{(F1+F2)^2}{\left(\frac{F1^2}{df_2\_of\_F2} + \frac{F2^2}{df_2\_of\_F1}\right)}$$

Applied to our example, this gives

$$i = 1$$

$$j = \frac{(35.646 + 2.920)^2}{\left(\frac{35.646^2}{8} + \frac{2.920^2}{9}\right)} = 9.3 \approx 9$$

$$\min F'(1,9) = \frac{35.646 * 2.920}{35.646 + 2.920} = 2.699$$

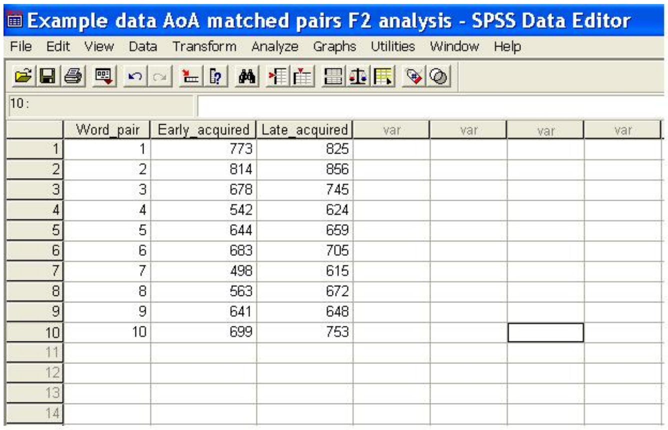
To find the p-value associated with minF', you can use the built-in Excel function [FDIST(2.699,1,9) = .135] or use a ready-made applet on the internet (see http://www.pallier.org//ressources/MinF/compminf.htm or http://users.ugent.be/~rhartsui/tools.html).

The minF' test informs us that we are not allowed on the basis of the data in Table 1 to argue for a reliable frequency effect that generalizes both across participants and stimuli, which was Clarke's message.

3. Getting overly excited about F2

In the years after Clark (1973) the importance of an F2 items analysis became generally accepted in psycholinguistics, but gradually psycholinguists forgot about minF' (Raaijmakers et al., 1999). Part of the reason for this was that psycholinguists were not aware of the fact that a significant F1 and a significant F2 do not suffice to get a significant minF' (what Raaijmakers et al. called the "F1 x F2 fallacy"). In addition, there were good reasons to expect that minF' would be a conservative test (i.e., more difficult to get significance with it), although subsequent simulations showed that this problem was less severe than feared at the onset.

Anyway, gradually psycholinguists moved away from minF' and limited their analyses to F1 (to check whether the findings could be generalized across participants) and F2 (to check whether the findings could be generalized across stimulus materials). In addition, psycholinguists became more and more 'sophisticated' in their use of F1 and F2. Looking at Table 3, we see that the F2 analysis in our example is one of the least powerful tests one could imagine: Because word frequency is a between-items variable, all noise due to the individual words is added to the error term and one needs large numbers of observations in the different conditions to find a significant F2. This is in particular a problem when pairs of words have been assembled that differ on one particular variable (e.g., age of acquisition, AoA) and are matched on a list of other variables (frequency, word length, number of orthographic neighbors, …). Because of the control variables, large differences between the word pairs are expected (otherwise one would not need to match the stimuli on these variables) and this variance should be partialed out before we start the F2 analysis. One solution is to use a repeated measures design for the F2 analysis as well. In this analysis the pairs of stimuli are considered as observations from the same 'entity' or 'block' (analog to the 'participant' in a repeated measures F1 analysis). Table 4 shows how the data of such an F2 design would look like for an experiment in which 10 pairs of words have been selected that differ in AoA (one word is acquired early in life, e.g., daffodil, the other word is acquired late in life, e.g., participant) and both are matched on a series of other measures (frequency, …).



	Word_pair	Early_acquired	Late_acquired	var	var	var	var
1	1	773	825				
2	2	814	856				
3	3	678	745				
4	4	542	624				
5	5	644	659				
6	6	683	705				
7	7	498	615				
8	8	563	672				
9	9	641	648				
10	10	699	753				
11							
12							
13							
14							

For these data, the F2 with repeated measures is F2(1,9) = 22.647, MSe = 710, p = .001, eta squared = .716), whereas with a between-items analysis it would be F2(1,18) = 1.922, MSe = 8366, p = .183, eta squared = .096). The reason for the lack of power of the between-items analysis becomes clear when you compare the mean squares of error of both tests (8366 vs. 710). In the repeated-items analysis, a lot of the variability between the stimuli is partialed out as variability between the blocks, due to variation in the control variables, whereas this variance is included in the error term of the between-items test, making it very hard to find a significant F2 (just like one needs at least 128 participants to look for a medium size effect in a between-participants F1 analysis with 1 IV and 2 conditions).

Another way to 'improve' the F2 analysis is to include a Latin-square variable (Pollatsek & Well, 1995). A technique psycholinguists often use, is to counterbalance their stimuli over participants. Imagine, for instance, that you want to investigate semantic priming. To do so, you search for target words with related and unrelated primes (e.g., using the Edinburgh Thesaurus, http://www.eat.rl.ac.uk/ or Nelson's Florida norms, http://w3.usf.edu/FreeAssociation/). These are some of the words you may come up with:

Target	Related prime	Unrelated prime
bread	butter	buffer
boy	girl	curl
nurse	doctor	danger
cat	dog	day
…		

Because you do not want to present your target words twice to the same participant, half of the participants see bread preceded by butter and the other half sees bread preceded by buffer, and so on. So, you will make two stimulus list:

List 1	List 2
butter-bread	buffer-bread
curl-boy	girl-boy
doctor-nurse	danger-nurse
day-cat	dog-cat
…	

Half of the participants will get list 1 and half list 2. Now, a typical problem in such a design is what to do with a slow or a fast participant. Table 4 illustrates what can happen:

data latin square design semantic priming - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

10 : cat_unrelated

	participant	bread_related	bread_unrelated	boy_related	boy_unrelated	nurse_related	nurse_unrelated	cat_related	cat_unrelated	var
1	1	624	.	.	694	588	.	.	658	
2	2	.	648	684	.	.	675	602	.	
3	3	1024	.	.	1124	968	.	.	1214	
4	4	.	655	625	.	.	645	668	.	
5	5	745	.	.	684	593	.	.	695	
6	6	.	698	652	.	.	689	656	.	
7	7	635	.	.	674	654	.	.	653	
8	8	.	704	705	.	.	695	687	.	
9	9	674	.	.	639	655	.	.	708	
10	10	.	657	658	.	.	597	604	.	
11										
12										
13										

The important person here is participant 3, who is considerably slower than everyone else. Because of the Latin-square design, this person will add extra RT to the related condition for the stimuli *butter-bread* and *doctor-nurse*; similarly s/he will add extra RT to the unrelated condition for the stimuli *curl-boy* and *day-cat*. This will show in the data that are entered in the F2 analysis, as can be seen below:

Untitled - SPSS Data Editor

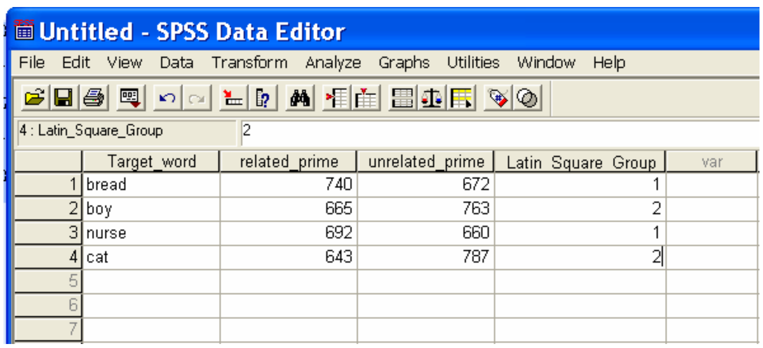
File Edit View Data Transform Analyze Graphs Utilities

14 :

	Target_word	related_prime	unrelated_prime	var
1	bread	740	672	
2	boy	665	763	
3	nurse	692	660	
4	cat	643	787	
5				
6				
7				
8				

For the target stimuli *boy* and *cat*, we find a huge effect in the expected direction, whereas for the stimuli *bread* and *nurse*, we find a small effect in the opposite direction, even though nearly all the individual participants showed the predicted semantic priming effect. If we do the calculations, we find $F2(1,3) = .489$, $MSe = 5158$, $p = .535$, eta squared = .140. Needless to say, such a low F2 value will also result in a low minF'.

