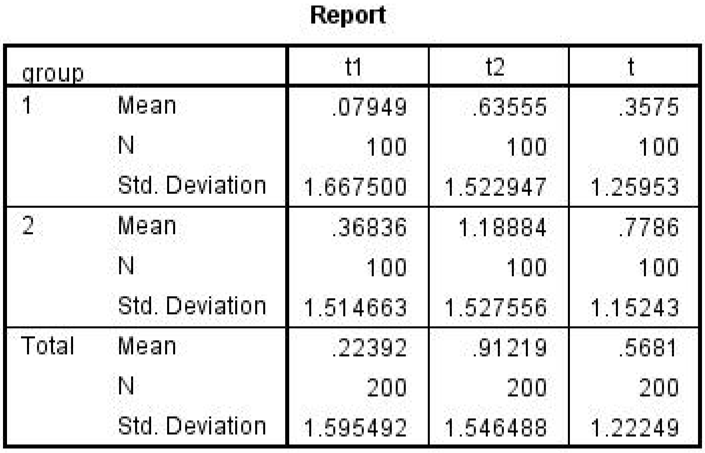
Standardised Effect Size in a Mixed/Multilevel Model

This note uses simple examples based on two or more groups (*group*), and measurements at two time points (*time*), to consider how standardised effect sizes can be derived from analyses using the mixed linear model. The emphasis is on obtaining measures of effect size which are equivalent to those which would be obtained from *t*-tests.

The means and standard deviations of the variables (*d data 1.sav*[1]):



Report

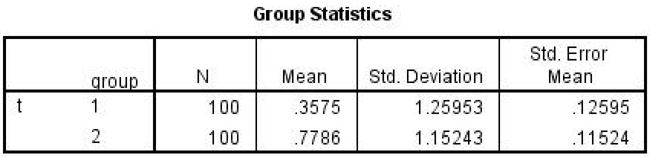
group		t1	t2	t
1	Mean	.07949	.63555	.3575
	N	100	100	100
	Std. Deviation	1.667500	1.522947	1.25953
2	Mean	.36836	1.18884	.7786
	N	100	100	100
	Std. Deviation	1.514663	1.527556	1.15243
Total	Mean	.22392	.91219	.5681
	N	200	200	200
	Std. Deviation	1.595492	1.546488	1.22249

t is the mean of *t1* and *t2*. The correlation between t1 and t2 is .211.

Between-subject effect

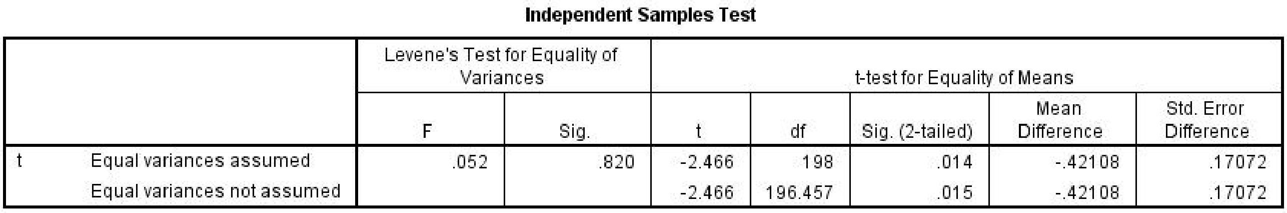
The main effect of *group* can be seen as the difference between the two groups in terms of scores which, for each subject, are the mean of their time 1 and time 2 scores (*t* above).

The results of a *t*-test comparing *t* (the mean of *t*1 and *t*2) the two groups (there are 100 cases in each group) are shown below.



Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
t	1	100	.3575	1.25953	.12595
	2	100	.7786	1.15243	.11524



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
t	Equal variances assumed	.052	.820	-2.466	198	.014	-.42108	.17072
	Equal variances not assumed			-2.466	196.457	.015	-.42108	.17072

Cohen's *d* is calculated as the difference between the means of the two groups divided by the pooled within-group standard deviation:

[1] The named datasets are downloadable from the same page as this document - *http://www.psy.mq.edu.au/psystat/SPSSGeneral.html*

d = (mean 1 – mean 2)/$\sqrt{}$[([n – 1]*σ_1^2 + [n – 1]* σ_2^2)/(n_1 + n_2 – 2)]

