Chapter 14: Repeated-measures designs

Smart Alex's Solutions

Task 1

> *It is common that lecturers obtain reputations for being 'hard' or 'light' markers (or to use the students' terminology, 'evil manifestations from Beelzebub's bowels' and 'nice people') but there is often little to substantiate these reputations. A group of students investigated the consistency of marking by submitting the same essays to four different lecturers. The mark given by each lecturer was recorded for each of the eight essays. The independent variable was the lecturer who marked the report and the dependent variable was the percentage mark given. The data are in the file* ***TutorMarks.sav****. Conduct a one-way ANOVA on these data by hand*

Essay	Tutor 1 (Prof Field)	Tutor 2 (Prof Smith)	Tutor 3 (Prof Scrote)	Tutor 4 (Prof Death)	Mean	s^2
1	62	58	63	64	**61.75**	**6.92**
2	63	60	68	65	**64.00**	**11.33**
3	65	61	72	65	**65.75**	**20.92**
4	68	64	58	61	**62.75**	**18.25**
5	69	65	54	59	**61.75**	**43.58**
6	71	67	65	50	**63.25**	**84.25**
7	78	66	67	50	**65.25**	**132.92**
8	75	73	75	45	**67.00**	**216.00**
Mean	**68.875**	**64.25**	**65.25**	**57.375**		

There were eight essays, each marked by four different lecturers. Their marks are shown in the table above. In addition, the mean mark given by each lecturer is shown in the table, and also the mean mark that each essay received and the variance of marks for a particular essay. Now, the total variance within essays will in part be caused by the fact that different lecturers are harder or softer markers (the manipulation), and in part by the fact that the essays themselves will differ in quality (individual differences).

The total sum of squares (SS_T)

Remember from one-way independent ANOVA that SS_T is calculated using the following equation:

$$\mathrm{SS_T} = s^2_{\mathrm{grand}}(N-1)$$

Well, in repeated-measures designs the total sum of squares is calculated in exactly the same way. The grand variance in the equation is simply the variance of all scores when we ignore the group to which they belong. So if we treated the data as one big group it would look as follows:

62	*58*	*63*	*64*
63	*60*	*68*	*65*
65	*61*	*72*	*65*
68	*64*	*58*	*61*
69	*65*	*54*	*59*
71	*67*	*65*	*50*
78	*66*	*67*	*50*
75	*73*	*75*	*45*

Grand Mean = 63.9375

The variance of these scores is 55.028 (try this on your calculator). We used 32 scores to generate this value, and so *N* is 32. As such the equation becomes:

$$\begin{aligned}\mathrm{SS_T} &= s^2_{\mathrm{grand}}(N-1)\\ &= 55.028(32-1)\\ &= 1705.868\end{aligned}$$

The degrees of freedom for this sum of squares, as with the independent ANOVA, will be *N* – 1, or 31.

The within-participant sum of squares (SS$_W$)

The crucial variation in this design is that there is a variance component called the within-participant variance (this arises because we've manipulated our independent variable within each participant). This is calculated using a sum of squares. Generally speaking, when we calculate any sum of squares we look at the squared difference between the mean and individual scores. This can be expressed in terms of the variance across a number of scores and the number of scores on which the variance is based. For example, when we calculated the residual sum of squares in independent ANOVA (SS$_R$) we used the following equation:

$$\mathrm{SS_R} = \sum (x_i - \bar{x}_i)^2$$

$$\mathrm{SS_R} = s^2(n-1)$$

This equation gave us the variance between individuals within a particular group, and so is an estimate of individual differences within a particular group. Therefore, to get the total value of individual differences we have to calculate the sum of squares within each group and then add them up:

$$\mathrm{SS_R} = s^2_{\text{group1}}(n_1-1) + s^2_{\text{group2}}(n_2-1) + s^2_{\text{group3}}(n_3-1)$$

This is all well and good when we have different people in each group, but in repeated-measures designs we've subjected people to more than one experimental condition, and therefore we're interested in the variation not within a group of people (as in independent ANOVA) but within an actual person. That is, how much variability is there within an individual? To find this out we actually use the same equation, but we adapt it to look at people rather than groups. So, if we call this sum of squares SS$_W$ (for within-participant SS) we could write it as:

$$\mathrm{SS_W} = s^2_{\text{person1}}(n_1-1) + s^2_{\text{person2}}(n_2-1) + s^2_{\text{person3}}(n_3-1) + \cdots + s^2_{\text{person}\,n}(n_n-1)$$

This equation simply means that were looking at the variation in an individual's scores and then adding these variances for all the people in the study. Some of you may have noticed that, in our example, we're using essays rather than people, and so to be pedantic we'd write this as:

$$\mathrm{SS_W} = s^2_{\text{essay1}}(n_1-1) + s^2_{\text{essay2}}(n_2-1) + s^2_{\text{essay3}}(n_3-1) \ldots + s^2_{\text{essay}\,n}(n_n-1)$$

The *n*s simply represent the number of scores on which the variances are based (i.e. the number of experimental conditions, or in this case the number of lecturers). All of the variances we need are in the table, so we can calculate SS$_W$ as:

$$\begin{aligned}
\mathrm{SS_W} &= s^2_{\text{essay1}}(n_1 - 1) + s^2_{\text{essay2}}(n_2 - 1) + s^2_{\text{essay3}}(n_3 - 1) + \cdots + s^2_{\text{essay}\,n}(n_n - 1) \\
&= (6.92)(4-1) + (11.33)(4-1) + (20.92)(4-1) + (18.25)(4-1) \\
&\quad + (43.58)(4-1) + (84.25)(4-1) + (132.92)(4-1) + (216)(4-1) \\
&= 20.76 + 34 + 62.75 + 54.75 + 130.75 + 252.75 + 398.75 + 648 \\
&= 1602.5
\end{aligned}$$

The degrees of freedom for each person are *n* – 1 (i.e. the number of conditions minus 1). To get the total degrees of freedom we add the *df* for all participants. So, with eight participants (essays) and four conditions (i.e. *n* = 4) we get 8 × 3 = 24 degrees of freedom.

The model sum of squares (SS$_M$)

So far, we know that the total amount of variation within the data is 1705.868 units. We also know that 1602.5 of those units are explained by the variance created by individuals' (essays') performances under different conditions. Now some of this variation is the result of our experimental manipulation, and some of this variation is simply random fluctuation. The next step is to work out how much variance is explained by our manipulation and how much is not.

In independent ANOVA, we worked out how much variation could be explained by our experiment (the model SS) by looking at the means for each group and comparing these to the overall mean. So, we measured the variance resulting from the differences between group means and the overall mean. We do exactly the same thing with a repeated-measures design. First we calculate the mean for each level of the independent variable (in this case the mean mark given by each lecturer) and compare these values to the overall mean of all marks. So, we calculate this SS in the same way as for independent ANOVA:

1. Calculate the difference between the mean of each group and the grand mean.
2. Square each of these differences.
3. Multiply each result by the number of subjects within that group (n_i).
4. Add the values for each group together:

$$\mathrm{SS_M} = \sum n_i\left(\bar{x}_i - \bar{x}_{\text{grand}}\right)^2$$

Using the means from the essay data, we can calculate SS$_M$ as follows:

$$
\begin{aligned}
\mathrm{SS_M} &= 8(68.875 - 63.9375)^2 + 8(64.25 - 63.9375)^2 + 8(65.25 - 63.9375)^2 \\
&\quad + 8(57.375 - 63.9375)^2 \\
&= 8(4.9375)^2 + 8(0.3125)^2 + 8(1.3125)^2 + 8(-6.5625)^2 \\
&= 554.125
\end{aligned}
$$

For $\mathrm{SS_M}$, the degrees of freedom (df_M) are again one less than the number of things used to calculate the sum of squares. For the model sums of squares we calculated the sum of squared errors between the four means and the grand mean. Hence, we used four things to calculate these sums of squares. So, the degrees of freedom will be 3. So, as with independent ANOVA, the model degrees of freedom is always the number of *groups* (*k*) minus 1:

$$df_\mathrm{M} = k - 1 = 3$$

The residual sum of squares ($\mathrm{SS_R}$)

We now know that there are 1706 units of variation to be explained in our data, and that the variation across our conditions accounts for 1602 units. Of these 1602 units, our experimental manipulation can explain 554 units. The final sum of squares is the residual sum of squares ($\mathrm{SS_R}$), which tells us how much of the variation cannot be explained by the model. This value is the amount of variation caused by extraneous factors outside of experimental control (such as natural variation in the quality of the essays). Knowing $\mathrm{SS_W}$ and $\mathrm{SS_M}$ already, the simplest way to calculate $\mathrm{SS_R}$ is to subtract $\mathrm{SS_M}$ from $\mathrm{SS_W}$:

$$
\begin{aligned}
\mathrm{SS_R} &= \mathrm{SS_W} - \mathrm{SS_M} \\
&= 1602.5 - 554.125 \\
&= 1048.375
\end{aligned}
$$

The degrees of freedom are calculated in a similar way:

$$
\begin{aligned}
df_\mathrm{R} &= df_\mathrm{W} - df_\mathrm{M} \\
&= 24 - 3 \\
&= 21
\end{aligned}
$$

The mean squares

$\mathrm{SS_M}$ tells us how much variation the model (e.g., the experimental manipulation) explains and $\mathrm{SS_R}$ tells us how much variation is due to extraneous factors. However, because both of these values are summed values the number of scores that were summed influences them. As with independent ANOVA, we eliminate this bias by calculating the average sum of squares (known as the *mean squares*, MS), which is simply the sum of squares divided by the degrees of freedom:

$$\mathrm{MS_M} = \frac{\mathrm{SS_M}}{df_\mathrm{M}} = \frac{554.125}{3} = 184.708$$

$$\mathrm{MS_R} = \frac{\mathrm{SS_R}}{df_\mathrm{R}} = \frac{1048.375}{21} = 49.923$$

$\mathrm{MS_M}$ represents the average amount of variation explained by the model (e.g. the systematic variation), whereas $\mathrm{MS_R}$ is a gauge of the average amount of variation explained by extraneous variables (the unsystematic variation).

The *F*-ratio

The *F*-ratio is a measure of the ratio of the variation explained by the model and the variation explained by unsystematic factors. It can be calculated by dividing the model mean squares by the residual mean squares. You should recall that this is exactly the same as for independent ANOVA:

$$F = \frac{\mathrm{MS_M}}{\mathrm{MS_R}}$$

So, as with the independent ANOVA, the *F*-ratio is still the ratio of systematic variation to unsystematic variation. As such, it is the ratio of the experimental effect to the effect on performance of unexplained factors. For the marking data, the *F*-ratio is:

$$F = \frac{\mathrm{MS_M}}{\mathrm{MS_R}} = \frac{184.708}{49.923} = 3.70$$

This value is greater than 1, which indicates that the experimental manipulation had some effect above and beyond the effect of extraneous factors. As with independent ANOVA this value can be compared against a critical value based on its degrees of freedom (which are df_M and df_R, which are 3 and 21 in this case).

Task 2

Repeat the analysis for Task 1 on SPSS and interpret the results.

Doing the analysis

To conduct an ANOVA using a repeated-measures design, activate the *define factors* dialog box by selecting Analyze General Linear Model ▸ Repeated Measures... . In the *define factors* dialog box (**Error! Reference source not found.**) you are asked to supply a name for the within-subject (repeated-measures) variable. In this case the repeated-measures variable was the tutor marking the essay, so replace the word *factor1* with the word TUTOR. Next, you have to tell SPSS how many levels there were (i.e., how many experimental conditions there were). In this case, there were four tutors, so enter the number 4 into the box labelled *Number of Levels*. Click on Add to add this variable to the list of repeated-measures variables. This variable will now appear in the white box at the bottom of the dialog box as *TUTOR(4)*. The finished dialog box is shown in the left-hand side of Figure . Next, click on Define to go to the main dialog box.

The main dialog box (see the right-hand side of Figure) has a space labelled *Within-Subjects Variables* that contains a list of four question marks followed by a number. These question marks are for the variables representing the four levels of the independent variable. The variables corresponding to these levels should be selected and placed in the appropriate space. We have only four variables in the data editor, so it is possible to select all four variables at once (by clicking on the variable at the top, pressing the *Shift* key and then clicking on the last variable that you want to select). The selected variables can then be dragged to the box labelled *Within-Subjects Variables* (or click on ➜). When all four variables have been transferred, you can select various options for the analysis. There are several options that can be accessed with the buttons at the side of the main dialog box. These options are similar to the ones we have already encountered.