One way to increase the power of this design is to add a Latin-square variable to the design. The words *bread* and *nurse* were seen in the related condition by one group of 5 participants, and in the unrelated condition by another group of 5 participants. And vice versa for the words *boy* and *cat*. Therefore, what we can do to get rid of the difference in average RTs between the groups, is to add the following between-items variable:


Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

4 : Latin_Square_Group 2

	Target_word	related_prime	unrelated_prime	Latin_Square_Group	var
1	bread	740	672	1	
2	boy	665	763	2	
3	nurse	692	660	1	
4	cat	643	787	2	
5					
6					
7					

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
factor1	Sphericity Assumed	2520.500	1	2520.500	5.910	.136	.747
	Greenhouse-Geisser	2520.500	1.000	2520.500	5.910	.136	.747
	Huynh-Feldt	2520.500	1.000	2520.500	5.910	.136	.747
	Lower-bound	2520.500	1.000	2520.500	5.910	.136	.747
factor1 * LS_group	Sphericity Assumed	14620.500	1	14620.500	34.280	.028	.945
	Greenhouse-Geisser	14620.500	1.000	14620.500	34.280	.028	.945
	Huynh-Feldt	14620.500	1.000	14620.500	34.280	.028	.945
	Lower-bound	14620.500	1.000	14620.500	34.280	.028	.945
Error(factor1)	Sphericity Assumed	853.000	2	426.500			
	Greenhouse-Geisser	853.000	2.000	426.500			
	Huynh-Feldt	853.000	2.000	426.500			
	Lower-bound	853.000	2.000	426.500			

Although the small number of stimuli in our example does not allow us to reach significance, the F2 test now looks much more convincing ($F2(1,2) = 5.910$, $MSe = 426.5$, $p = .136$). A look at the ANOVA table shows that a lot of the noise in the F2 analysis caused by the slow participant 3 has been captured by the interaction effect between semantic priming and Latin-square group. In the same way, unintended variation between the stimuli that make up list 1 and list 2 can be partialed out by including a Latin-Square variable in the F1 analysis. Another way to get rid of unintended variation due to slow participants, is to use the z-scores per participant (i.e., the (RTs – $M_{participant}$)/$sd_{participant}$, a technique used by Balota and Besner).

4. Being put down again

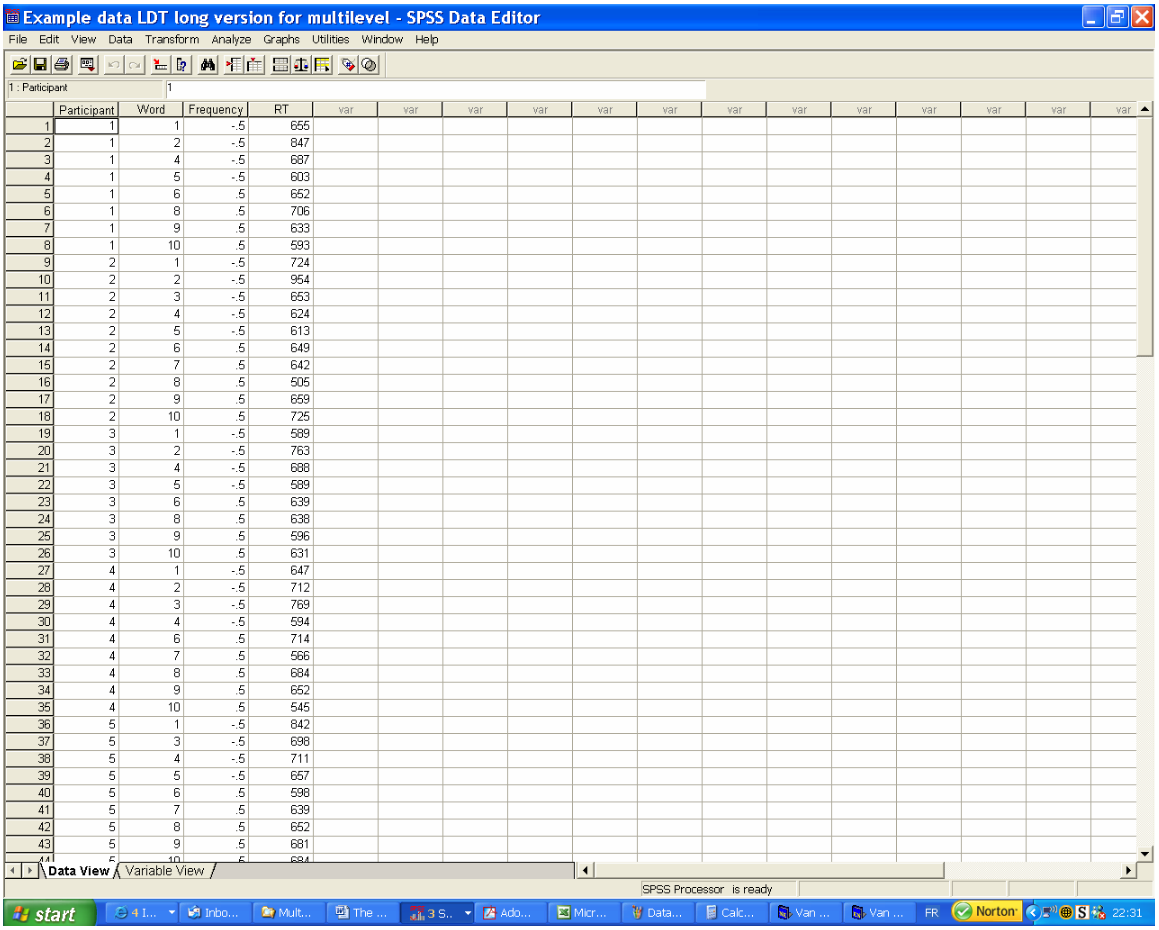
Just when psycholinguists thought they were getting savvy enough to run proper analyses, they were attacked anew. First, there was Raaijmakers' comment that a significant F1 and a significant F2 were not enough to generalize across participants and stimuli. This prompted JML to require all its authors to report minF' in addition to F1 and F2. As a kind of consolation, Raaijmakers et al. (1999) added that an F2 analysis is not always required and in some cases can even lead to a needless loss of power. Ironically, by doing so Raaijmakers et al. repeated Clark's (1973) mistake, because ever since I've seen more references to Raaijmakers et al. by authors claiming that their non-significant F2 analysis is of no real concern than by authors arguing why they believe minF' is more important than separate F1 and F2 analyses.

At the same time, Baayen started to launch the claim that the minF' analysis as a combination of F1 and F2 is needlessly complicated and should be replaced by mixed-effects (or multi-level) modeling (Baayen, 2007; Baayen et al., 2006). Unfortunately, Baayen's language is so specialized that it took me a few months and the help of others to realize what he was talking about. In particular, I've been able to make headway by comparing Baayen et al. (2006) with Locker, Hoffman, and Bovaird (2007) and by trying to understand what Van den Noortgate and Onghena (2006) were doing. Below you find my current understanding of these techniques. It may be wrong in a number of details (in which case I would appreciate your feedback), but at least it looks pretty convincing to me (at the moment). Here we go.

5. Jumping a few levels higher

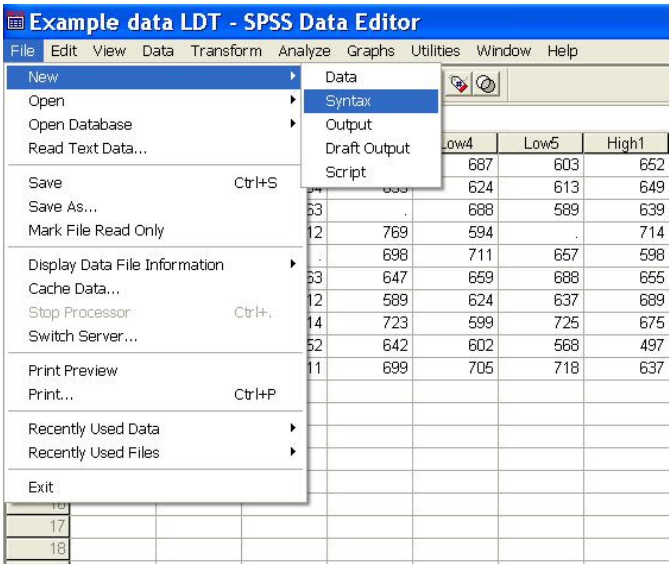
Just like an ANOVA at its basis is nothing else than a multiple regression, so you can approach the problem of random participants and random stimuli as a regression problem. You try to predict an observed RT as the end result of (i) a participant, (ii) a stimulus, and (iii) the contribution of one (or more) IVs. So, what you try to do is to see whether your manipulation is explaining anything more than what could be predicted on the basis of the participants and the stimuli. The only real thing you need is an algorithm that goes beyond simple linear regression. Turns out that SPSS has such an algorithm! (At least from version 11 on). It is called MIXED. I will go through the procedure on the basis of Table 1 (LDT to high and low frequency words).