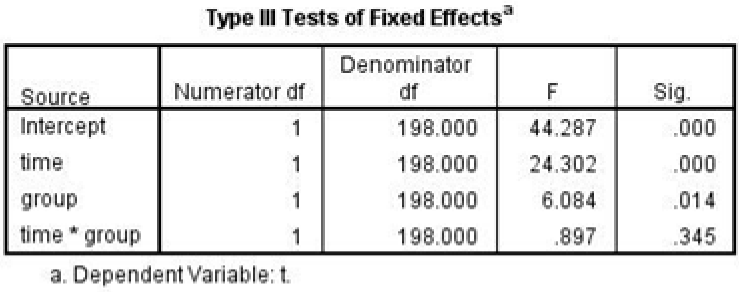
d = (.7786 - .3575)/$\sqrt{}$[([100 – 1]*1.25953^2 + [100 – 1]*1.15243^2)/200 – 2] = .34883.

In this calculation, the difference between the means is .42110 and the pooled SD is 1.20717.

If the data in this example are stacked (*d data 1 – stacked.sav*), so that each person is represented by two lines of data, one for time 1 and the other for time 2, a linear mixed model analysis can be carried out in SPSS with the following syntax:

```
mixed t by time group/
 fixed=intercept time group time by group/
 random=intercept | subject(id)/
 print=solution testcov/
 emmeans=table(group).
```

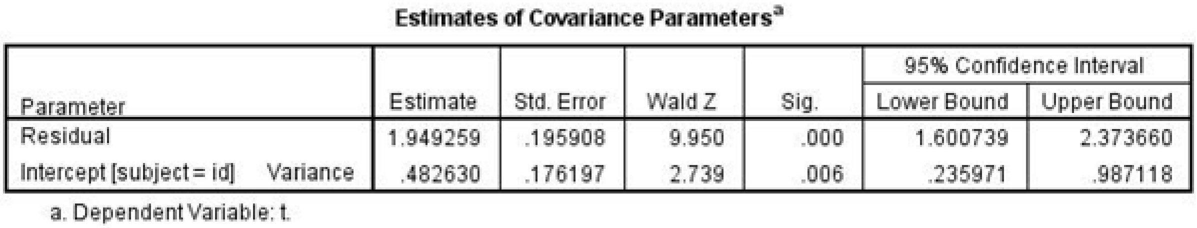
The results of an analysis in which the factors are *time* and *group* and both main effects and the interaction are tested are shown below.



Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	198.000	44.287	.000
time	1	198.000	24.302	.000
group	1	198.000	6.084	.014
time * group	1	198.000	.897	.345

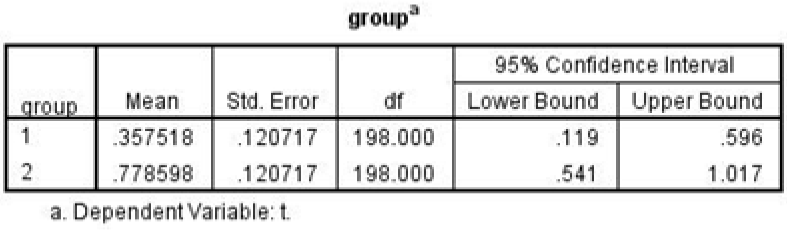
a. Dependent Variable: t.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		1.949259	.195908	9.950	.000	1.600739	2.373660
Intercept [subject= id]	Variance	.482630	.176197	2.739	.006	.235971	.987118

a. Dependent Variable: t.



group[a]

group	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
1	.357518	.120717	198.000	.119	.596
2	.778598	.120717	198.000	.541	1.017

a. Dependent Variable: t.