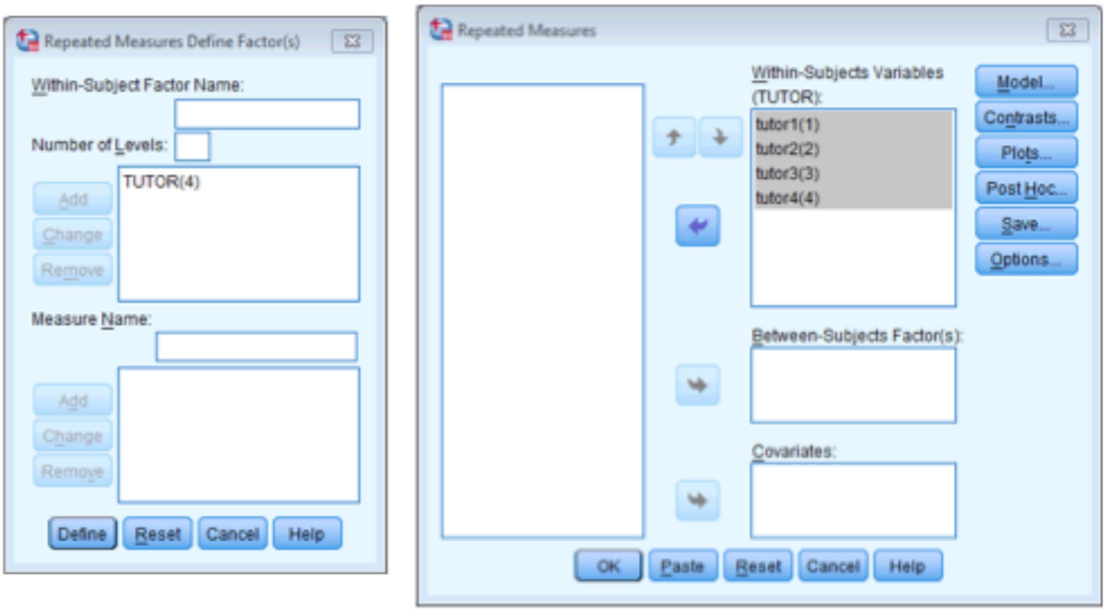
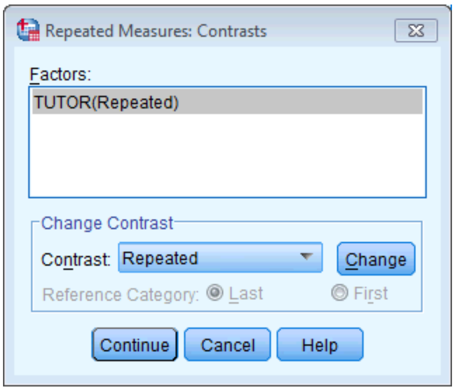


Figure 1

If you click on Contrasts... in the main dialog box you can access the *contrasts* dialog box (**Error! Reference source not found.**). There is no particularly good contrast for the data we have (the simple contrast is not very useful because we have no control category) so let's use the *repeated* contrast, which will compare each tutor's marks against the previous tutor. When you have selected this contrast, click on Continue to return to the main dialog box.

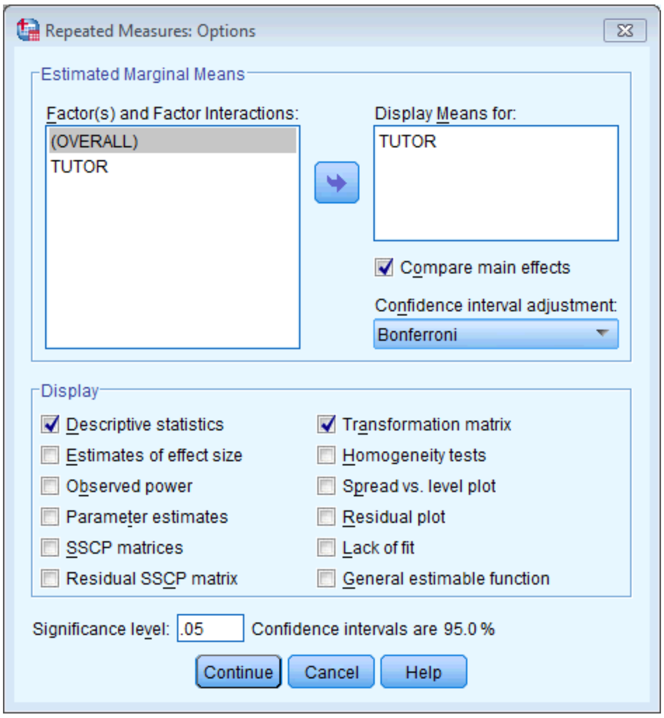

Repeated Measures: Contrasts
Change Contrast
Change
Reference Category: ○ Last ○ First
Continue
Cancel
Help

Factors: TUTOR(Repeated)

Contrast: Repeated

Figure 2

Clicking on Options... in the main dialog box will open the GLM *Repeated Measures: Options* dialog box (Figure). To specify *post hoc* tests, select the repeated-measures variable (in this case **TUTOR**) from the box labelled *Estimated Marginal Means: Factor(s) and Factor Interactions* and drag it to the box labelled *Display Means for* (or click on ➡). Once a variable has been transferred, you will be able to select ☑ Compare main effects. Once this option is selected, the box labelled *Confidence interval adjustment* becomes active and you can click on LSD(none) ▾ to see a choice of three adjustment levels. I am going to select the Bonferroni correction (see the book chapter). I am also going to ask for descriptive statistics and a transformation matrix. The transformation matrix provides the coding values for any contrast selected in the contrasts dialog box (Figure). When you have selected the options of interest, click on Continue to return to the main dialog box, and then click on OK to run the analysis.


Repeated Measures: Options

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
TUTOR

Display Means for:
TUTOR

☑ Compare main effects

Confidence interval adjustment: Bonferroni

Display

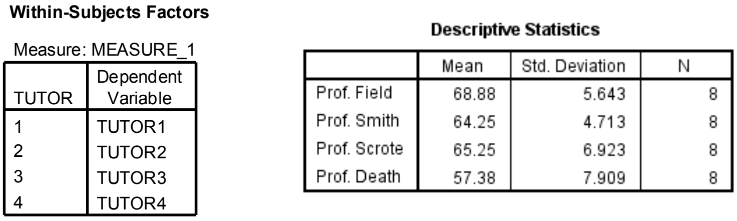
☑ Descriptive statistics	☑ Transformation matrix
□ Estimates of effect size	□ Homogeneity tests
□ Observed power	□ Spread vs. level plot
□ Parameter estimates	□ Residual plot
□ SSCP matrices	□ Lack of fit
□ Residual SSCP matrix	□ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Figure 3

Initial output for one-way repeated-measures ANOVA

First, we are told the variables that represent each level of the independent variable (Output 1). This box is useful to check that the variables were entered in the correct order. The next table provides basic descriptive statistics for the four levels of the independent variable. From this table we can see that, on average, Professor Field gave the highest marks to the essays (that's because I'm so nice, you see … or it could be because I'm stupid and so have low academic standards?). Professor Death, on the other hand, gave very low grades. These mean values are useful for interpreting any effects that may emerge from the main analysis.



Within-Subjects Factors

Measure: MEASURE_1

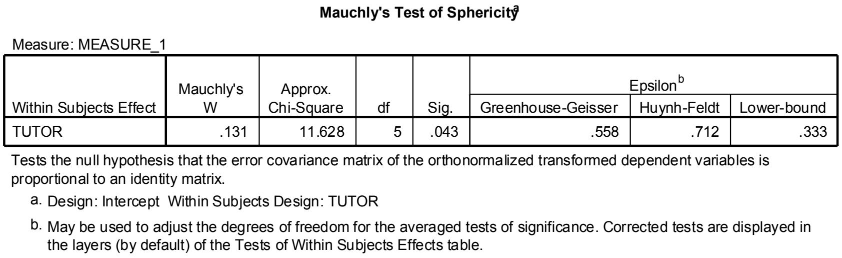
TUTOR	Dependent Variable
1	TUTOR1
2	TUTOR2
3	TUTOR3
4	TUTOR4

Descriptive Statistics

	Mean	Std. Deviation	N
Prof. Field	68.88	5.643	8
Prof. Smith	64.25	4.713	8
Prof. Scrote	65.25	6.923	8
Prof. Death	57.38	7.909	8

Output 1

Output 2 contains information about Mauchly's test. This test should be non-significant if we are to assume that the condition of sphericity has been met. The output shows Mauchly's test for the tutor data, and the important column is the one containing the significance value. The significance value (.043) is less than the critical value of .05, so we accept that the variances of the differences between levels are significantly different. In other words, the assumption of sphericity has been violated. Knowing that we have violated this assumption, a pertinent question is: how should we proceed?



Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TUTOR	.131	11.628	5	.043	.558	.712	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept Within Subjects Design: TUTOR

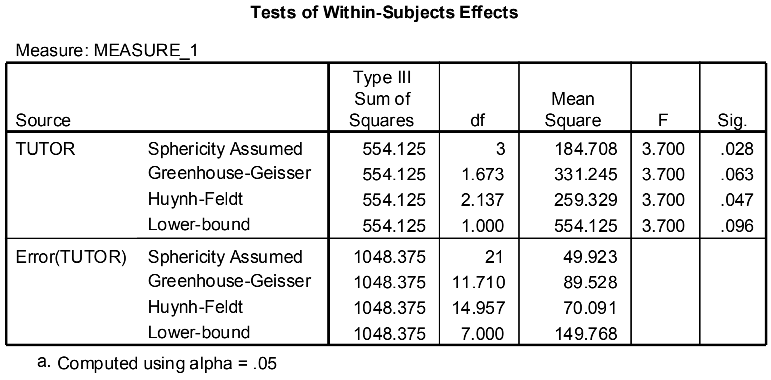
b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the layers (by default) of the Tests of Within Subjects Effects table.

Output 2

SPSS produces three corrections based upon the estimates of sphericity advocated by Greenhouse and Geisser (1959) and Huynh and Feldt (1976). Both of these estimates give rise to a correction factor that is applied to the degrees of freedom used to assess the observed *F*-ratio. The *Greenhouse–Geisser correction* varies between 1/(*k*–1) (where *k* is the number of repeated-measures conditions) and 1. The closer that $\hat{\varepsilon}$ is to 1.00, the more homogeneous the variances of differences, and hence the closer the data are to being spherical. In a situation in which there are four conditions (as with our data) the lower limit of $\hat{\varepsilon}$ will be 1/(4–1), or 0.33 (known as the lower-bound estimate of sphericity). The calculated value of $\hat{\varepsilon}$ in the output is .558. This is closer to the lower limit of 0.33 than it is to the upper limit of 1, and it therefore represents a substantial deviation from sphericity. We will see how these values are used in the next section.

The main ANOVA

Output 3 shows the results of the ANOVA for the within-subjects variable. This table can be read much the same as for one-way between-group ANOVA. There is a sum of squares for the repeated-measures effect of **tutor**, which tells us how much of the total variability is explained by the experimental effect. Note the value is 554.125, which is the model sum of squares (SS_M) that we calculated in Task 1. There is also an error term, which is the amount of unexplained variation across the conditions of the repeated-measures variable. This is the residual sum of squares (SS_R) that was calculated earlier, and note the value is 1048.375 (which is the same value as calculated). As I explained earlier, these sums of squares are converted into mean squares by dividing by the degrees of freedom. As we saw before, the *df* for the effect of **tutor** are simply $k - 1$, where k is the number of levels of the independent variable. The error *df* are $(n - 1)(k - 1)$, where n is the number of participants (or in this case, the number of essays) and k is as before. The *F*-ratio is obtained by dividing the mean squares for the experimental effect (184.708) by the error mean squares (49.923). As with between-group ANOVA, this test statistic represents the ratio of systematic variance to unsystematic variance. The value of *F* (184.71/49.92 = 3.70) is then compared against a critical value for 3 and 21 degrees of freedom. SPSS displays the exact significance level for the *F*-ratio. The significance of *F* is .028, which is significant because it is less than the criterion value of .05. We can, therefore, conclude that there was a significant difference between the marks awarded by the four lecturers. However, this main test does not tell us which lecturers differed from each other in their marking.



Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TUTOR	Sphericity Assumed	554.125	3	184.708	3.700	.028
	Greenhouse-Geisser	554.125	1.673	331.245	3.700	.063
	Huynh-Feldt	554.125	2.137	259.329	3.700	.047
	Lower-bound	554.125	1.000	554.125	3.700	.096
Error(TUTOR)	Sphericity Assumed	1048.375	21	49.923		
	Greenhouse-Geisser	1048.375	11.710	89.528		
	Huynh-Feldt	1048.375	14.957	70.091		
	Lower-bound	1048.375	7.000	149.768		

a. Computed using alpha = .05

Output 3

Although this result seems very plausible, we have learnt that the violation of the sphericity assumption makes the *F*-test inaccurate. We know from Mauchly's test that these data were non-spherical and so we need to make allowances for this violation. The SPSS output shows the *F*-ratio and associated degrees of freedom when sphericity is assumed, and the significant *F*-statistic indicated some difference(s) between the mean marks given by the four lecturers. In versions of SPSS after version 8, this table also contains several additional rows giving the corrected values of *F* for the three different types of adjustment (Greenhouse–Geisser, Huynh–Feldt and lower-bound).