The first thing to do is rewrite everything as you would for a multiple regression analysis. So, you have three predictor variables: participant, stimulus word, and frequency condition (the latter is recoded as -.5 for a low frequency word +.5 for a high frequency item; by using this code, you can easily interpret the regression weight). So, this gives the following input file:

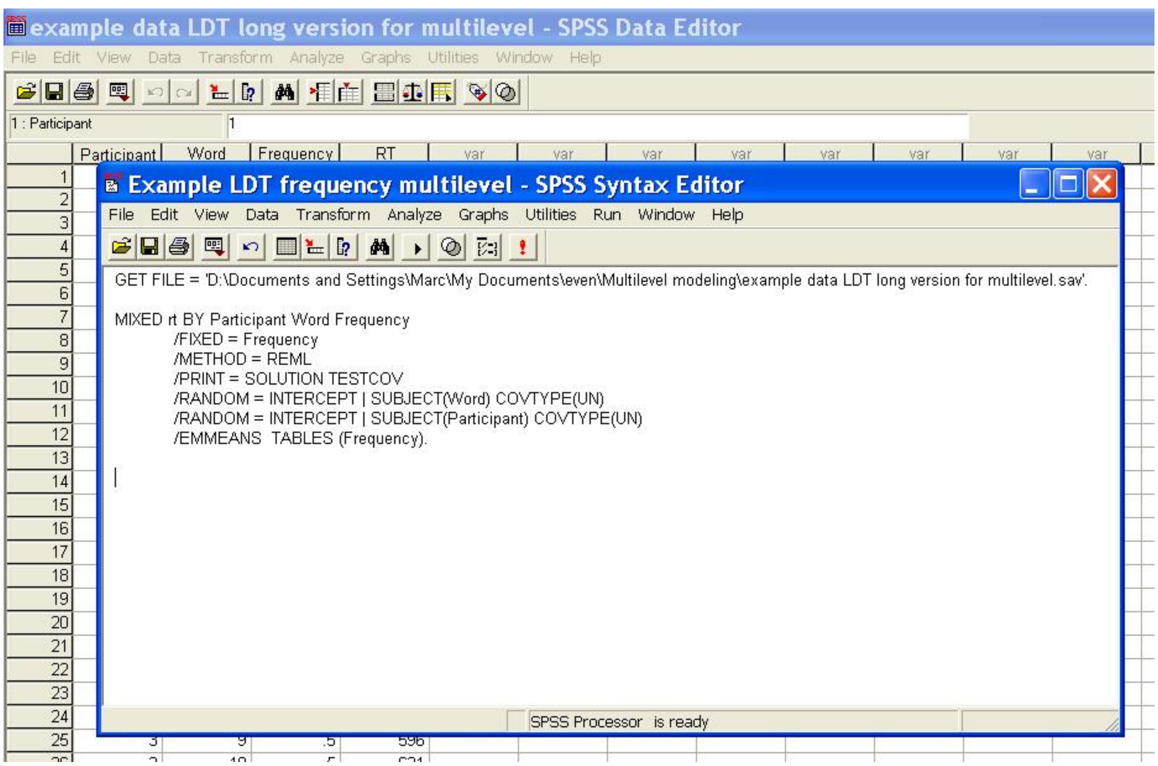


	Participant	Word	Frequency	RT
1	1	1	-.5	655
2	1	2	-.5	847
3	1	4	-.5	687
4	1	5	-.5	603
5	1	6	.5	652
6	1	8	.5	706
7	1	9	.5	633
8	1	10	.5	593
9	2	1	-.5	724
10	2	2	-.5	954
11	2	3	-.5	653
12	2	4	-.5	624
13	2	5	-.5	613
14	2	6	.5	649
15	2	7	.5	642
16	2	8	.5	505
17	2	9	.5	659
18	2	10	.5	725
19	3	1	-.5	589
20	3	2	-.5	763
21	3	4	-.5	688
22	3	5	-.5	589
23	3	6	.5	639
24	3	8	.5	638
25	3	9	.5	596
26	3	10	.5	631
27	4	1	-.5	647
28	4	2	-.5	712
29	4	3	-.5	769
30	4	4	-.5	594
31	4	6	.5	714
32	4	7	.5	566
33	4	8	.5	684
34	4	9	.5	652
35	4	10	.5	545
36	5	1	-.5	842
37	5	3	-.5	698
38	5	4	-.5	711
39	5	5	-.5	657
40	5	6	.5	598
41	5	7	.5	639
42	5	8	.5	652
43	5	9	.5	681
44	5	10	.5	684

The nice thing about this input is that it makes no great deal if there are a few missing observations. You just skip the line (e.g., word 3 for participant 1). The regression method is reasonably robust against empty cells (at least that's what I've read). Then we have to enter our model. Here it is a bit tricky because you must enter the syntax editor. You do this as follows:



This opens a syntax file. Another, more easy way to open a syntax file is to open a ready made file (or to click on it in windows explorer). Then everything opens automatically in SPSS. In the syntax file you write the following:



First, you have to indicate where the computer can find your data file. Then, you indicate what the dependent variable is of your MIXED program (RT) and which predictor variables (Participant, Word, Frequency). Participant and Word are random variables (i.e., a random sample from the population). Frequency is a fixed effect (you are interested in these two levels). Basically this is all you have to do. You indicate that each participant and each stimulus word can have a different intercept value (i.e. need more or less time to process) and in addition you want to see whether frequency adds enough weight to be significant. The /EMMEANS command gives you the maximum likelihood estimator of the condition means. Once you've entered everything (do not forget the full stops!) you click on RUN. If everything goes well, this is what you should get (among other garbage):

Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	11.747	1284.077	.000
Frequency	1	7.988	2.860	.129

a Dependent Variable: RT.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	637.8481631	24.4844447	10.318	26.051	.000	583.5204508	692.1758754
[Frequency=-.5]	54.1145501	31.9990880	7.988	1.691	.129	-19.6941046	127.9232048
[Frequency=.5]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: RT.

Frequency(a)

Frequency	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
-.5	691.963	24.517	10.374	637.602	746.324
.5	637.848	24.484	10.318	583.520	692.176

a Dependent Variable: RT.

The F-value is given in the first table. The second table contains the t-values of the planned comparisons. The F-value is:

$F(1,7.988) = 2.860$, $p = .129$

For the sake of comparison, this was the minF' value we obtained:

$minF'(1,9) = 2.699$, $p = .135$.

Not bad if you look at the ease with which you can do this analysis!!! Baayen et al. (2006) have done quite some simulations with this technique (albeit on an R version of theirs, which gives the same results) and they claim that it is safe (i.e., does not result in spurious significant effects and is not too conservative). In addition, once you know the technique, it is very versatile. Below, I give a few more examples.

6. Getting carried away (again)

One way to check the adequacy of a procedure is to apply it to the classic data sets that have been used in the literature on F2 effects. Most of them come from Raaijmakers et al. (1999).

For instance, Raaijmakers et al. (1999) give the following example (also analyzed by Baayen et al., 2006). It concerns a hypothetical study in which 4 participants take part in a priming study and see 4 items with a short SOA and 4 (different) items with a long SOA.