Note that the difference between the means shown in the estimated marginal means table (.778598 - .357518 = .4211), and the *p*-value for the group comparison (.014), are the same as those for the *t*-test.

d from the variance estimates in the mixed model

In the Estimates *of Covariance Parameters* table, the *Residual* parameter (1.949259) is the variance within subjects, while the *Intercept* parameter (.482630) is the between-subject variance. In principle, the square root of the intercept parameter (.694716) could be used to calculate a measure of effect size, but this measure would be considerably larger than that obtained with the information from the *t*-test. The reason for this is that the estimate of within-cell variance based on the *t*-test results, as well as reflecting between-subject variability, also reflects the variability between the two observations for each subject. In order to obtain a *d*-value equal to that obtained from the *t*-test based on the averaged observations from the mixed analysis, the intercept variance must be included in the calculation. Also, because the two observations are averaged (see Hedges, 2007, p. 347), the appropriate estimate of the contribution of within-subject variance is the within-subject variance estimate from the mixed model output, divided by the number of observations for each subject, plus the estimate of intercept variance. The estimate of the between-subject standard deviation equivalent to that derived from the *t*-test analysis is therefore:

$\sqrt{}(1.949259/2 + .482630) = 1.20717$.

The general rule is that a SD equivalent to that obtained from the *t*-test analysis can be obtained from the mixed model output by dividing the residual variance estimate by the number of observations for each subject (or in each cluster), adding the result to the between-subject variance estimate, and taking the square root of the sum.

d from the Estimated Marginal Means table

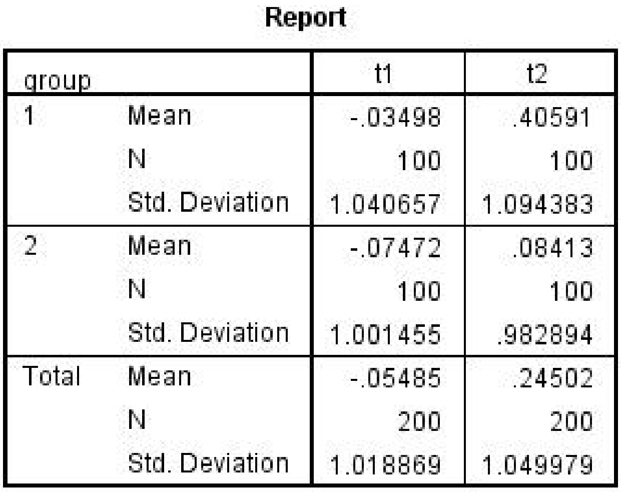
Another way of obtaining the SD necessary to calculate the effect size is to use the standard errors in the estimated marginal means table. Given that the standard error is equal to the standard deviation over the square root of the sample size, the SD in this case is equal to $.120717 * \sqrt{}100$ which is 1.20717. This is the same result as that obtained from the variance estimate.

In this example, the standard error is the same in both groups; with unequal sample sizes, it would be necessary to obtain a pooled estimate based on the standard errors for each group. Differences in the standard deviations of the groups are sometimes taken into account when calculating *d*.

Within-subject effect

When two independent groups are compared, the pooled within-group standard deviation is used as the denominator when calculating *d*. This SD is the basis for the standard error which is used to calculate the *t*-statistic when testing the significance of differences. The equivalent quantities when carrying out a paired *t*-test are the standard deviation of the *difference score* and the standard error of the difference. It would therefore be logical to use the SD of the difference score in the calculation of *d* in paired *t*-test analyses. However, it has been argued by Dunlap *et al* (1996) that, because the standard deviation of the difference score is affected by the correlation between the paired measurements (with higher correlations leading to smaller standard deviations, as shown by the expression for the variance of the difference between two variables: $\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2$ where *r* is the correlation coefficient) the calculation of *d* in the paired case should use the same standard deviations as those used in the independent groups case.

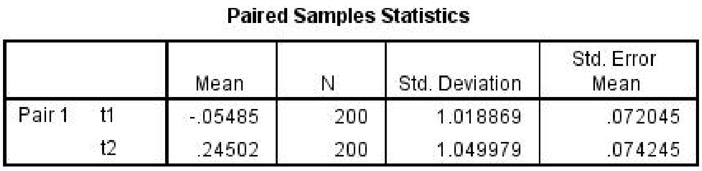
The following output, similar in structure to that used above, but with a higher correlations (.7) between the *t1* and *t2* measures, will be used to illustrate. The means and standard deviations for this dataset (*d data 2.sav*)are:



Report

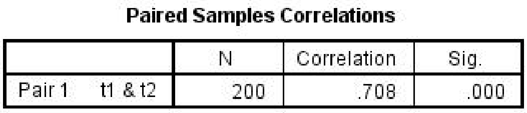
group		t1	t2
1	Mean	-.03498	.40591
	N	100	100
	Std. Deviation	1.040657	1.094383
2	Mean	-.07472	.08413
	N	100	100
	Std. Deviation	1.001455	.982894
Total	Mean	-.05485	.24502
	N	200	200
	Std. Deviation	1.018869	1.049979

The correlation between *t*1 and *t*2 is .708.