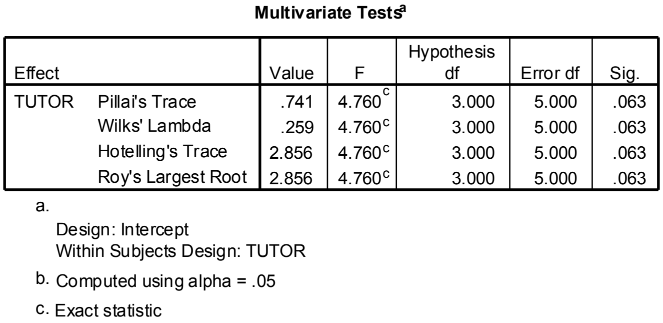
Notice that in all cases the *F*-ratios remain the same; it is the degrees of freedom that change (and hence the critical value against which the *F*-statistic is compared). The degrees of freedom have been adjusted using the estimates of sphericity calculated by SPSS. The adjustment is made by multiplying the degrees of freedom by the estimate of sphericity.[1] The new degrees of freedom are then used to ascertain the significance of *F*. For these data the corrections result in the observed *F* being non-significant when using the Greenhouse–Geisser correction (because $p > .05$). However, it was noted earlier that this correction is quite conservative, and so can miss effects that genuinely exist. It is, therefore, useful to consult the Huynh–Feldt-corrected *F*-statistic. Using this correction, the *F*-value is still significant because the probability value of .047 is just below the criterion value of .05. So, by this correction we would accept the hypothesis that the lecturers differed in their marking. However, it was also noted earlier that this correction is quite liberal and so tends to accept values as significant when, in reality, they are not significant. This leaves us with the puzzling dilemma of whether or not to accept this *F*-statistic as significant. I mentioned earlier that Stevens (2002) recommends taking an average of the two estimates, and certainly when the two corrections give different results (as is the case here) this is wise advice. If the two corrections give rise to the same conclusion it makes little difference which you choose to report (although if you accept the *F*-statistic as significant it is best to report the conservative Greenhouse–Geisser estimate to avoid criticism!). Although it is easy to calculate the average of the two correction factors and to correct the degrees of freedom accordingly, it is not so easy to then calculate an exact probability for those degrees of freedom. Therefore, should you ever be faced with this perplexing situation (and, to be honest, that's fairly unlikely) I recommend taking an average of the two significance values to give you a rough idea of which correction is giving the most accurate answer. In this case, the average of the two *p*-values is (.063 + .047)/2 = .055.

[1] For example, the Greenhouse–Geisser estimate of sphericity was .558. The original degrees of freedom for the model were 3; this value is corrected by multiplying by the estimate of sphericity (3 × 0.558 = 1.674). Likewise the error *df* were 21; this value is corrected in the same way (21 × 0.558 = 11.718). The *F*-ratio is then tested against a critical value with these new degrees of freedom (1.674, 11.718). The other corrections are applied in the same way.

Therefore, we should probably go with the Greenhouse–Geisser correction and conclude that the *F*-ratio is non-significant.

These data illustrate how important it is to use a valid critical value of *F*: it can mean the difference between a statistically significant result and a non-significant result. More important, it can mean the difference between making a Type I error and not. Had we not used the corrections for sphericity we would have concluded erroneously that the markers gave significantly different marks. However, I should quantify this statement by saying that this example also highlights how arbitrary it is that we use a .05 level of significance. These two corrections produce significance values only marginally less than or more than .05, and yet they lead to completely opposite conclusions! So, we might be well advised to look at an effect size to see whether the effect is substantive regardless of its significance.

We also saw earlier that a final option, when you have data that violate sphericity, is to use multivariate test statistics (MANOVA) because they do not make this assumption (see O'Brien & Kaiser, 1985). The repeated-measures procedure in SPSS automatically produces multivariate test statistics. Output 4 shows the multivariate test statistics for this example. The column displaying the significance values clearly shows that the multivariate tests are non-significant (because *p* is .063, which is greater than the criterion value of .05). Bearing in mind the loss of power in these tests, this result supports the decision to accept the null hypothesis and conclude that there are no significant differences between the marks given by different lecturers. The interpretation of these results should stop now because the main effect is non-significant. However, we will look at the output for contrasts to illustrate how these tests are displayed in the SPSS Viewer.



Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
TUTOR	Pillai's Trace	.741	4.760[c]	3.000	5.000	.063
	Wilks' Lambda	.259	4.760[c]	3.000	5.000	.063
	Hotelling's Trace	2.856	4.760[c]	3.000	5.000	.063
	Roy's Largest Root	2.856	4.760[c]	3.000	5.000	.063

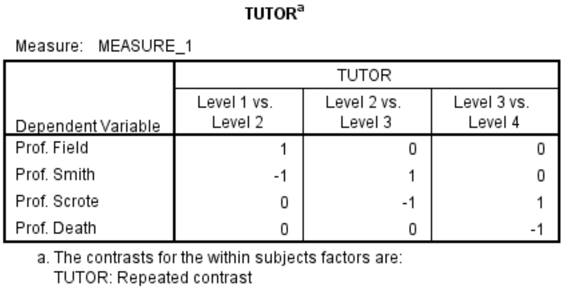
a.
Design: Intercept
Within Subjects Design: TUTOR

b. Computed using alpha = .05

c. Exact statistic

Output 4

The transformation matrix requested in the options is shown in Output 5, and we have to draw on our knowledge of contrast coding to interpret this table. The first thing to remember is that a code of 0 means that the group is not included in a contrast. Therefore, contrast 1 (labelled *Level 1 vs. Level 2* in the table) ignores Prof Scrote and Prof Death. The next thing to remember is that groups with a negative weight are compared to groups with a positive weight. In this case this means that the first contrast compares Prof Field against Prof Smith. Using the same logic, contrast 2 (labelled *Level 2 vs. Level 3*) ignores Prof Field and Prof Death and compares Prof Smith and Prof Scrote. Finally, contrast 3 (*Level 3 vs. Level 4*) compares Prof Death with Prof Scrote. This pattern of contrasts is consistent with what we expect to get from a repeated contrast (i.e. all groups except the first are compared to the preceding category). The transformation matrix, which appears at the bottom of the output, is used primarily to confirm what each contrast represents.



TUTOR[a]

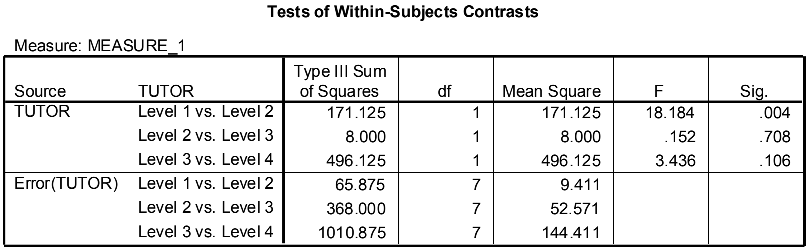
Measure: MEASURE_1

	TUTOR		
Dependent Variable	Level 1 vs. Level 2	Level 2 vs. Level 3	Level 3 vs. Level 4
Prof. Field	1	0	0
Prof. Smith	-1	1	0
Prof. Scrote	0	-1	1
Prof. Death	0	0	-1

a. The contrasts for the within subjects factors are:
TUTOR: Repeated contrast

Output 5

Above the transformation matrix, we should find a summary table of the contrasts (Output 6). Each contrast is listed in turn, and as with between-group contrasts, an *F*-test is performed that compares the two chunks of variation. So, looking at the significance values from the table, we could say that Prof Field marked significantly more highly than Prof Smith (*Level 1 vs. Level 2*), but that Prof Smith's marks were roughly equal to Prof Scrote's (*Level 2 vs. Level 3*) and Prof Scrote's marks were roughly equal to Prof Death's (*Level 3 vs. Level 4*). However, the significant contrast should be ignored because of the non-significant main effect (remember that the data did not obey sphericity). The important point to note is that the sphericity in our data has led to some important issues being raised about correction factors, and about applying discretion to your data (it's comforting to know that the computer does not have all of the answers, but it's slightly alarming to realize that this means we have to actually know some of the answers ourselves). In this example we would have to conclude that no significant differences existed between the marks given by different lecturers. However, the ambiguity of our data might make us consider running a similar study with a greater number of essays being marked.



Tests of Within-Subjects Contrasts

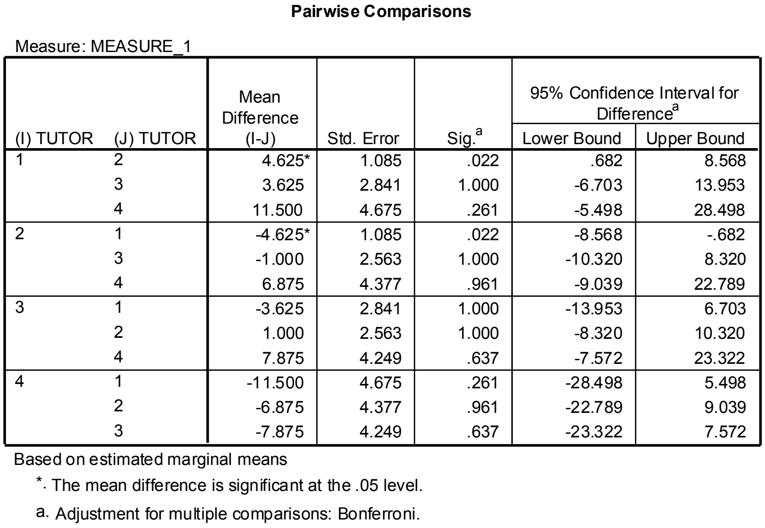
Measure: MEASURE_1

Source	TUTOR	Type III Sum of Squares	df	Mean Square	F	Sig.
TUTOR	Level 1 vs. Level 2	171.125	1	171.125	18.184	.004
	Level 2 vs. Level 3	8.000	1	8.000	.152	.708
	Level 3 vs. Level 4	496.125	1	496.125	3.436	.106
Error(TUTOR)	Level 1 vs. Level 2	65.875	7	9.411		
	Level 2 vs. Level 3	368.000	7	52.571		
	Level 3 vs. Level 4	1010.875	7	144.411		

Output 6

Post hoc tests

If you selected *post hoc* tests for the repeated-measures variable in the options dialog box, then Output 7 will be produced in the SPSS Viewer.



Pairwise Comparisons

Measure: MEASURE_1

					95% Confidence Interval for Difference[a]	
(I) TUTOR	(J) TUTOR	Mean Difference (I-J)	Std. Error	Sig.[a]	Lower Bound	Upper Bound
1	2	4.625*	1.085	.022	.682	8.568
	3	3.625	2.841	1.000	-6.703	13.953
	4	11.500	4.675	.261	-5.498	28.498
2	1	-4.625*	1.085	.022	-8.568	-.682
	3	-1.000	2.563	1.000	-10.320	8.320
	4	6.875	4.377	.961	-9.039	22.789
3	1	-3.625	2.841	1.000	-13.953	6.703
	2	1.000	2.563	1.000	-8.320	10.320
	4	7.875	4.249	.637	-7.572	23.322
4	1	-11.500	4.675	.261	-28.498	5.498
	2	-6.875	4.377	.961	-22.789	9.039
	3	-7.875	4.249	.637	-23.322	7.572

Based on estimated marginal means

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

a. Adjustment for multiple comparisons: Bonferroni.

Output 7

The difference between group means is displayed and also the standard error, the significance value and a confidence interval for the difference between means. By looking at the significance values we can see that the only difference between group means is between Prof Field and Prof Smith. Looking at the means of these groups, we can see that I give significantly higher marks than Prof Smith. However, there is a rather anomalous result in that there is no significant difference between the marks given by Prof Death and myself, even though the mean difference between our marks is higher (11.5) than the mean difference between myself and Prof Smith (4.6). The reason for this result is the sphericity in the data. The interested reader might like to run some correlations between the four tutors' grades. You will find that there is a very high positive correlation between the marks given by Prof Smith and myself (indicating a low level of variability in our data). However, there is a very low correlation between the marks given by Prof Death and myself (indicating a high level of variability between our marks). It is this large variability between Prof Death and myself that has produced the non-significant result despite the average marks being very different (this observation is also evident from the standard errors).

Task 3

Calculate the effect sizes for the analysis in Task 1.