TABLE 2

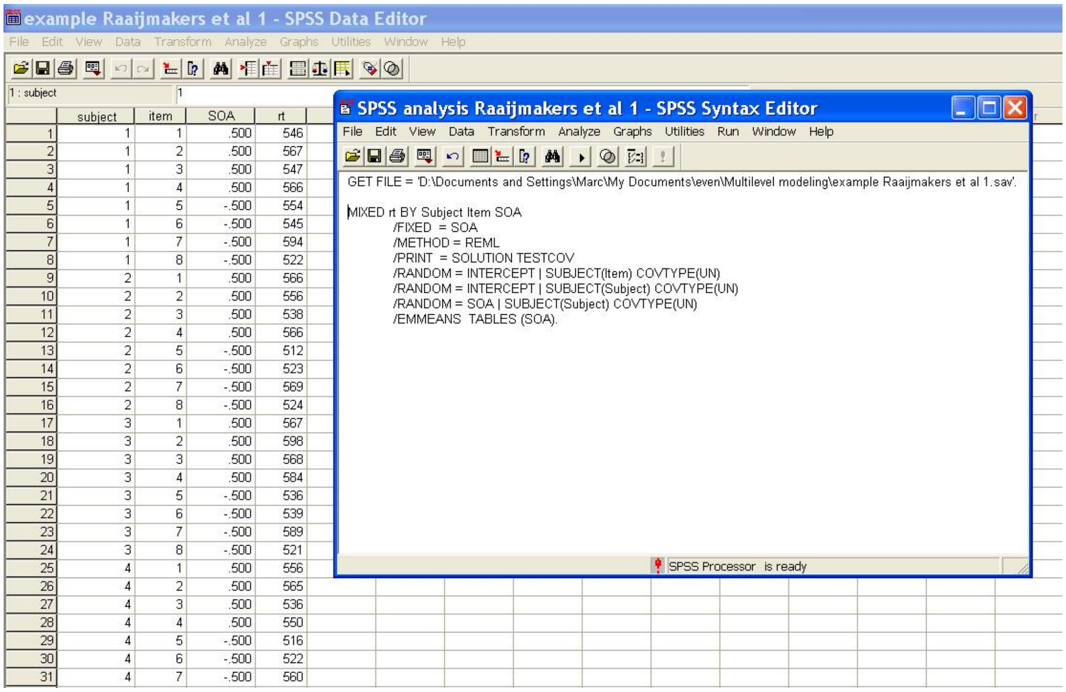
Simulated Data for Repeated-Measurements ANOVA with Words Sampled Randomly

	Short SOA				Long SOA			
Subject	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	546	567	547	566	554	545	594	522
2	566	556	538	566	512	523	569	524
3	567	598	568	584	536	539	589	521
4	556	565	536	550	516	522	560	486
5	595	609	585	588	578	540	615	546
6	569	578	560	583	501	535	568	514
7	527	554	535	527	480	467	540	473
8	551	575	558	556	588	563	631	558

For this table Raaijmakers et al. report:

F1(1,7) = 7.41, p = .0297
F2(1,6) = 2.17, p = .1912
minF'(1,10) = 1.68, p = .224

So, how does the multilevel analysis cope? To find out, we again have to write the table in a long form and then run the analysis.



These are the results:

Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	12.590	2655.786	.000
SOA	1	8.250	1.717	.225

a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	563.3125000	12.3815143	9.543	45.496	.000	535.5445293	591.0804707
[SOA=-.500]	-22.4062500	17.1014526	8.250	-1.310	.225	-61.6356214	16.8231214
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

These findings [F(1,8.25) = 1.72, p = .225] agree pretty well with those of minF'.

The second example Raaijmakers et al. (1999) gave was a priming study in which the SOA between prime and target was manipulated and in which the items were matched in 4 pairs (called blocks). This is how the data looked like:

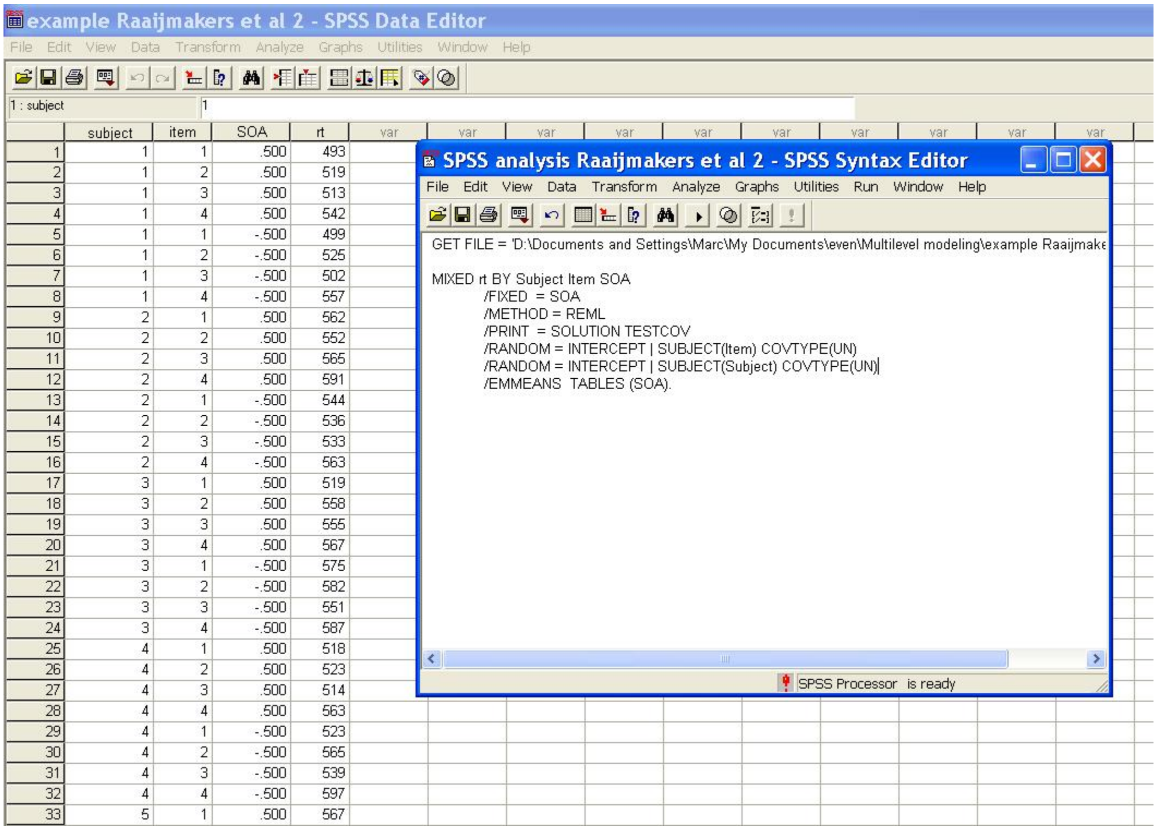
TABLE 4

Simulated Data for Repeated-Measurements ANOVA with Matched Items

	Short SOA				Long SOA			
Subject	Block 1	Block 2	Block 3	Block 4	Block 1	Block 2	Block 3	Block 4
1	493	519	513	542	499	525	502	557
2	562	552	565	591	544	536	533	563
3	519	558	555	567	575	582	551	587
4	518	523	514	563	523	565	539	597
5	567	562	577	595	521	563	559	575
6	520	534	527	568	512	541	531	559
7	516	544	513	575	555	569	550	601
8	525	528	528	559	551	542	529	578

$F1(1,7) = 0.86$, $p = .385$
$F2(1,3) = 7.19$, $p = .075$ (by making use of a repeated measures design; see the semantic priming experiment above)
$minF'(1,3) = 0.77$, $p = .445$

The picture below shows how to do the multilevel analysis:



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	4.721	2637.865	.000
SOA	1	52.000	3.411	.070