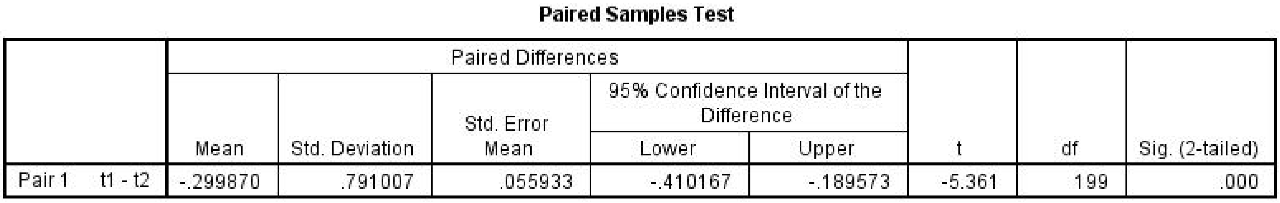
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	t1	-.05485	200	1.018869	.072045
	t2	.24502	200	1.049979	.074245



Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	t1 & t2	200	.708	.000



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	t1 - t2	-.299870	.791007	.055933	-.410167	-.189573	-5.361	199	.000

If *d* is calculated with the mean difference over the standard deviation of the difference score, we obtain .299870/.791007 = .379099, whereas if we use the mean standard deviation of the original scores (1.018869 + 1.049979)/2 = 1.034424) we obtain a *d* of .289891.

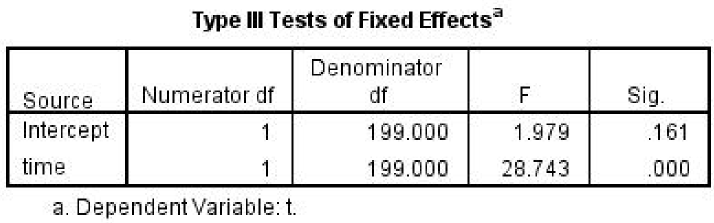
d from the variance estimates in the mixed model

If the data in the above example are stacked (*d data 2 –stacked*.sav) and the following mixed model analysis is carried
out,

mixed t by time/
  fixed=intercept time/

print=solution testcov/
emmeans=table(time) compare(time).

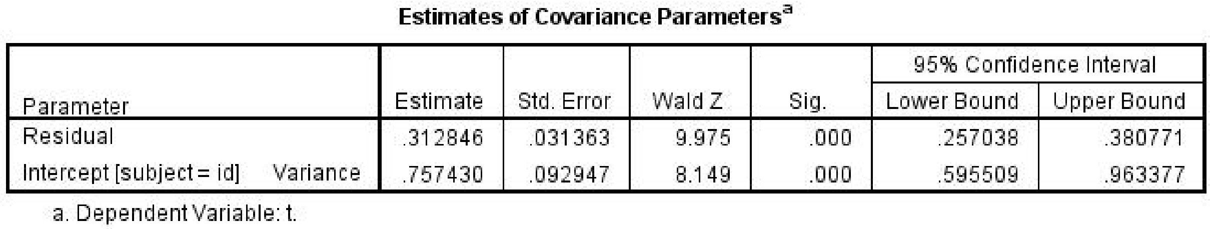
the following results are obtained:



Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	199.000	1.979	.161
time	1	199.000	28.743	.000

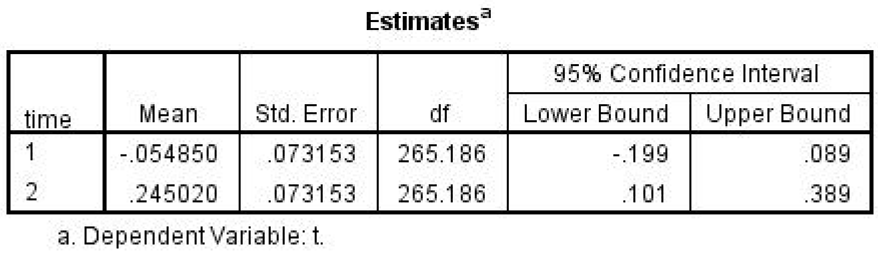
a. Dependent Variable: t.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		.312846	.031363	9.975	.000	.257038	.380771
Intercept [subject = id]	Variance	.757430	.092947	8.149	.000	.595509	.963377