In repeated-measures ANOVA, the equation for ω^2 is:

$$\omega^2 = \frac{\left[\frac{k-1}{nk}(\mathrm{MS_M} - \mathrm{MS_R})\right]}{\mathrm{MS_R} + \frac{\mathrm{MS_B} - \mathrm{MS_R}}{k} + \left[\frac{k-1}{nk}(\mathrm{MS_M} - \mathrm{MS_R})\right]}$$

SPSS doesn't give us $\mathrm{SS_W}$ in the output, but we know that this is made up of $\mathrm{SS_M}$ and $\mathrm{SS_R}$, which we are given. By substituting these terms, and rearranging the equation, we get:

$$\mathrm{SS_T} = \mathrm{SS_B} + \mathrm{SS_M} + \mathrm{SS_R}$$
$$\mathrm{SS_B} = \mathrm{SS_T} - \mathrm{SS_M} - \mathrm{SS_R}$$

The next problem is that SPSS, which is clearly trying to hinder us at every step, doesn't give us $\mathrm{SS_T}$ and, I'm afraid (unless I've missed something in the output), you're just going to have to calculate it by hand. From the values we calculated earlier, you should get:

$$\begin{aligned}\mathrm{SS_B} &= 1705.868 - 554.125 - 1048.375 \\ &= 103.37\end{aligned}$$

The next step is to convert this to a mean squares by dividing by the degrees of freedom, which in this case are the number of people in the experiment minus 1 (n – 1):

$$\begin{aligned}\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{df_B}} &= \frac{\mathrm{SS_B}}{N-1} \\ &= \frac{103.37}{8-1} \\ &= 14.77\end{aligned}$$

Having done all this and probably died of boredom in the process we must now resurrect ourselves with renewed vigour for the effect size equation, which becomes:

$$
\begin{aligned}
\omega^2 &= \frac{\left[\frac{4-1}{8\times 4}(184.71-49.92)\right]}{49.92+\frac{14.77-49.92}{4}+\left[\frac{4-1}{8\times 4}(184.71-49.92)\right]} \\
&= \frac{12.64}{53.77} \\
&= .24
\end{aligned}
$$

So, we get ω^2 = .24. If you calculate it the same way as for the independent ANOVA you should get a slightly bigger answer (.25 in fact).

I've mentioned at various other points that it's actually more useful to have effect size measures for focused comparisons anyway (rather than the main ANOVA), and so a slightly easier approach to calculating effect sizes is to calculate them for the contrasts we did. For these we can use the equation that we've seen before to convert the *F*-values (because they all have 1 degree of freedom for the model) to *r*:

$$
r = \sqrt{\frac{F(1, \mathrm{df_R})}{F(1, \mathrm{df_R}) + \mathrm{df_R}}}
$$

For the three comparisons we did, we would get:

$$
r_{\text{Field vs. Smith}} = \sqrt{\frac{18.18}{18.18+7}} = .85
$$

$$
r_{\text{Smith vs. Scrote}} = \sqrt{\frac{0.15}{0.15+7}} = .14
$$

$$
r_{\text{Scrote vs. Death}} = \sqrt{\frac{3.44}{3.44+7}} = .57
$$

Therefore, the differences between Profs Field and Smith and between Scrote and Death were both large effects, but the differences between Profs. Smith and Scrote were small.

Reporting one-way repeated-measures ANOVA

We could report the main finding as follows:

- The results show that the mark of an essay was not significantly affected by the lecturer who marked it, *F*(1.67, 11.71) = 3.70, *p* = .063.

If you choose to report the sphericity test as well, you should report the chi-square approximation, its degrees of freedom and the significance value. It's also nice to report the degree of sphericity by reporting the epsilon value. We'll also report the effect size in this improved version:

- Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(5)$ = 11.63, p < .05, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity (ε = .56). The results show that the mark of an essay was not significantly affected by the lecturer who marked it, F(1.67, 11.71) = 3.70, p = .063, ω^2 = .24.

Remember that because the main ANOVA was not significant we shouldn't report any further analysis.

Task 4

The 'roving eye' effect is the propensity of people in relationships to 'eye up' members of the opposite sex. I took 20 men and fitted them with incredibly sophisticated glasses that could track their eye movements and record both the movement and the object being observed (this is the point at which it should be apparent that I'm making it up as I go along). Over four different nights I plied these poor souls with 1, 2, 3 or 4 pints of strong lager in a nightclub. Each night I measured how many different women they eyed up (a woman was categorized as having been eyed up if the man's eye moved from her head to her toe and back up again). The data are in the file ***RovingEye.sav****. Analyse them with a one-way ANOVA.*

SPSS output

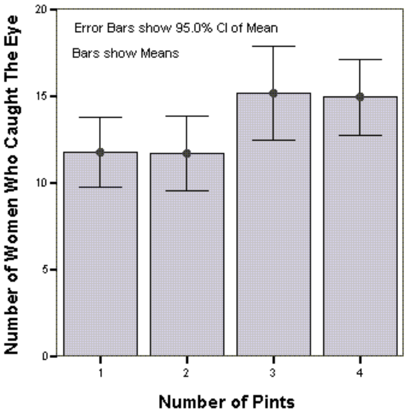
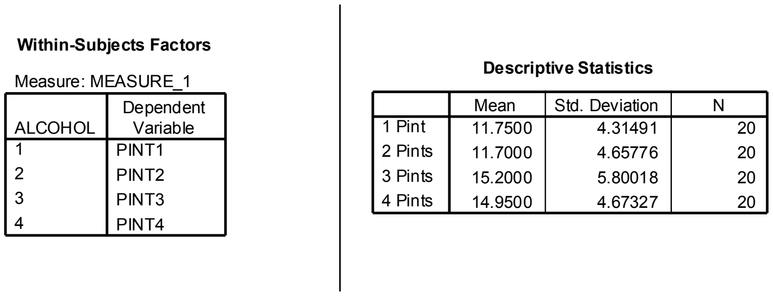


Figure 4

The error bar chart of the roving eye data in Figure 4 shows the mean number of women who were eyed up after different doses of alcohol. It's clear from this chart that the mean number of women is pretty similar between 1 and 2 pints, and for 3 and 4 pints, but there is a jump after 2 pints.



Within-Subjects Factors

Measure: MEASURE_1

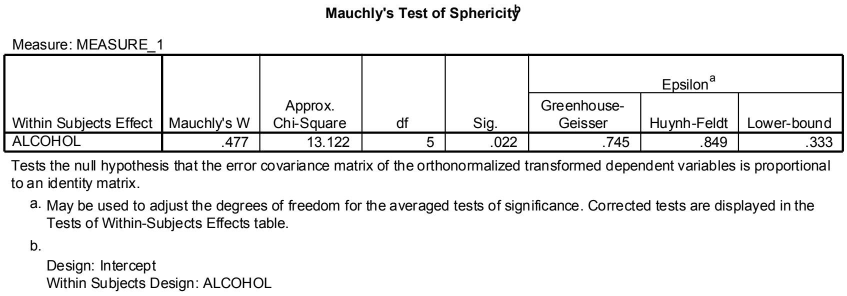
ALCOHOL	Dependent Variable
1	PINT1
2	PINT2
3	PINT3
4	PINT4

Descriptive Statistics

	Mean	Std. Deviation	N
1 Pint	11.7500	4.31491	20
2 Pints	11.7000	4.65776	20
3 Pints	15.2000	5.80018	20
4 Pints	14.9500	4.67327	20

Output 8

Output 8 shows the initial diagnostic statistics. First, we are told the variables that represent each level of the independent variable. This box is useful to check that the variables were entered in the correct order. The next table provides basic descriptive statistics for the four levels of the independent variable. This table confirms what we saw in Figure 4.



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
ALCOHOL	.477	13.122	5	.022	.745	.849	.333

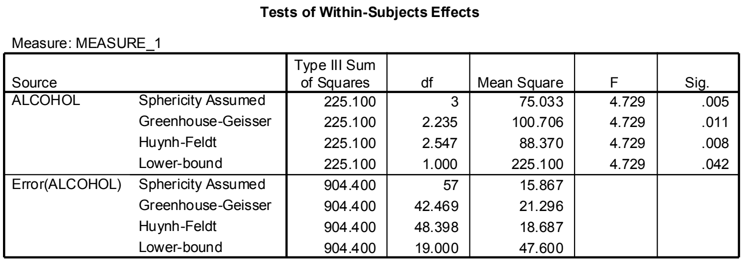
Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: ALCOHOL

Output 9

Output 9 contains Mauchly's test, and we hope to find that it's non-significant if we are to assume that the condition of sphericity has been met. However, the significance value (.022) is less than the critical value of .05, so we accept that the assumption of sphericity has been violated.



Tests of Within-Subjects Effects

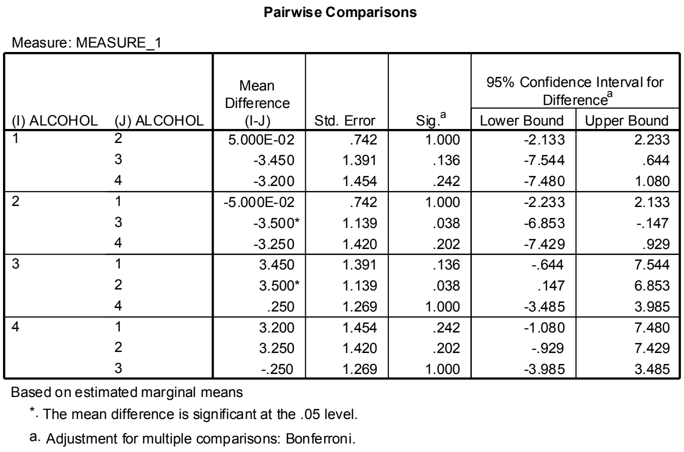
Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
ALCOHOL	Sphericity Assumed	225.100	3	75.033	4.729	.005
	Greenhouse-Geisser	225.100	2.235	100.706	4.729	.011
	Huynh-Feldt	225.100	2.547	88.370	4.729	.008
	Lower-bound	225.100	1.000	225.100	4.729	.042
Error(ALCOHOL)	Sphericity Assumed	904.400	57	15.867		
	Greenhouse-Geisser	904.400	42.469	21.296		
	Huynh-Feldt	904.400	48.398	18.687		
	Lower-bound	904.400	19.000	47.600		

Output 1

Output 10 shows the main result of the ANOVA. The significance of *F* is .005, which is significant because it is less than the criterion value of .05. We can, therefore, conclude that alcohol had a significant effect on the average number of women that were eyed up. However, this main test does not tell us which quantities of alcohol made a difference to the number of women eyed up.

This result is all very nice, but as of yet we haven't done anything about our violation of the sphericity assumption. This table contains several additional rows giving the corrected values of *F* for the three different types of adjustment (Greenhouse–Geisser, Huynh–Feldt and lower-bound). First we decide which correction to apply, and to do this we need to look at the estimates of sphericity: if the Greenhouse–Geisser and Huynh–Feldt estimates are less than .75 we should use Greenhouse–Geisser, and if they are above .75 we use Huynh–Feldt. We discovered in the book that based on these criteria we should use Huynh–Feldt here. Using this corrected value we still find a significant result because the observed *p* (.008) is still less than the criterion of .05.



Pairwise Comparisons

Measure: MEASURE_1

(I) ALCOHOL	(J) ALCOHOL	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	5.000E-02	.742	1.000	-2.133	2.233
	3	-3.450	1.391	.136	-7.544	.644
	4	-3.200	1.454	.242	-7.480	1.080
2	1	-5.000E-02	.742	1.000	-2.233	2.133
	3	-3.500*	1.139	.038	-6.853	-.147
	4	-3.250	1.420	.202	-7.429	.929
3	1	3.450	1.391	.136	-.644	7.544
	2	3.500*	1.139	.038	.147	6.853
	4	.250	1.269	1.000	-3.485	3.985
4	1	3.200	1.454	.242	-1.080	7.480
	2	3.250	1.420	.202	-.929	7.429
	3	-.250	1.269	1.000	-3.985	3.485

Based on estimated marginal means

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

a. Adjustment for multiple comparisons: Bonferroni.

Output 2

The main effect of alcohol doesn't tell us anything about which doses of alcohol produced different results to other doses. So, we might do some *post hoc* tests as well. Output 11 shows the table from SPSS that contains these tests. We read down the column labelled *Sig.* and look for values less than .05. By looking at the significance values, we can see that the only difference between condition means is between 2 and 3 pints of alcohol.

Interpreting and writing the result

We could report the main finding as follows:

✓ Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(5)$ = 13.12, *p* = .022, therefore degrees of freedom were corrected using Huynh–Feldt estimates of sphericity (*ε* = .85). The results show that the number of women eyed up was significantly affected by the amount of alcohol drunk, *F*(2.55, 48.40) = 4.73, *p* = .008, *r* = .40. Bonferroni *post hoc* tests revealed a significant difference in the number of women eyed up only between 2 and 3 pints, 95% CI (–6.85, –0.15), *p* = .038, but not between 1 and 2 pints (*p* = 1.00), 1 and 3 pints (*p* = .136), 1 and 4 pints (*p* = .242), 2 and 4 pints (*p* = .202) or 3 and 4 pints (*p* = 1.00).