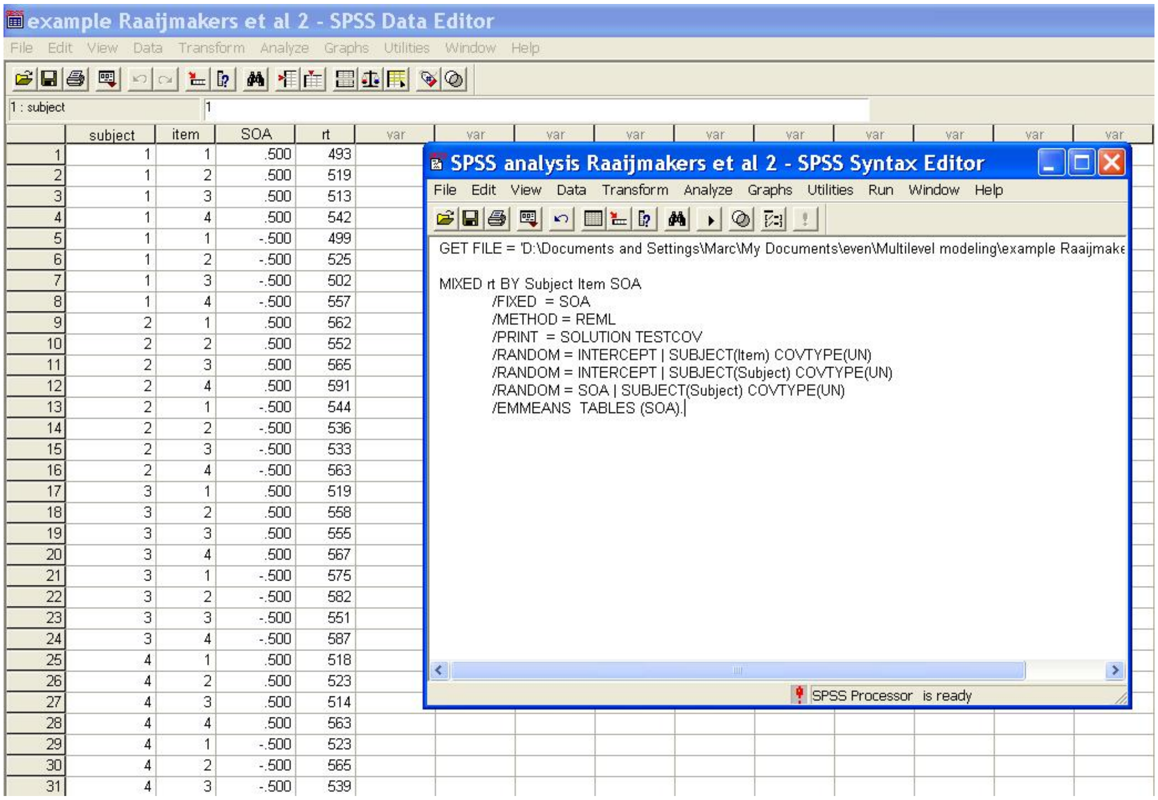
a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	543.50000	10.814031	5.019	50.259	.000	515.7339981	571.2660019
[SOA=-.500]	6.9375000	3.7564648	52.000	1.847	.070	-.6003980	14.4753980
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

Here we see something 'strange': The multilevel analysis is much more 'lenient' than minF' (F(1,52) = 3.41, p = .07). What is happening here? To be honest, I don't know. The only thing I know is that when Baayen et al. (2006) discussed this example, they included an additional random variable, next to participants and items, namely SOA (which is random by participant; the authors do not explain why). If we do so, we get the following:



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	4.857	2600.649	.000
SOA	1	6.714	.862	.385

a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	543.50000	11.600214	6.163	46.853	.000	515.2966380	571.7033620
[SOA=-.500]	6.9375000	7.4729796	6.714	.928	.385	-10.8872079	24.7622079
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

Now, the F-value looks much more like what one would expect: F(1,6.7) = .862, p = .385. Apparently in a blocked design you need to define the block as a random variable (I hope to clear this out in a later version).

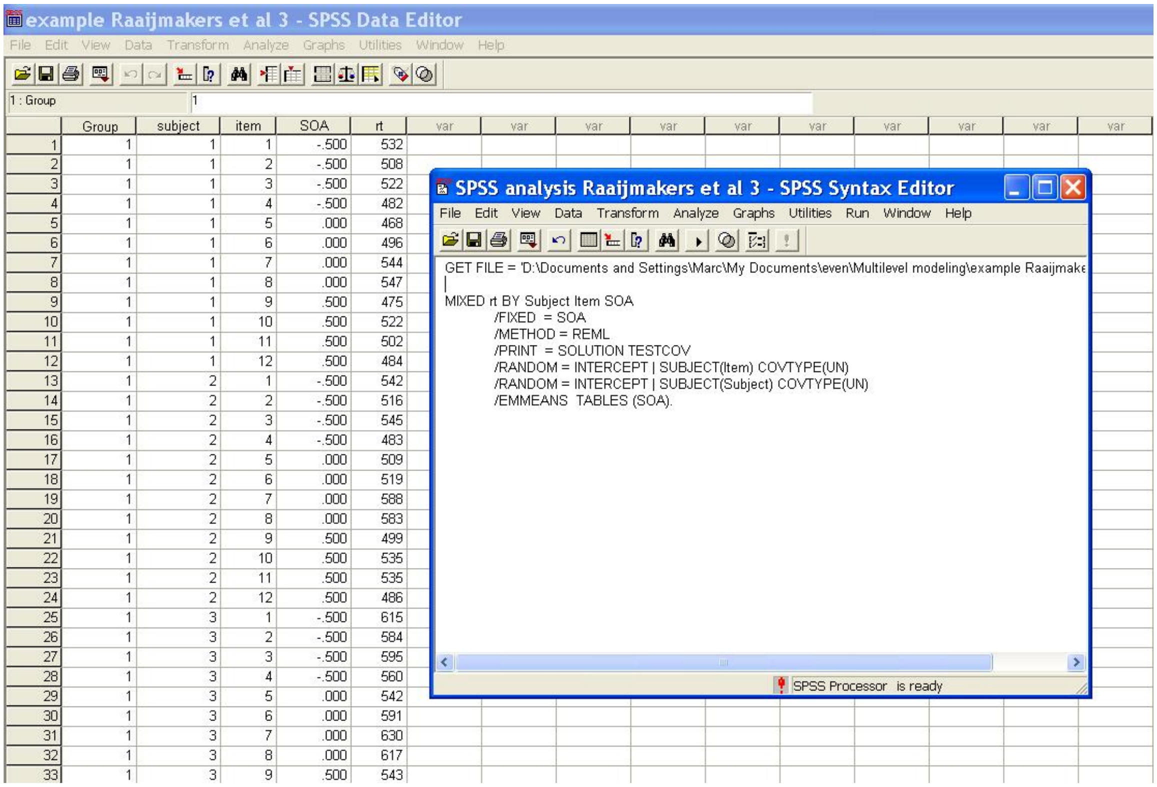
The final example Raaijmakers et al. (1999) gave was an example in which a Latin-square design is used. It was a priming study with 3 SOA levels (short, medium, and long) and 12 items that were rotated over the three conditions.

TABLE 7

Simulated Data for Design Using Counterbalanced Lists

		Short SOA				Medium SOA				Long SOA			
Group	Subject	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
1	1	532	508	522	482	468	496	544	547	475	522	502	484
	2	542	516	545	483	509	519	588	583	499	535	535	486
	3	615	584	595	560	542	591	630	617	543	606	560	545
	4	547	553	584	535	514	555	591	606	538	565	546	527
		Item 9	Item 10	Item 11	Item 12	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
2	5	553	598	581	551	619	576	606	561	548	590	614	631
	6	464	502	485	451	484	479	499	471	447	486	514	523
	7	481	511	492	472	531	506	542	475	471	510	539	556
	8	541	588	551	533	582	556	589	515	538	545	601	576
		Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 1	Item 2	Item 3	Item 4
3	9	482	530	571	563	501	561	500	506	543	539	558	497
	10	559	570	632	639	551	592	572	561	617	587	616	549
	11	462	497	546	538	487	546	491	470	529	508	525	473
	12	460	463	511	528	457	506	487	453	498	479	512	443

Raaijmakers et al. calculated a reasonably complicated F-statistic for this design, which yielded $F(2,20) = .896$, $p = .424$ (see also below for the 'usual' F1, F2, and minF'). The multilevel analysis gave the following results.



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	19.798	1530.967	.000
SOA	2	119.000	.944	.392

a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	533.95833	13.709801	20.084	38.947	.000	505.3678104	562.5488563
[SOA=-.500]	-.4583333	2.0058829	119.000	-.228	.820	-4.4301819	3.5135152
[SOA=.000]	2.1250000	2.0058829	119.000	1.059	.292	-1.8468486	6.0968486
[SOA=.500]	0(a)	0	.	.	.	.	.