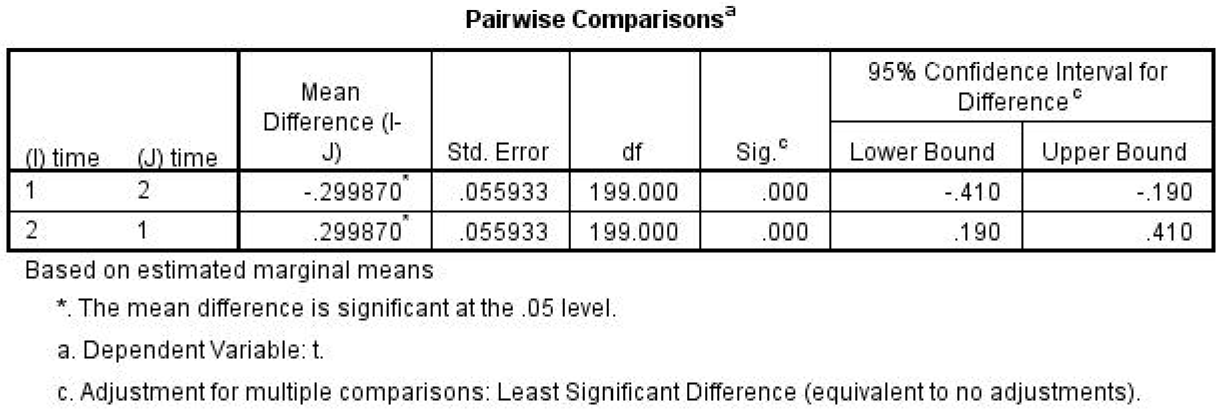
a. Dependent Variable: t.



Estimates[a]

time	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
1	-.054850	.073153	265.186	-.199	.089
2	.245020	.073153	265.186	.101	.389

a. Dependent Variable: t.



Pairwise Comparisons[a]

(I) time	(J) time	Mean Difference (I-J)	Std. Error	df	Sig.[c]	95% Confidence Interval for Difference[c]	
						Lower Bound	Upper Bound
1	2	-.299870*	.055933	199.000	.000	-.410	-.190
2	1	.299870*	.055933	199.000	.000	.190	.410

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Dependent Variable: t.

c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Note that the square root of the *F*-ratio of 28.743 for the time effect (from the table for the *Type III Tests of Fixed Effects*) is equal to the square of the *t*-statistic obtained in the paired *t*-test, and that the standard error for the pairwise comparisons (.055933) is the same as the standard error of the difference in the *t*-test.

Bearing in mind the suggestions in the Dunlap *et al* article, we would like to find an appropriate standard deviation on which to base a calculation of *d*.

We want a standard deviation which combines both within- and between-subject variance, so the residual and intercept variances are summed and the square root taken to obtain $\sqrt{}$(.312846 + .757430) = 1.034541. Because in this case we are not dealing with an averaged quantity (the mean of the *t1* and *t2*), the residual variance is not divided by the number of cases in each cluster. The resulting value is approximately equal to the average of the standard deviations of *t1* and *t2* shown in the *t*-test output. Using this standard deviation we obtain a *d* value equal to .289891.

d from the Estimated Marginal Means table

As was the case with the between-subject factor, it is possible to obtain the appropriate standard deviation for the calculation of *d* from the table of estimated marginal means. The standard error in the table is .073153. If this is multiplied by the square root of the number of cases (200), the result is 1.034540 and the value of *d* is as calculated above.

Conclusion

Probably the simplest way to calculate *d* values from mixed model output is to request estimated marginal means and use the standard errors to derive standard deviations. It has been shown above that the estimates arrived at in this way are consistent with those derived from the situation for which *d* was developed, that involving two independent groups.