Task 5

In the previous chapter we came across the beer-goggles effect, a severe perceptual distortion occurring after imbibing several pints of alcohol that makes previously unattractive people suddenly become the hottest thing since Spicy Gonzalez's extra-hot Tabasco-marinated chillies. In short, one minute you're standing in a zoo admiring the orang-utans, and the next you're wondering why someone would put the adorable Zoë Field in a cage. Anyway, in that chapter, we demonstrated that the beer-goggles effect was stronger for men than for women, and took effect only after 2 pints. Imagine we followed this finding up. We took a sample of 26 men (because the effect is stronger in men) and gave them various doses of ***Alcohol*** *over four different weeks (0 pints, 2 pints, 4 pints and 6 pints of lager). Each week (and, therefore, in each state of drunkenness) participants were asked to select a mate in a normal club (that had dim lighting) and then select a second mate in a specially designed club that had bright lighting. The second independent variable was whether the club had dim or bright lighting. The outcome measure was the attractiveness of each mate as assessed by a panel of independent judges. The data are in the file* ***BeerGogglesLighting.sav****. Analyse them with a two-way repeated-measures ANOVA.*

SPSS output

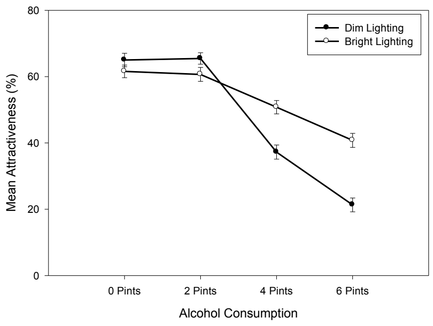
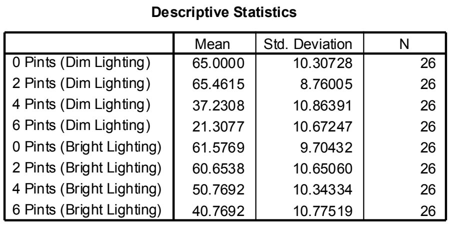


Figure 5

Figure 5 displays the mean attractiveness of the partner selected (with error bars) in dim and brightly lit clubs after the different doses of alcohol. The chart shows that in both dim and brightly lit clubs there is a tendency for men to select less attractive mates as they consume more and more alcohol.

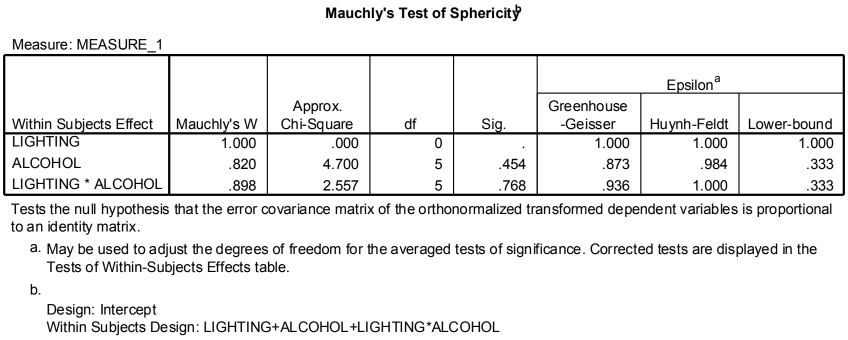


Descriptive Statistics

	Mean	Std. Deviation	N
0 Pints (Dim Lighting)	65.0000	10.30728	26
2 Pints (Dim Lighting)	65.4615	8.76005	26
4 Pints (Dim Lighting)	37.2308	10.86391	26
6 Pints (Dim Lighting)	21.3077	10.67247	26
0 Pints (Bright Lighting)	61.5769	9.70432	26
2 Pints (Bright Lighting)	60.6538	10.65060	26
4 Pints (Bright Lighting)	50.7692	10.34334	26
6 Pints (Bright Lighting)	40.7692	10.77519	26

Output 3

Output 12 the means for all conditions in a table. These means correspond to those plotted in the graph.



Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

					Epsilon[a]		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse -Geisser	Huynh-Feldt	Lower-bound
LIGHTING	1.000	.000	0	.	1.000	1.000	1.000
ALCOHOL	.820	4.700	5	.454	.873	.984	.333
LIGHTING * ALCOHOL	.898	2.557	5	.768	.936	1.000	.333

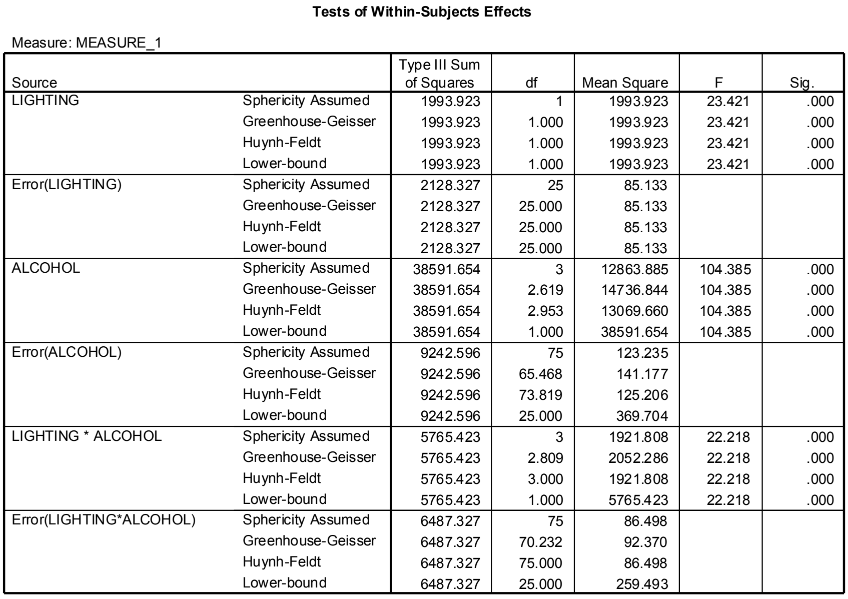
Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: LIGHTING+ALCOHOL+LIGHTING*ALCOHOL

Output 4

The lighting variable had only two levels (dim or bright), and so the assumption of sphericity doesn't apply and SPSS doesn't produce a significance value (Output 13). However, for the effects of alcohol consumption and the interaction of alcohol consumption and lighting, we do have to look at Mauchly's test. The significance values are both above .05 (they are .454 and .768, respectively), and so we know that the assumption of sphericity has been met for both alcohol consumption and the interaction of alcohol consumption and lighting.



Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
LIGHTING	Sphericity Assumed	1993.923	1	1993.923	23.421	.000
	Greenhouse-Geisser	1993.923	1.000	1993.923	23.421	.000
	Huynh-Feldt	1993.923	1.000	1993.923	23.421	.000
	Lower-bound	1993.923	1.000	1993.923	23.421	.000
Error(LIGHTING)	Sphericity Assumed	2128.327	25	85.133		
	Greenhouse-Geisser	2128.327	25.000	85.133		
	Huynh-Feldt	2128.327	25.000	85.133		
	Lower-bound	2128.327	25.000	85.133		
ALCOHOL	Sphericity Assumed	38591.654	3	12863.885	104.385	.000
	Greenhouse-Geisser	38591.654	2.619	14736.844	104.385	.000
	Huynh-Feldt	38591.654	2.953	13069.660	104.385	.000
	Lower-bound	38591.654	1.000	38591.654	104.385	.000
Error(ALCOHOL)	Sphericity Assumed	9242.596	75	123.235		
	Greenhouse-Geisser	9242.596	65.468	141.177		
	Huynh-Feldt	9242.596	73.819	125.206		
	Lower-bound	9242.596	25.000	369.704		
LIGHTING * ALCOHOL	Sphericity Assumed	5765.423	3	1921.808	22.218	.000
	Greenhouse-Geisser	5765.423	2.809	2052.286	22.218	.000
	Huynh-Feldt	5765.423	3.000	1921.808	22.218	.000
	Lower-bound	5765.423	1.000	5765.423	22.218	.000
Error(LIGHTING*ALCOHOL)	Sphericity Assumed	6487.327	75	86.498		
	Greenhouse-Geisser	6487.327	70.232	92.370		
	Huynh-Feldt	6487.327	75.000	86.498		
	Lower-bound	6487.327	25.000	259.493		

Output 5

Output 14 shows the main ANOVA summary table. The main effect of lighting is shown by the *F*-ratio in the row labelled **lighting**. The significance of this value is reported as .000 (i.e., *p* < .001), which is well below the usual cut-off point of .05. We can conclude that average attractiveness ratings were significantly affected by whether mates were selected in a dim or well-lit club. We can easily interpret this result further because there were only two levels: attractiveness ratings were higher in the well-lit clubs, so we could conclude that when we ignore how much alcohol was consumed, the mates selected in well-lit clubs were significantly more attractive than those chosen in dim clubs.

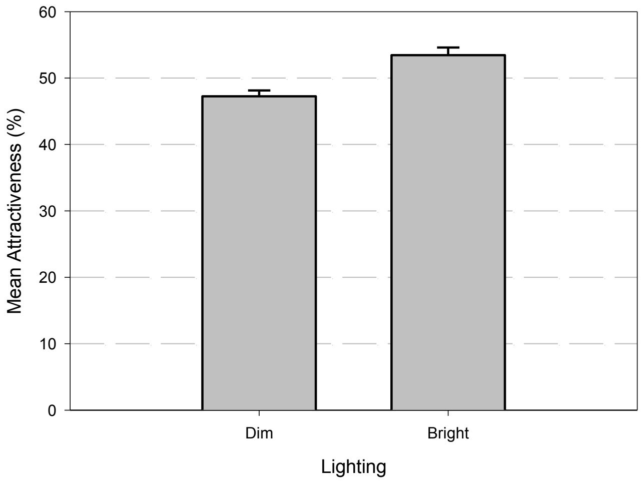


Figure 6

The main effect of alcohol consumption is shown by the *F*-ratio in the row labelled **alcohol**. The probability associated with this *F*-ratio is reported as .000, which is well below the critical value of .05. We can conclude that there was a significant main effect of the amount of alcohol consumed on the attractiveness of the mate selected. We know that generally there was an effect, but without further tests (e.g., *post hoc* comparisons) we can't say exactly which doses of alcohol had the most effect. I've plotted the means for the four doses in Figure 7. This graph shows that when you ignore the lighting in the club, the attractiveness of mates is similar after no alcohol and 2 pints of lager but starts to rapidly decline at 4 pints and continues to decline after 6 pints.

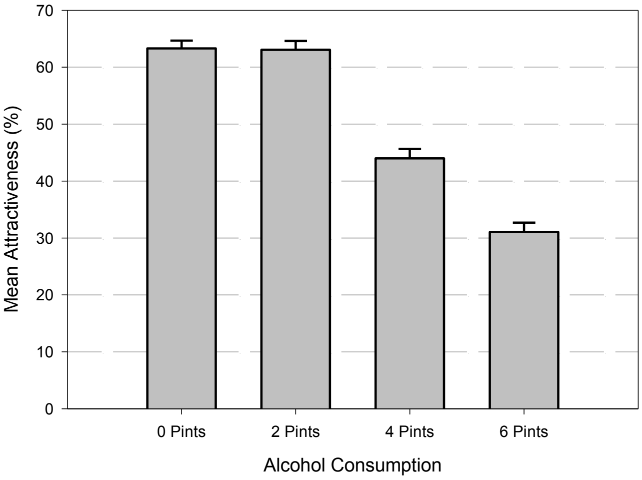
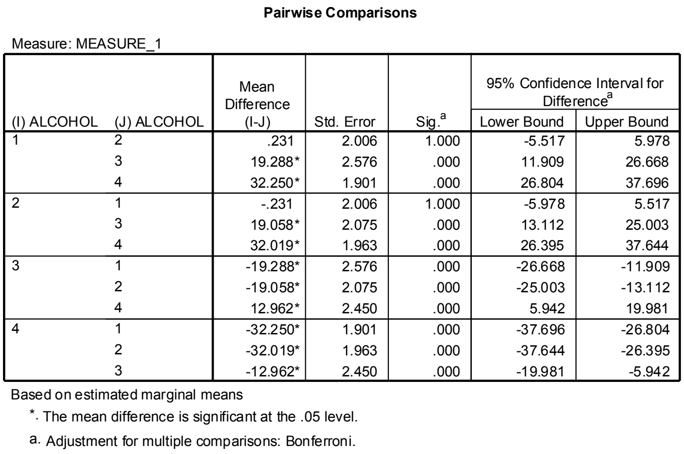


Figure 7

Output 15 shows some *post hoc* tests for the main effect of alcohol. In this example I've chosen a Bonferroni correction. The main column of interest is the one labelled *Sig.*, but the confidence intervals also tell us the likely difference between means if we were to take other samples. The mean attractiveness was significantly higher after no pints than it was after 4 pints and 6 pints (both *p*s are less than .001). We can also see that the mean attractiveness after 2 pints was significantly higher than after 4 pints and 6 pints (again, both *p*s are less than .001). Finally, the mean attractiveness after 4 pints was significantly higher than after 6 pints (*p* is less than .001). So, we can conclude that the beer goggles effect doesn't kick in until after 2 pints, and that it has an ever-increasing effect (well, up to 6 pints at any rate!).