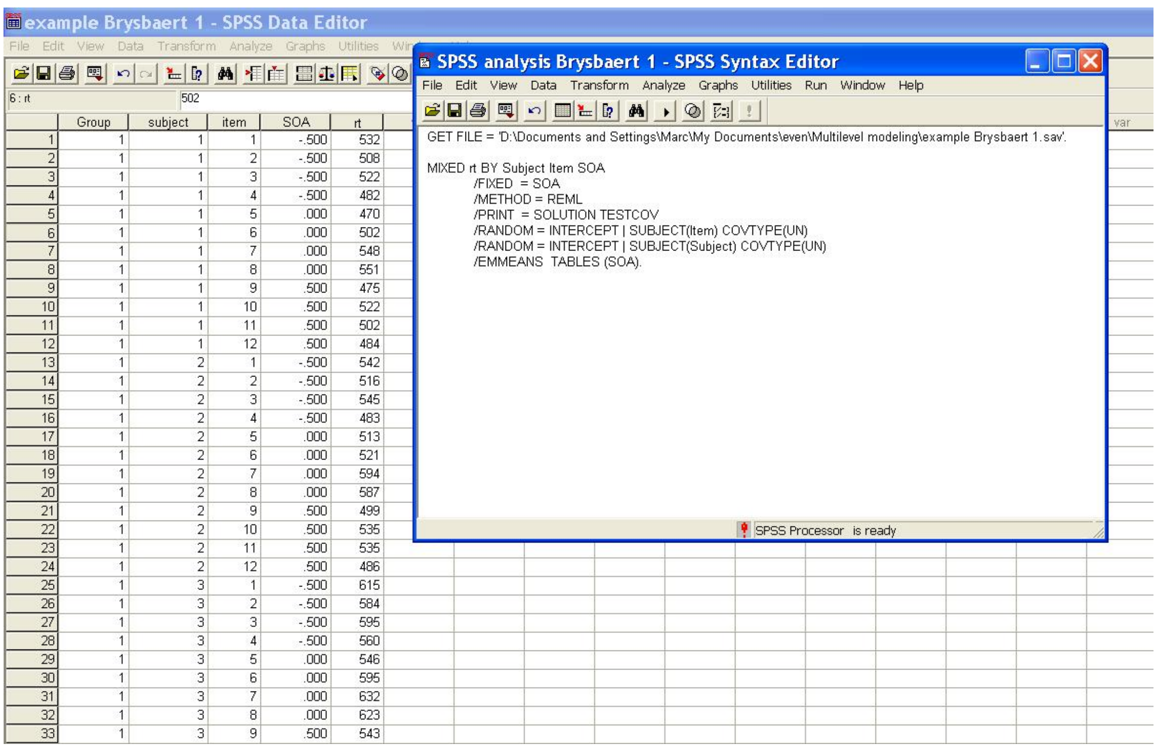
a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

Interestingly, in this analysis Baayen et al. (2006) do not define an extra random variable to capture the repeated measures element. Still, the F-value is quite comparable to the one obtained by Raaijmakers et al., even though it has a much bigger df2 (due to the fact that much less parameters must be estimated).

A concern about the above analysis may be that it doesn't matter that much which analysis you use when the effect is small. So, to see how the different analyses compare when the effects are slightly more interesting, I added 4 ms to the medium SOA condition (half of the data got +4, one quarter +2, and the remaining quarter +6). Given that the variability of the data is quite low, this should suffice to find significance, which is indeed what I found when I ran the usual F1, F2, and minF':

$F1(2,18) = 5.481$, $p = .014$
$F2(2,18) = 9.426$, $p = .002$
$minF'(2,34) = 3.456$, $p = .043$
multilevel $F(2,119) = 6.742$, $p = .002$

The multilevel analysis gave the following:



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	19.797	1538.707	.000
SOA	2	119.000	6.742	.002

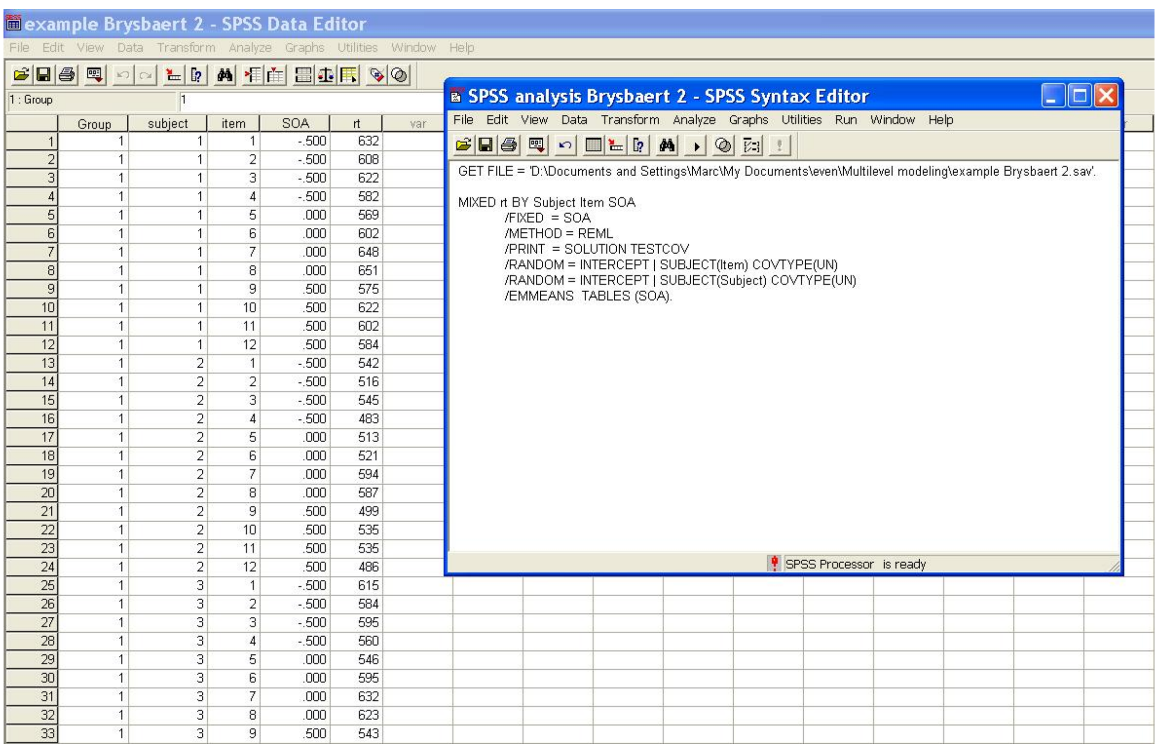
a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	533.95833	13.709200	20.082	38.949	.000	505.3689197	562.5477469
[SOA=-.500]	-.4583333	2.0020390	119.000	-.229	.819	-4.4225705	3.5059039
[SOA=.000]	6.1250000	2.0020390	119.000	3.059	.003	2.1607628	10.0892372
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

Something else I tried, was see what happens if one participant gets an extra 100 ms on all items (see the example above for the slow participant). If the underlying reasoning of the technique is what it claims to be, then this should have no effect on the F-statistic for SOA, because the change can easily be captured by a different intercept for the participant involved. So, we should get rid of the requirement to introduce between-items Latin-square variables or the necessity to work with z-scores. This is exactly what happened, as can be seen in the following tables:



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	18.743	1383.900	.000
SOA	2	119.000	6.742	.002

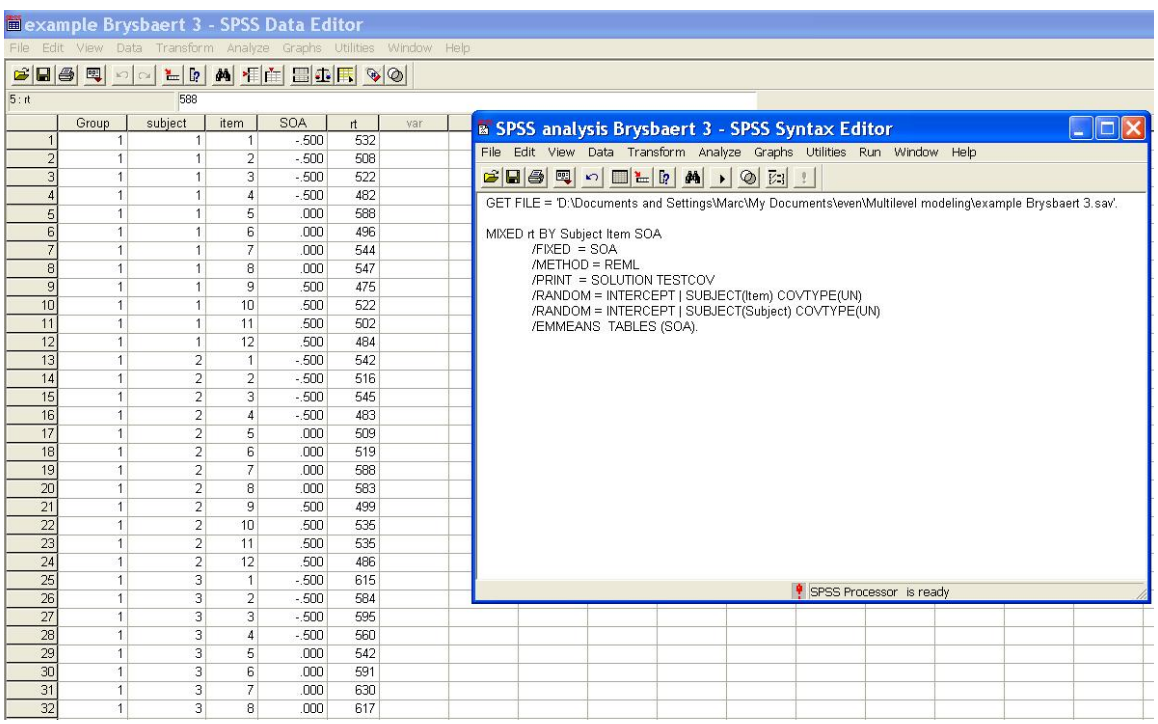
a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	542.29167	14.673791	18.978	36.956	.000	511.5766442	573.0066891
[SOA=-.500]	-.4583333	2.0020390	119.000	-.229	.819	-4.4225705	3.5059039
[SOA=.000]	6.1250000	2.0020390	119.000	3.059	.003	2.1607628	10.0892372
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

Finally, I wanted to see what happens when 1 observation in Raaijmakers et al.'s table got a much higher value (participant 1, item 5 +120 ms). Will this turn the multilevel F-statistic into a spurious significance?