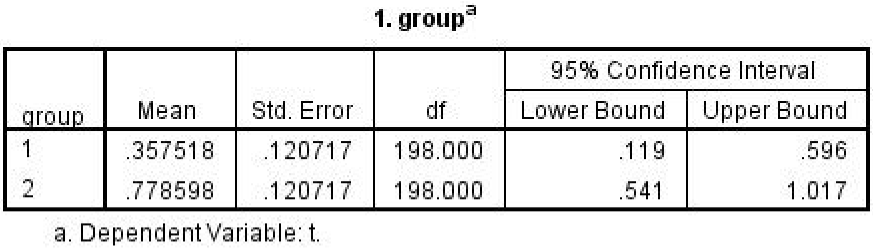
When deriving the standard deviations from standard errors, it is important to use the correct number of observations in the formula SD = SE*√*n*. In the first example, *n*=100 because there were 100 subjects in each group, with their time 1 and time 2 scores averaged. In the second example, there were 200 observations at each time point, so *n*=200. A further point is that because, in these examples, there were equal numbers of cases in each group, the standard error was the same for both, so it was not necessary to pool the estimates; when there are different numbers of cases in each group this would be necessary for the between-subjects case. For the within-subjects case, it would probably best to base the standard deviation on the standard error for the first time point.

Tests of Simple Effects

When tests of simple effects are used, and it is necessary to calculate *d* for, say, the difference between groups at each time point or between times for each group, it is not appropriate to use the standard deviation obtained for the main effect of each variable as the denominator for the respective calculations. Returning to the dataset used in the first example (*d data 1 – stacked.sav*), and running these commands:

```
mixed t by time group/
 fixed=intercept time group time by group/
 random=intercept | subjects(id)/
 print=solution testcov/
 emmeans=table(group)/
 emmeans=table(time)/
 emmeans=table(group*time).
```

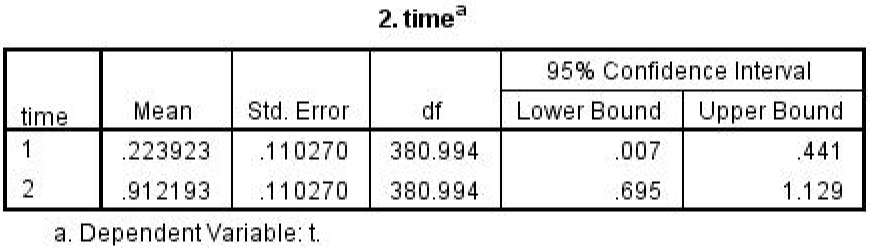
we obtain these tables of estimated marginal means:



1. group[a]

group	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
1	.357518	.120717	198.000	.119	.596
2	.778598	.120717	198.000	.541	1.017

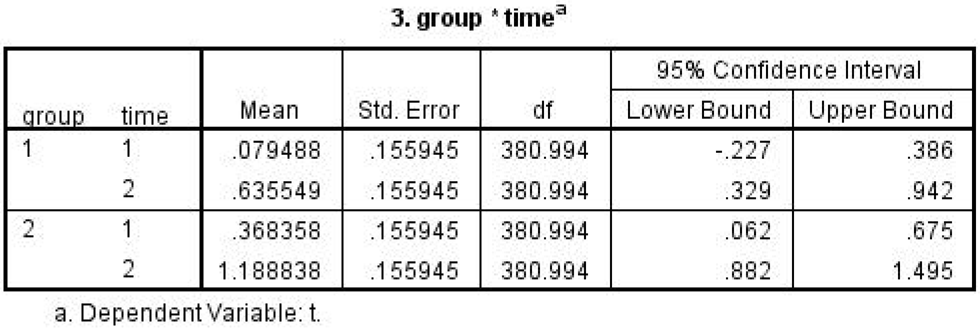
a. Dependent Variable: t.



2. time[a]

time	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
1	.223923	.110270	380.994	.007	.441
2	.912193	.110270	380.994	.695	1.129

a. Dependent Variable: t.