Pairwise Comparisons

Measure: MEASURE_1

(I) ALCOHOL	(J) ALCOHOL	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.231	2.006	1.000	-5.517	5.978
	3	19.288*	2.576	.000	11.909	26.668
	4	32.250*	1.901	.000	26.804	37.696
2	1	-.231	2.006	1.000	-5.978	5.517
	3	19.058*	2.075	.000	13.112	25.003
	4	32.019*	1.963	.000	26.395	37.644
3	1	-19.288*	2.576	.000	-26.668	-11.909
	2	-19.058*	2.075	.000	-25.003	-13.112
	4	12.962*	2.450	.000	5.942	19.981
4	1	-32.250*	1.901	.000	-37.696	-26.804
	2	-32.019*	1.963	.000	-37.644	-26.395
	3	-12.962*	2.450	.000	-19.981	-5.942

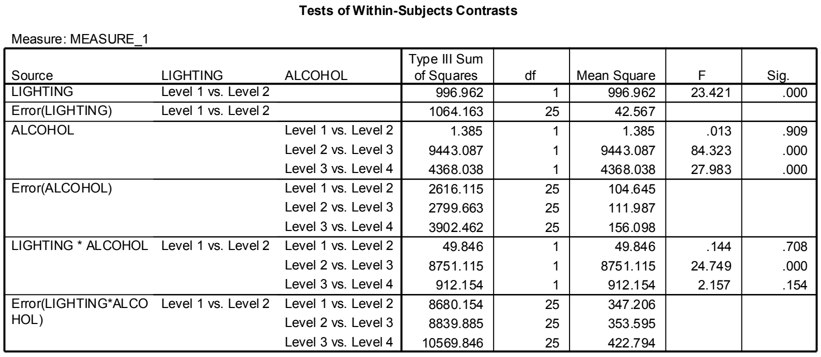
Based on estimated marginal means

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

a. Adjustment for multiple comparisons: Bonferroni.

Output 6

The interaction effect is shown in Output 14 by the *F*-ratio in the row labelled **Lighting*Alcohol**. The resulting *F*-ratio is 22.22 (1921.81/86.50), which has an associated probability value reported as .000. As such, there is a significant interaction between the amount of alcohol consumed and the lighting in the club on the attractiveness of the mate selected.



Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	LIGHTING	ALCOHOL	Type III Sum of Squares	df	Mean Square	F	Sig.
LIGHTING	Level 1 vs. Level 2		996.962	1	996.962	23.421	.000
Error(LIGHTING)	Level 1 vs. Level 2		1064.163	25	42.567		
ALCOHOL		Level 1 vs. Level 2	1.385	1	1.385	.013	.909
		Level 2 vs. Level 3	9443.087	1	9443.087	84.323	.000
		Level 3 vs. Level 4	4368.038	1	4368.038	27.983	.000
Error(ALCOHOL)		Level 1 vs. Level 2	2616.115	25	104.645		
		Level 2 vs. Level 3	2799.663	25	111.987		
		Level 3 vs. Level 4	3902.462	25	156.098		
LIGHTING * ALCOHOL	Level 1 vs. Level 2	Level 1 vs. Level 2	49.846	1	49.846	.144	.708
		Level 2 vs. Level 3	8751.115	1	8751.115	24.749	.000
		Level 3 vs. Level 4	912.154	1	912.154	2.157	.154
Error(LIGHTING*ALCOHOL)	Level 1 vs. Level 2	Level 1 vs. Level 2	8680.154	25	347.206		
		Level 2 vs. Level 3	8839.885	25	353.595		
		Level 3 vs. Level 4	10569.846	25	422.794		

Output 7

Output 16 shows a set of contrasts that compare each level of the alcohol variable to the previous level of that variable (this is called a *repeated* contrast in SPSS). So, it compares no pints with 2 pints (*Level 1 vs. Level 2*), 2 pints with 4 pints (*Level 2 vs. Level 3*) and 4 pints with 6 pints (*Level 3 vs. Level 4*). As you can see from the output, if we just look at the main effect of group these contrasts tell us what we already know from the *post hoc* tests, that is, the attractiveness after no alcohol doesn't differ from the attractiveness after 2 pints, $F(1, 25) < 1$, the attractiveness after 4 pints does differ from that after 2 pints, $F(1, 25) = 84.32$, $p < .001$, and the attractiveness after 6 pints does differ from that after 4 pints, $F(1, 25) = 27.98$, $p < .001$.

More interesting is to look at the interaction term in the table. This compares the same levels of the alcohol variable, but for each comparison it is also comparing the difference between the means for the dim and brightly lit clubs. One way to think of this is to look at the interaction graph and note the vertical differences between the means for dim and bright clubs at each level of alcohol. When nothing was drunk the distance between the bright and dim means is quite small (it's actually 3.42 units on the attractiveness scale), and when 2 pints of alcohol were drunk the difference between the dim and well-lit club is still quite small (4.81 units to be precise). The first contrast is comparing the difference between dim and bright clubs when nothing was drunk with the difference between dim and bright clubs when 2 pints were drunk. So, it is asking 'is 3.42 significantly different from 4.81?' The answer is 'no', because the *F*-ratio is non-significant – in fact, it's less than 1, $F(1, 25) < 1$. The second contrast for the interaction is looking at the difference between dim and bright clubs when 2 pints were drunk (4.81) and the difference between dim and bright clubs when 4 pints were drunk (this difference is –13.54; note that the direction of the difference has changed as indicated by the lines crossing in the graph). This difference is significant, $F(1, 25) = 24.75$, $p < .001$. The final contrast for the interaction is looking at the difference between dim and bright clubs when 4 pints were drunk (−13.54) and the difference between dim and bright clubs when 6 pints were drunk (this difference is –19.46). This contrast is not significant, $F(1, 25) = 2.16$, *ns*. So, we could conclude that there was a significant interaction between the amount of alcohol drunk and the lighting in the club. Specifically, the effect of alcohol after 2 pints on the attractiveness of the mate was much more pronounced when the lights were dim.

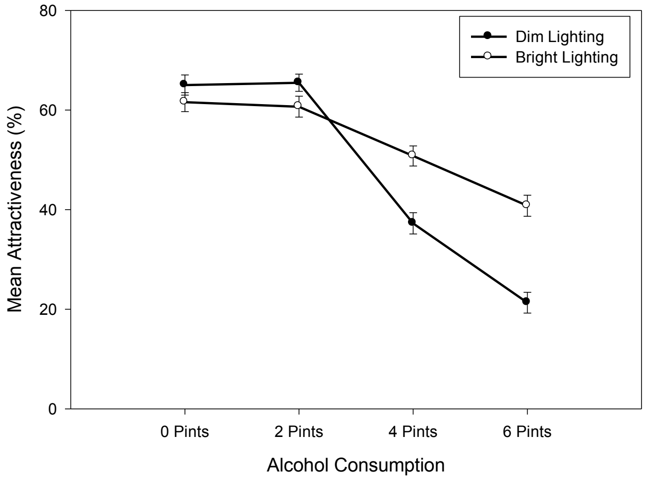


Figure 8

Writing the result

We can report the three effects from this analysis as follows:

- ✓ The results show that the attractiveness of the mates selected was significantly lower when the lighting in the club was dim compared to when the lighting was bright, *F*(1, 25) = 23.42, *p* < .001.
- ✓ The main effect of alcohol on the attractiveness of mates selected was significant, *F*(3, 75) = 104.39, *p* < .001. This indicated that when the lighting in the club was ignored, the attractiveness of the mates selected differed according to how much alcohol was drunk before the selection was made. Specifically, *post hoc* tests revealed that, compared to a baseline of when no alcohol had been consumed, the attractiveness of selected mates was not different after 2 pints (*p* > .05), but was significantly lower after 4 and 6 pints (*p* < .001 in both cases). The mean attractiveness after 2 pints was also significantly higher than after 4 pints and 6 pints (*p* < .001 in both cases), and the mean attractiveness after 4 pints was significantly higher than after 6 pints (*p* < .001). To sum up, the beer-goggles effect seems to take effect after 2 pints have been consumed and has an increasing impact until 6 pints are consumed.
- ✓ The lighting × alcohol interaction was significant, *F*(3, 75) = 22.22, *p* < .001, indicating that the effect of alcohol on the attractiveness of the mates selected differed when lighting was dim compared to when it was bright. Contrasts on this interaction term

revealed that when the difference in attractiveness ratings between dim and bright clubs was compared after no alcohol and after 2 pints had been drunk there was no significant difference, *F*(1, 25) < 1. However, when comparing the difference between dim and bright clubs when 2 pints were drunk with the difference after 4 pints were drunk a significant difference emerged, *F*(1, 25) = 24.75, *p* < .001. A final contrast revealed that the difference between dim and bright clubs after 4 pints were drunk compared to after 6 pints was not significant, *F*(1, 25) = 2.16, *ns*. To sum up, there was a significant interaction between the amount of alcohol drunk and the lighting in the club: the decline in the attractiveness of the selected mate seen after 2 pints (compared to after 4) was significantly more pronounced when the lights were dim.

Task 6

Using SPSS Tip 14.2, change the syntax in ***SimpleEffectsAttitude.sps*** *to look at the effect of drink at different levels of imagery.*

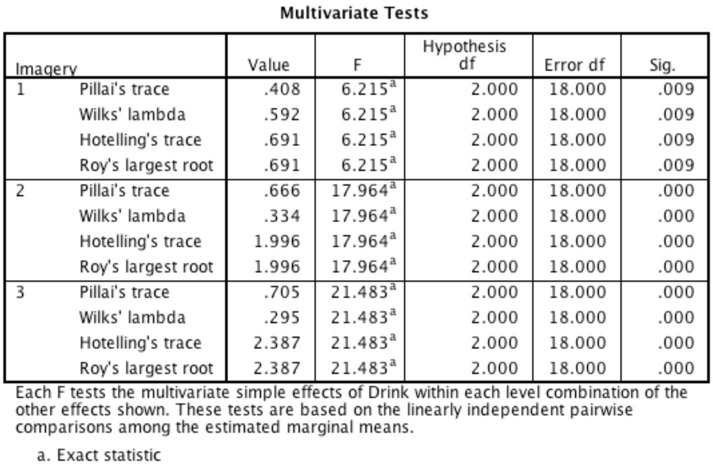
The correct syntax to use is:

```
GLM beerpos beerneg beerneut winepos wineneg wineneut waterpos waterneg waterneut

 /WSFACTOR=Drink 3  Imagery 3

/EMMEANS = TABLES(Drink*Imagery) COMPARE(Drink).
```

SPSS output



Multivariate Tests

Imagery		Value	F	Hypothesis df	Error df	Sig.
1	Pillai's trace	.408	6.215[a]	2.000	18.000	.009
	Wilks' lambda	.592	6.215[a]	2.000	18.000	.009
	Hotelling's trace	.691	6.215[a]	2.000	18.000	.009
	Roy's largest root	.691	6.215[a]	2.000	18.000	.009
2	Pillai's trace	.666	17.964[a]	2.000	18.000	.000
	Wilks' lambda	.334	17.964[a]	2.000	18.000	.000
	Hotelling's trace	1.996	17.964[a]	2.000	18.000	.000
	Roy's largest root	1.996	17.964[a]	2.000	18.000	.000
3	Pillai's trace	.705	21.483[a]	2.000	18.000	.000
	Wilks' lambda	.295	21.483[a]	2.000	18.000	.000
	Hotelling's trace	2.387	21.483[a]	2.000	18.000	.000
	Roy's largest root	2.387	21.483[a]	2.000	18.000	.000

Each F tests the multivariate simple effects of Drink within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Output 8

Output 17 shows is a significant effect of drink at level 1 of imagery. So, the ratings of the three drinks significantly differed when positive imagery was used. Because there are three levels of drink, though, this isn't that helpful in untangling what's going on. There is also a significant effect of drink at level 2 of imagery. So, the ratings of the three drinks significantly differed when negative imagery was used. Finally, there is also a significant effect of drink at level 3 of imagery. So, the ratings of the three drinks significantly differed when neutral imagery was used.

Task 7

> *A lot of my research looks at the effect of giving children information about animals. In one particular study (Field, 2006), I used three novel animals (the quoll, quokka and cuscus) and children were told negative things about one of the animals, positive things about another, and were given no information about the third (our control). I then asked the children to place their hands in three wooden boxes each of which they believed contained one of the aforementioned animals. The data are in the file* ***Field(2006).sav****. Draw an error bar graph of the means, then do some normality tests on the data.*

Error bar graph

You really ought to know how to do an error bar graph by now, so all I will say is that it should look something like Figure 9.