Type III Tests of Fixed Effects(a)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	19.421	1610.356	.000
SOA	2	119.000	2.055	.133

a Dependent Variable: Response Time in Milliseconds.

Estimates of Fixed Effects(b)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	533.95833	13.436372	19.984	39.740	.000	505.9291461	561.9875206
[SOA=-.500]	-.4583333	2.7740527	119.000	-.165	.869	-5.9512348	5.0345681
[SOA=.000]	4.6250000	2.7740527	119.000	1.667	.098	-.8679014	10.1179014
[SOA=.500]	0(a)	0	.	.	.	.	.

a This parameter is set to zero because it is redundant.
b Dependent Variable: Response Time in Milliseconds.

The obtained F-value [F(2,119)=2.055, p < .133] compares favorably to what happens with F1 and F2 (although not to minF', which is a good reminder that the F1 x F2 criterion may give the wrong impression):

F1(2,18) = 2.739, p = .092
F2(2,18) = 3.061, p = .072
minF'(2,36) = 1.44, p = .249

Finally, the multilevel design is not limited to a single IV. Locker et al. (2007) give an example of an LDT experiment in which the effects of phonological neighborhood frequency and semantic neighborhood size were measured. This is their code (which can easily be adapted).

```
MIXED rt BY Subject Item Freq Size
    /FIXED  = Freq Size Freq*Size
    /METHOD = REML
    /PRINT  = SOLUTION TESTCOV
    /RANDOM = INTERCEPT | SUBJECT(Item) COVTYPE(UN)
    /RANDOM = INTERCEPT | SUBJECT(Subject) COVTYPE(UN)
    /EMMEANS  TABLES (Freq*Size).
```

In summary, I am becoming more and more convinced that multilevel modeling is the way forward. The analyses are easier than than the F1, F2, and minF' calculations and they seem to be of a higher quality. In the final section, I refer to one more advantage of the multilevel approach.

7. Beyond dichotomizing

For someone with a bit of experience in analyzing psycholinguistic data, the idea of simultaneously controlling for item and participant variation must ring a bell. In 1990, Lorch and Myers published an article on how to do a proper linear regression in a repeated measures design. The problem is analogue to the one discussed in Figure 1, although now it involves generalization over participants.

The problem is illustrated in Table 4, where the results are shown for 6 participants on 10 items that vary in log10(frequency).

	LogFreq	RTpart1	RTpart2	RTpart3	RTpart4	RTpart5	RTpart6
1	.25	900	625	601	706	821	489
2	.50	850	654	609	652	812	512
3	.75	800	699	614	717	845	497
4	1.00	750	599	610	642	854	468
5	1.25	700	652	630	713	823	501
6	1.50	650	603	624	695	832	466
7	1.75	600	631	637	689	861	484
8	2.00	550	622	629	664	815	503
9	2.25	500	669	643	703	769	498
10	2.50	450	599	641	678	804	527

Table 4 : Example of regression data in a design with a repeated measure (LDT to 10 words varying in frequency).

If we average the data over the 6 participants and calculate the regression analysis, we get:

RT = 702 – 33.5 LogFreq (LogFreq: $t(8) = -7.588$, $p < .001$, $R^2 = .88$).

A look at Table 4 makes clear where this huge frequency effect comes from (and how things can go pear-shaped). Only one of the participants (i.e., part1) shows a substantial linear frequency effect. All the others show either no effect or even a slight opposite effect. Unfortunately, this variability is lost when the regression is based on the mean RT over participants.

To counter this problem, Lorch & Myers (1990) suggested to do a separate analysis per participant and then to run a t-test on the regression weights obtained. So, they would do the following calculations:

Part1 : 950 – 200 LogFreq
Part2 : 651 – 11.3 LogFreq
Part3 : 599 + 18.1 LogFreq
Part4 : 687 – .9 LogFreq
Part5 : 843 – 13.9 LogFreq
Part6 : 485 + 7.0 LogFreq

A simple one-sample t-test reveals that in the Lorch & Myers (1990) analysis, the effect of LogFreq is not significant ($t(5) = -.996$, $p = .365$).

Ever since many psycholinguists have happily spent days calculating regression weights of individual participants and running one-sample t-tests on them, even though apparently there is a simpler way to get at it directly from the ANOVA table.

If you want to have a go at this type of analysis, here is the example Lorch & Myers worked with in their article. It deals with sentence reading times as a function of the rank order of the sentence, the number of words in the sentence, and the number of new words in the sentence.

OBSERVATIONS 153

Table 3
Subjects' Reading Times and Values of Predictor Variables for Each Sentence of the Experimental Text

SNT	SP	WRDS	NEW	SBJ 1	SBJ 2	SBJ 3	SBJ 4	SBJ 5	SBJ 6	SBJ 7	SBJ 8	SBJ 9	SBJ 0
1	1	13	1	3429	2795	4161	3071	3625	3161	3232	7161	1536	4063
2	2	16	3	6482	5411	4491	5063	9295	5643	8357	4313	2946	6652
3	3	9	2	1714	2339	3018	2464	6045	2455	4920	3366	1375	2179
4	4	9	2	3679	3714	2866	2732	4205	6241	3723	6330	1152	3661
5	5	10	3	4000	2902	2991	2670	3884	3223	3143	6143	2759	3330
6	6	18	4	6973	8018	6625	7571	8795	13188	11170	6071	7964	7866
7	7	6	1	2634	1750	2268	2884	3491	3688	2054	1696	1455	3705

Note. SNT = sentence; SP = serial position of sentence; WRDS = number of words in sentence; NEW = number of new arguments in sentence; SBJ = subject.

This is the analysis Lorch & Myers reported:

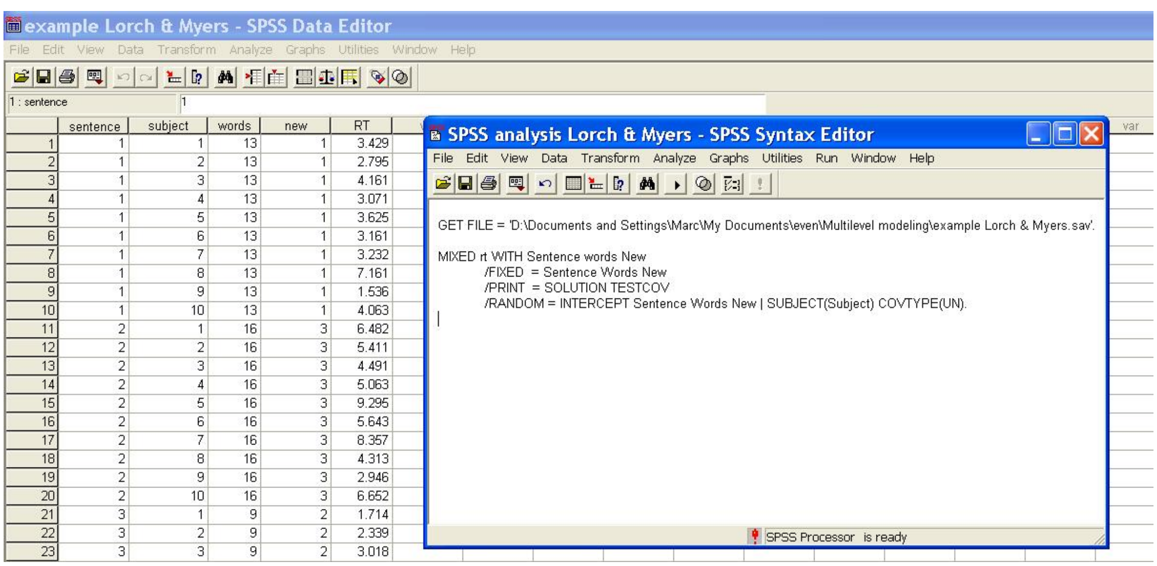
Table 4
Regression Coefficients From Individual Analyses of Subjects' Data in Reading Experiment