3. group * time[a]

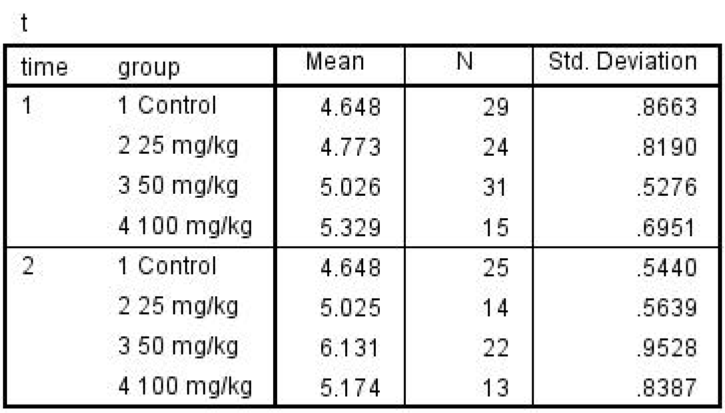
group	time	Mean	Std. Error	df	95% Confidence Interval	
					Lower Bound	Upper Bound
1	1	.079488	.155945	380.994	-.227	.386
	2	.635549	.155945	380.994	.329	.942
2	1	.368358	.155945	380.994	.062	.675
	2	1.188838	.155945	380.994	.882	1.495

a. Dependent Variable: t.

If the standard error shown in the third table (*group * time*) was used, the SD obtained for use in calculating *d* would be .155945*√100 = 1.55945. This is the same as that obtained from the table for *time*: .110270*√200 = 1.559453. That obtained from the information in the first (*group*) is different: .120717*√100 = 1.20717. The reason for this is that the between-subject test is assumed to be based on the mean of the time 1 and time 2 variables.

An example of testing simple effects based on more realistic data

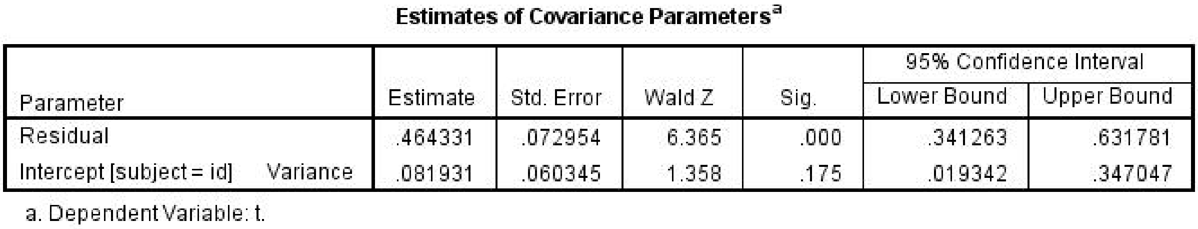
In this example (*d data 3.sav*) there are still two time points, but there are now four groups, containing varying numbers of cases (and each group has different numbers at time 1 and time 2, which is realistic for non-experimental studies involving follow-up) and for which the standard deviations differ, as below.



t

time	group	Mean	N	Std. Deviation
1	1 Control	4.648	29	.8663
	2 25 mg/kg	4.773	24	.8190
	3 50 mg/kg	5.026	31	.5276
	4 100 mg/kg	5.329	15	.6951
2	1 Control	4.648	25	.5440
	2 25 mg/kg	5.025	14	.5639
	3 50 mg/kg	6.131	22	.9528
	4 100 mg/kg	5.174	13	.8387

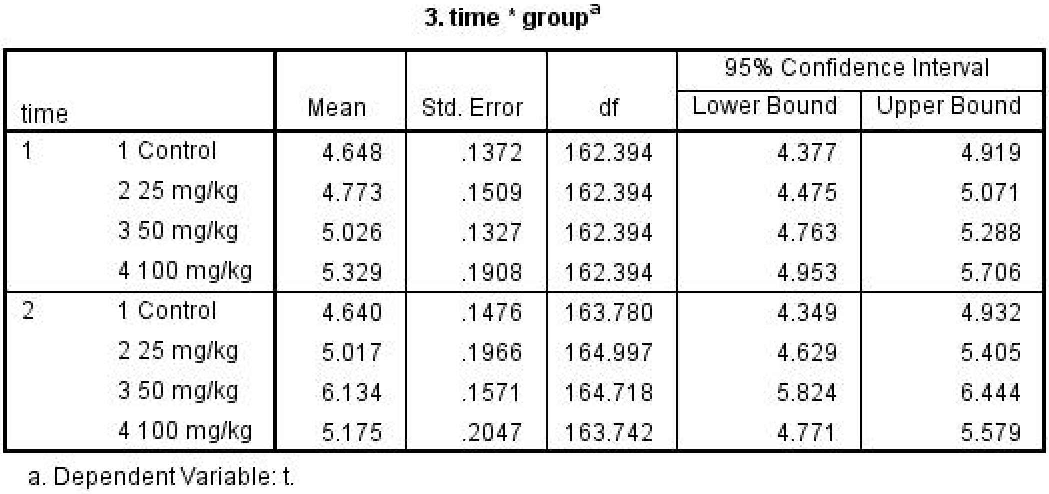
When a mixed model analysis is carried out, the following results (among others) are obtained.



Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		.464331	.072954	6.365	.000	.341263	.631781
Intercept [subject = id]	Variance	.081931	.060345	1.358	.175	.019342	.347047

a. Dependent Variable: t.



3. time * group[a]

time		Mean	Std. Error	df	95% Confidence Interval	
					Lower Bound	Upper Bound
1	1 Control	4.648	.1372	162.394	4.377	4.919
	2 25 mg/kg	4.773	.1509	162.394	4.475	5.071
	3 50 mg/kg	5.026	.1327	162.394	4.763	5.288
	4 100 mg/kg	5.329	.1908	162.394	4.953	5.706
2	1 Control	4.640	.1476	163.780	4.349	4.932
	2 25 mg/kg	5.017	.1966	164.997	4.629	5.405
	3 50 mg/kg	6.134	.1571	164.718	5.824	6.444
	4 100 mg/kg	5.175	.2047	163.742	4.771	5.579

a. Dependent Variable: t.

Following the calculation given earlier, the standard deviation derived from the *Estimates* table is √(.464331 + .081931) = .7391. This value is pooled over both groups and times. It can be used for calculating *d* for the simple effects of both *time* and *group* and provides a consistent denominator for the calculations. This might be seen as a drawback when calculating *d* for groups differing in size and/or standard deviation when the magnitude of the effect size may not seem to be consistent with the significance of the difference.

Turning now to the *time* by *group* table of estimated means, we can take a more piecemeal approach to calculating *d*. To obtain the pooled standard deviation needed to calculate *d* for the comparison of two groups at a particular time, or of two times for a particular group, we can use the formula:

$$\mathrm{SD}_{\mathrm{pooled}} = \surd\ (((se_1 * \surd n_1)^2 * (n_1 - 1) + (se_2 * \surd n_2)^2 * (n_2 - 1))/(n_1 + n_2 - 2))$$

$$= \surd\ ((se_1^2 * n_1 * (n_1 - 1) + se_2^2 * n_2 * (n_2 - 1))/(n_1 + n_2 - 2))$$

$$\approx \surd\ ((se_1^2 * n_1^2 + se_2^2 * n_2^2)/(n_1 + n_2))$$

$$= \surd\ ((se_1 * n_1)^2 + (se_2 * n_2)^2/(n_1 + n_2)) \text{ for large } n_s.$$

For example, *d* for the difference between the means of group 1 and group 4 at time 1 is:

(5.329 – 4.648)/

$\sqrt{}$ ((.1908^2 * 15 * (15 – 1) + .1372^2 * 29 * (29 – 1))/(15 + 29 – 2))

= .6810/.7389 = .9216.

This compares to .6810/.7391 = .9214 when the overall estimate of SD is used.

The difference between times 1 and 2 for group 3 is:

(6.134 – 5.026)/

$\sqrt{}$ ((.1571^2 * 22 * (22 – 1) + .1327^2 * 31 * (31 – 1))/(22 + 31 – 2))

= 1.108/.7380 = 1.5014

This compares to 1.108/.7391 = 1.4991 when the overall estimate of SD is used.

Note that if the number of cases in each group is the same at each time point, as they might be in an experimental study, the estimated standard deviation is the same for any *se* * $\sqrt{n}$ pair.

Alan Taylor
21st of March 2014
Additions and amendments 7th of September 2015

Thanks to Greg Baker for reading and commenting on an earlier version of this note and to Judith Fethney for prompting, and helping with, the additional section on testing simple effects, and for reading and commenting on the revised document.

References

Dunlap, W.P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1, 170-177

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 41-370.