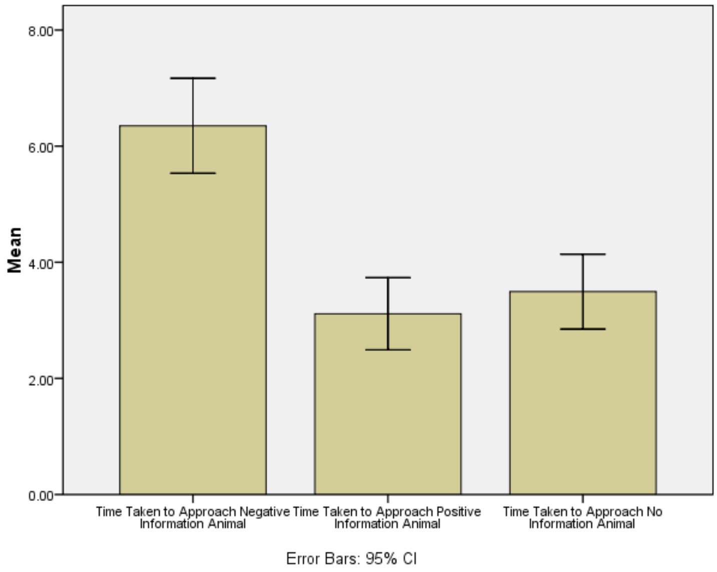
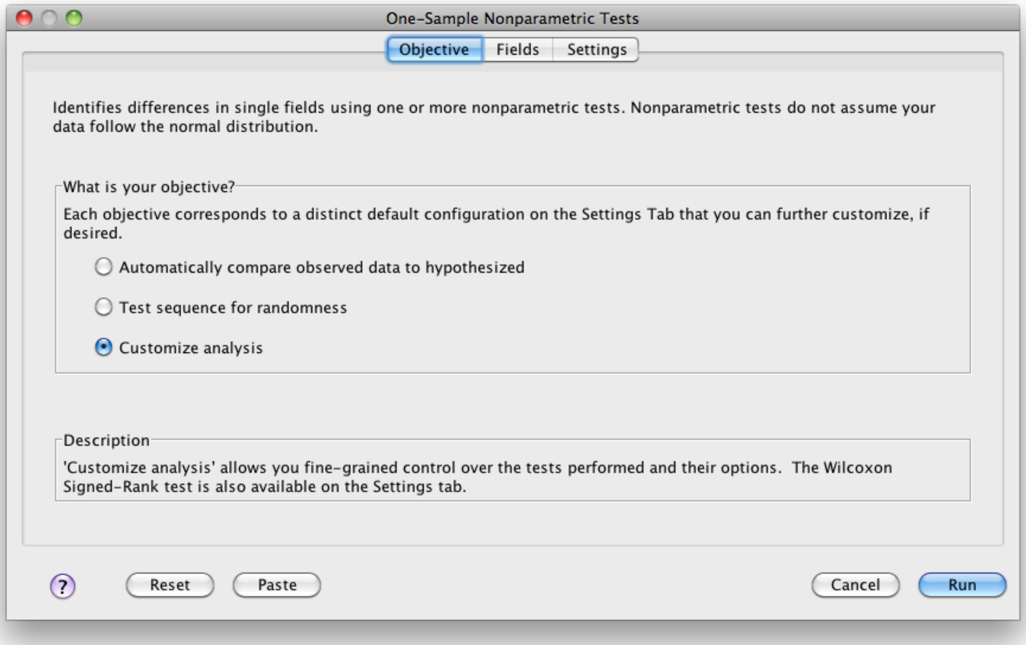


Figure 9

Normality tests

To get the normality tests I used the Kolmogorov–Smirnov test from the Nonparametric⇒One Sample… menu. The reason I did this is because I had a fairly large sample. The K-S test executed through this menu differs from that obtained through the Explore procedure because it does not use the Lilliefors correction (see the additional materials for Chapter 5). To get this test complete the dialog boxes as follows:



Identifies differences in single fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

○ Automatically compare observed data to hypothesized

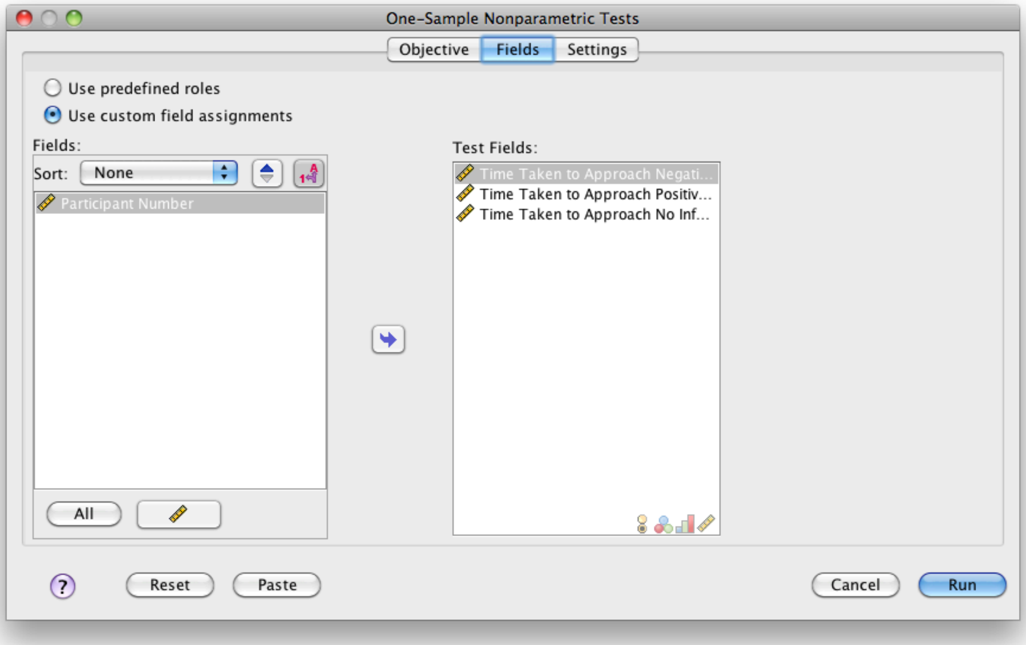
○ Test sequence for randomness

● Customize analysis

Description

'Customize analysis' allows you fine-grained control over the tests performed and their options. The Wilcoxon Signed-Rank test is also available on the Settings tab.

Figure 1



○ Use predefined roles
● Use custom field assignments

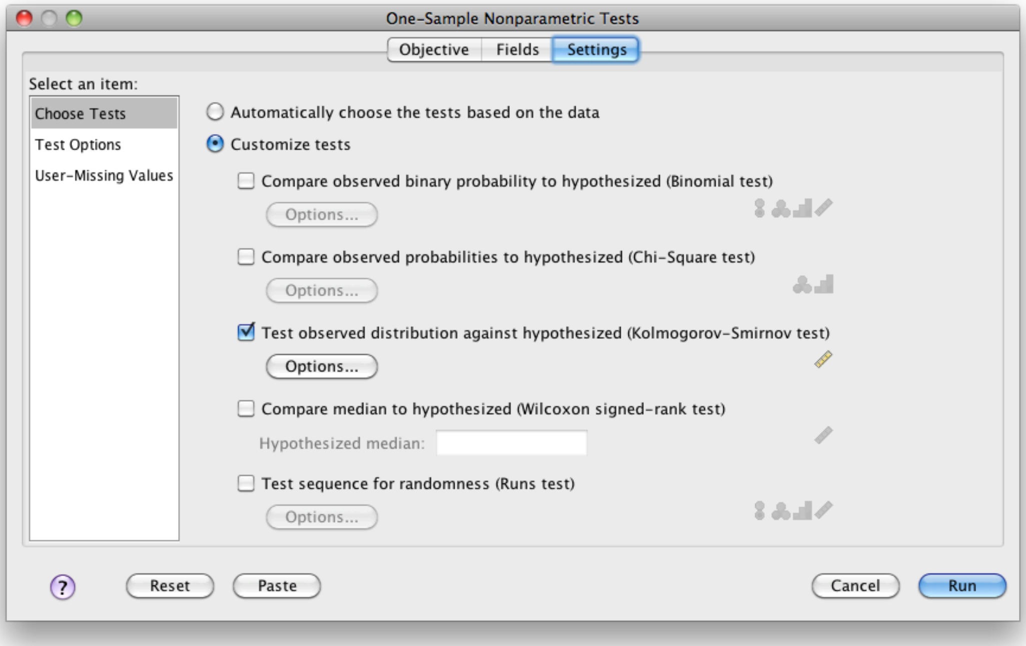
Fields:
Sort: None

Participant Number

Test Fields:

Time Taken to Approach Negati...
Time Taken to Approach Positiv...
Time Taken to Approach No Inf...

Figure 2



Select an item:

Choose Tests
Test Options
User-Missing Values

○ Automatically choose the tests based on the data
● Customize tests

□ Compare observed binary probability to hypothesized (Binomial test)

□ Compare observed probabilities to hypothesized (Chi-Square test)

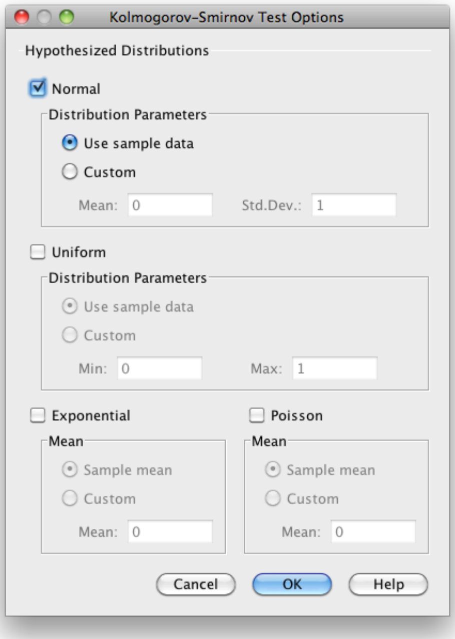
☑ Test observed distribution against hypothesized (Kolmogorov-Smirnov test)

□ Compare median to hypothesized (Wilcoxon signed-rank test)

Hypothesized median:

□ Test sequence for randomness (Runs test)

Figure 3



Kolmogorov-Smirnov Test Options

Hypothesized Distributions

☑ Normal

Distribution Parameters

● Use sample data

○ Custom

Mean: 0 Std.Dev.: 1

□ Uniform

Distribution Parameters

● Use sample data

○ Custom

Min: 0 Max: 1

□ Exponential

Mean

● Sample mean

○ Custom

Mean: 0

□ Poisson

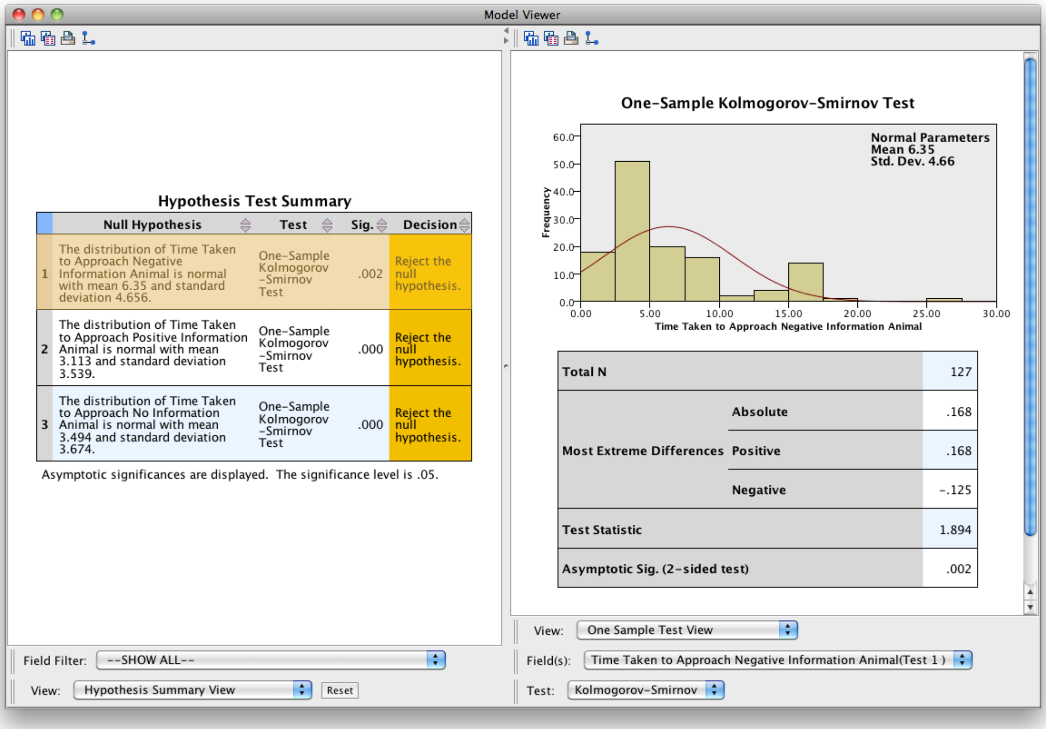
Mean

● Sample mean

○ Custom

Mean: 0

Figure 13



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Time Taken to Approach Negative Information Animal is normal with mean 6.35 and standard deviation 4.656.	One-Sample Kolmogorov-Smirnov Test	.002	Reject the null hypothesis.
2	The distribution of Time Taken to Approach Positive Information Animal is normal with mean 3.113 and standard deviation 3.539.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
3	The distribution of Time Taken to Approach No Information Animal is normal with mean 3.494 and standard deviation 3.674.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N		127
Most Extreme Differences	Absolute	.168
	Positive	.168
	Negative	-.125
Test Statistic		1.894
Asymptotic Sig. (2-sided test)		.002