Subject	SP	WORDS	NEW
1	0.23124	0.39103	0.22161
2	0.30533	0.43415	0.34637
3	0.20637	0.40360	−.25294
4	0.48300	0.50203	−.27683
5	−0.06210	0.28778	0.92680
6	1.10982	0.80850	−.23336
7	0.25448	0.57498	0.79643
8	−0.33147	0.11341	0.33124
9	0.66786	0.50078	0.16320
10	0.46921	0.56964	−.50621
M	0.33337	0.45859	0.15163
SE	0.12417	0.05855	0.14982
t	2.6849	7.83289	1.01210

Note. SP = serial position of sentence; WORDS = number of words in sentence; NEW = number of new arguments in sentence.

From this they concluded that the serial position of the sentence and the number of words were significant predictors of reading time, but not the number of new words.

Van den Noortgate and Onghena (2006) used this example to show how much easier multilevel programming is. The nice thing about the MIXED function is that it not only works with discrete variables but also with continuous variables (the only thing you have to change is to use WITH instead of BY in the model specification). This is the program Van den Noortgate & Onghena used:



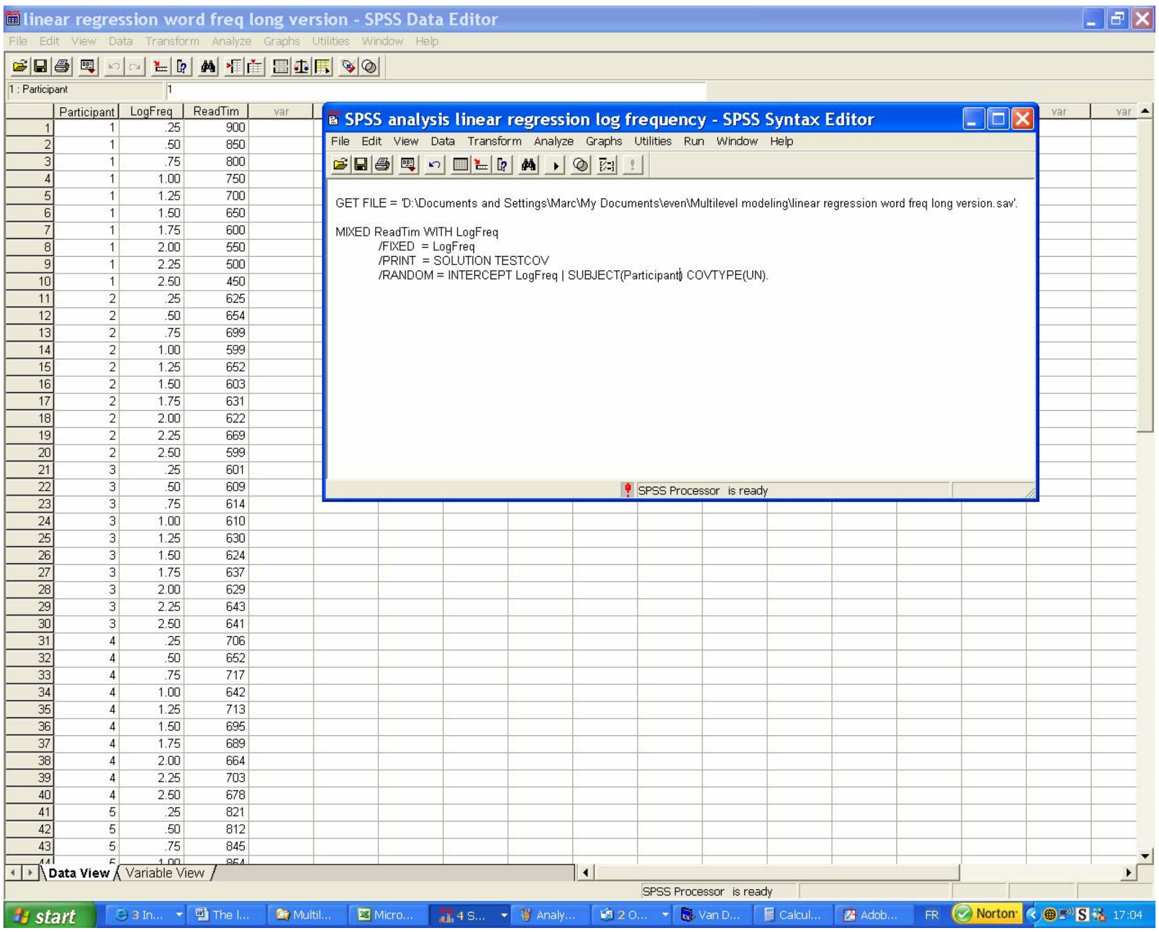
with the following results:

Estimates of Fixed Effects(a)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	-2.586950	.7425953	19.755	-3.484	.002	-4.1372114	-1.0366896
sentence	.3333728	.0989789	36.617	3.368	.002	.1327516	.5339941
words	.4585893	.0680731	36.617	6.737	.000	.3206113	.5965673
new	.1516299	.2560739	36.617	.592	.557	-.3674087	.6706684

a Dependent Variable: Reading time.

When we do the same analysis on on our simple example with the word frequency data, we get



Estimates of Fixed Effects(a)

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	702.42222	68.77472	5.000	10.213	.000	525.6311788	879.2132657
LogFreq	-33.50708	33.646936	5.000	-.996	.365	-119.9992728	52.9851314

a Dependent Variable: ReadTim.

8. Conclusion

There is an ongoing complaint among teachers and lecturers that students nowadays know less than students some time ago (despite the Flynn-effect). Until recently I thought this was because teachers and lecturers were good students themselves and therefore have a biased view of the motivation and the level of knowledge of their cohort (as they did not tend to interact a lot with the 'bad' students). A few months ago, however, I came across an article in which an educational psychologist gave another explanation. According to him, teachers in particular see the lack of knowledge in students for what *they themselves* know well on the basis of their education (e.g., history, geography, correct spelling, algebra, elementary statistics, …), but they fail to notice the knowledge pupils/students have that is not shared by teachers/lecturers. When it comes to acquiring new knowledge and skills, teachers are no better than students if the immediate use of the knowledge is not obvious.

This view has crossed my mind a few times in the past couple of days: Is it possible that we keep on clutching to the familiar F1 and F2, because we've learned to calculate them in our undergraduate studies (in my case even by hand)? My present journey most certainly has convinced me that I seem to have missed a few steps in current statistical sophistication. It certainly is an incentive to explore the ***lme4*** package (http://cran.r-project.org), which has many more goodies and possibilities than what is on offer in SPSS (Baayen, 2007; Baayen et al., 2006). The present review shows that a better understanding of multilevel analysis techniques (or mixed-effects techniques) is likely to be rewarding, although it is amazing how much is already available in the statistical program we use daily, at no larger clicking cost than we are doing now (often quite the contrary as I have found out)!

9. References

Baayen, R.H. (2007). Analyzing linguistic data: A practical introduction to statistics. Cambridge: Cambridge University Press (in press).

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2006). *Mixed-effects modeling with crossed random effects for subjects and items.* Available on the internet (copy the title in google).

Clark, H.H. (1973). The language-as-fixed effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Locker, L., Hoffman, L., & Bovaird, J. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*.

Lorch, R.F., & Myers, J.L. (1990). Regression analysis of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 149-157.

Pollatsek, A., & Well, A.D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 785-794.

Raaijmakers, J.G.W. (2003). A further look at the "language-as-fixed-effect fallacy'. *Canadian Journal of Experimental Psychology, 57*, 141-151.

Raaijmakers, J.G.W., Schrijnemakers, J.M.C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Van den Noortgate, W., & Onghena, P. (2006). Analysing repeated measures in cognitive research: A comment on regression cofficient analyses. *European Journal of Cognitive Psychology, 18*, 937-952.