Field Filter: --SHOW ALL--

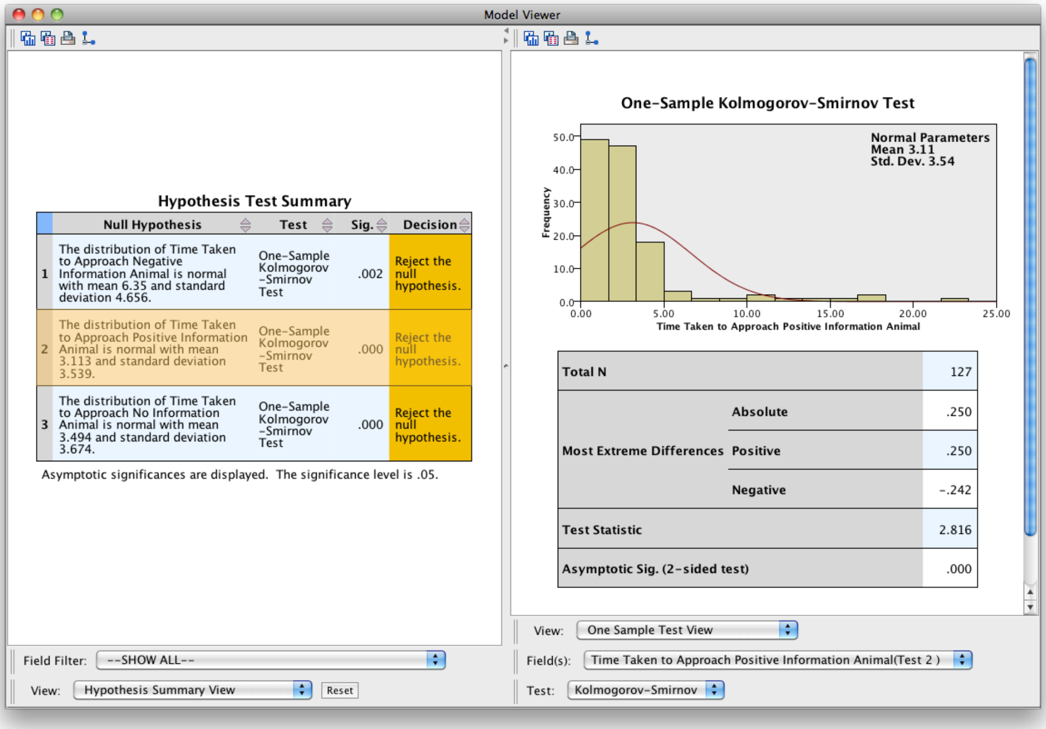
View: Hypothesis Summary View Reset

View: One Sample Test View

Field(s): Time Taken to Approach Negative Information Animal(Test 1)

Test: Kolmogorov-Smirnov

Figure 14



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Time Taken to Approach Negative Information Animal is normal with mean 6.35 and standard deviation 4.656.	One-Sample Kolmogorov-Smirnov Test	.002	Reject the null hypothesis.
2	The distribution of Time Taken to Approach Positive Information Animal is normal with mean 3.113 and standard deviation 3.539.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
3	The distribution of Time Taken to Approach No Information Animal is normal with mean 3.494 and standard deviation 3.674.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N		127
Most Extreme Differences	Absolute	.250
	Positive	.250
	Negative	-.242
Test Statistic		2.816
Asymptotic Sig. (2-sided test)		.000

Field Filter: --SHOW ALL--

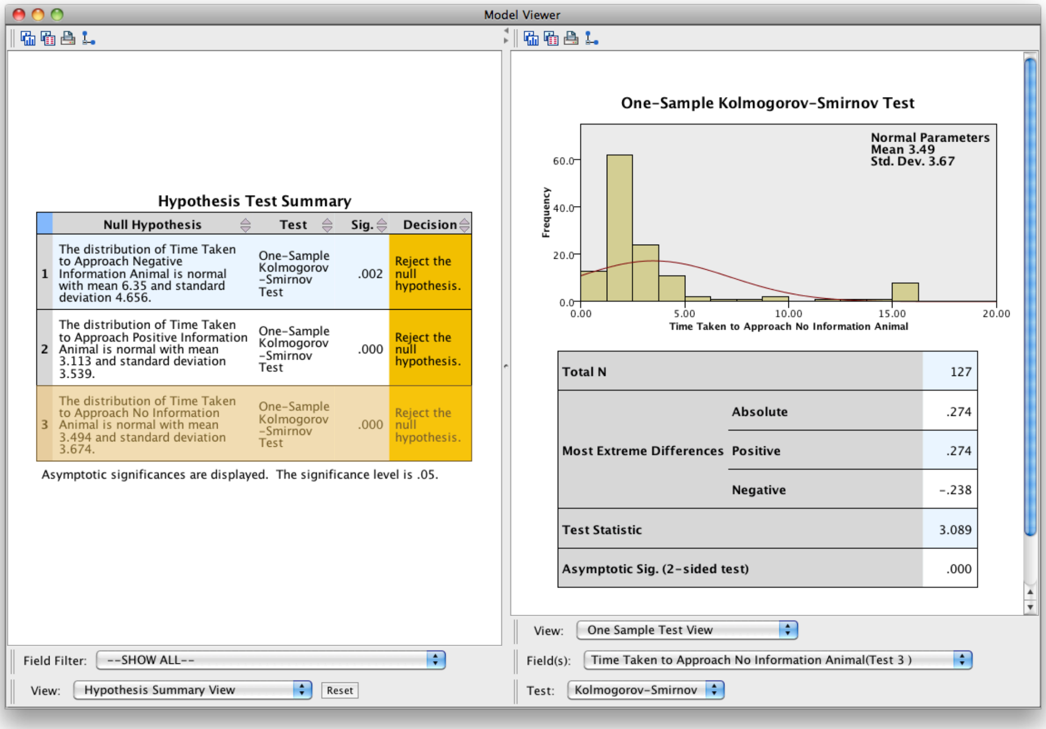
View: Hypothesis Summary View

View: One Sample Test View

Field(s): Time Taken to Approach Positive Information Animal(Test 2)

Test: Kolmogorov-Smirnov

Figure 15



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Time Taken to Approach Negative Information Animal is normal with mean 6.35 and standard deviation 4.656.	One-Sample Kolmogorov-Smirnov Test	.002	Reject the null hypothesis.
2	The distribution of Time Taken to Approach Positive Information Animal is normal with mean 3.113 and standard deviation 3.539.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
3	The distribution of Time Taken to Approach No Information Animal is normal with mean 3.494 and standard deviation 3.674.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N		127
Most Extreme Differences	Absolute	.274
	Positive	.274
	Negative	-.238
Test Statistic		3.089
Asymptotic Sig. (2-sided test)		.000

Field Filter: --SHOW ALL--

View: Hypothesis Summary View Reset

View: One Sample Test View

Field(s): Time Taken to Approach No Information Animal(Test 3)

Test: Kolmogorov-Smirnov

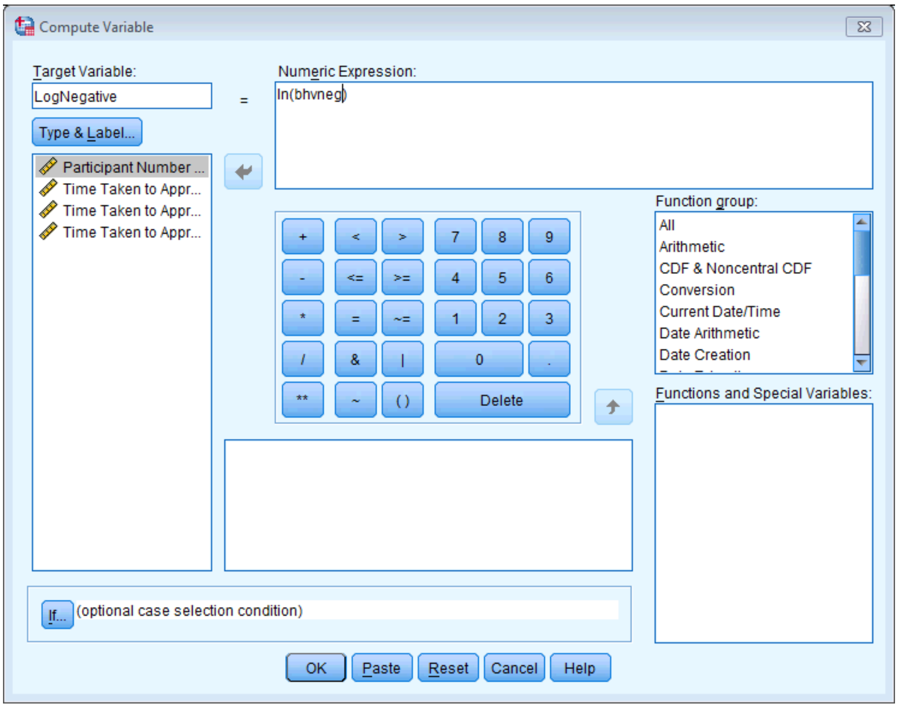
Figure 16

The resulting K–S tests show that the data are very heavily non-normal. If you look at the Q–Q and P–P plots you will see that the data are very heavily skewed. This will be, in part, because if a child didn't put their hand in the box after 15 seconds we gave them a score of 15 and asked them to move on to the next box (this was for ethical reasons: if a child hadn't put their hand in the box after 15 s we assumed that they did not want to do the task).

Task 8

Log-transform the scores in Task 7 and repeat the normality tests.

To log-transform the scores we need to use the *compute* function (Figure 17).



Target Variable: LogNegative

Numeric Expression: ln(bhvneg)

Figure 17

We need to do this three times (once for each variable). Alternatively, we could use the following syntax:

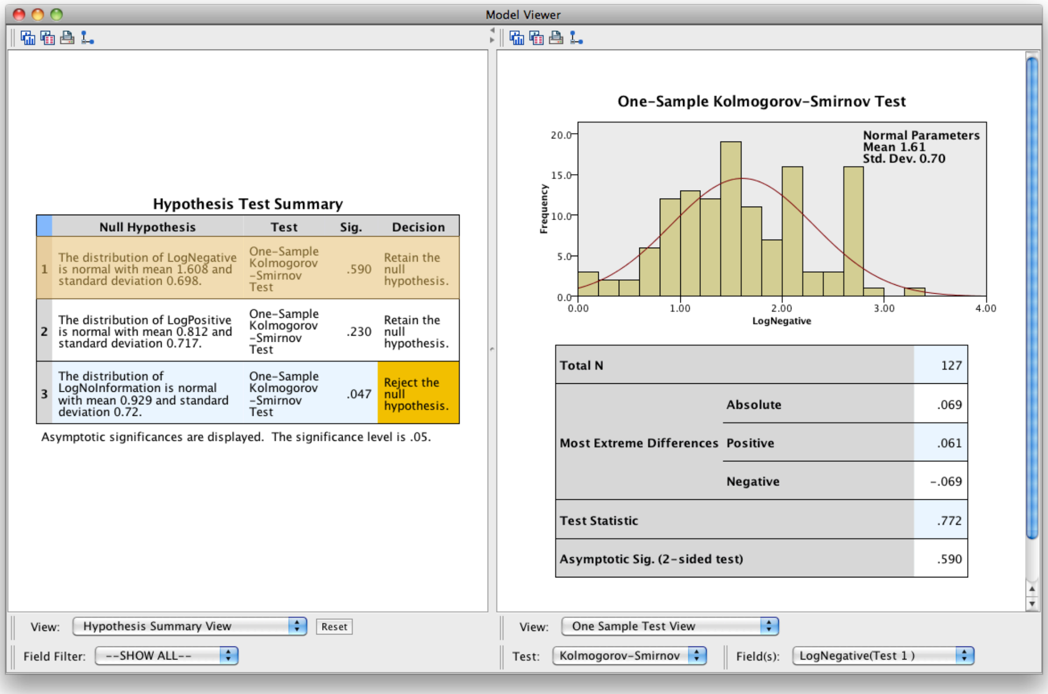
```
COMPUTE LogNegative=ln(bhvneg).

COMPUTE LogPositive=ln(bhvpos).

COMPUTE LogNoInformation=ln(bhvnone).

EXECUTE.
```

If we rerun the K-S test on these transformed scores we get the output shown in Figures 18–20.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of LogNegative is normal with mean 1.608 and standard deviation 0.698.	One-Sample Kolmogorov -Smirnov Test	.590	Retain the null hypothesis.
2	The distribution of LogPositive is normal with mean 0.812 and standard deviation 0.717.	One-Sample Kolmogorov -Smirnov Test	.230	Retain the null hypothesis.
3	The distribution of LogNoInformation is normal with mean 0.929 and standard deviation 0.72.	One-Sample Kolmogorov -Smirnov Test	.047	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

View: Hypothesis Summary View

Field Filter: --SHOW ALL--

One-Sample Kolmogorov-Smirnov Test

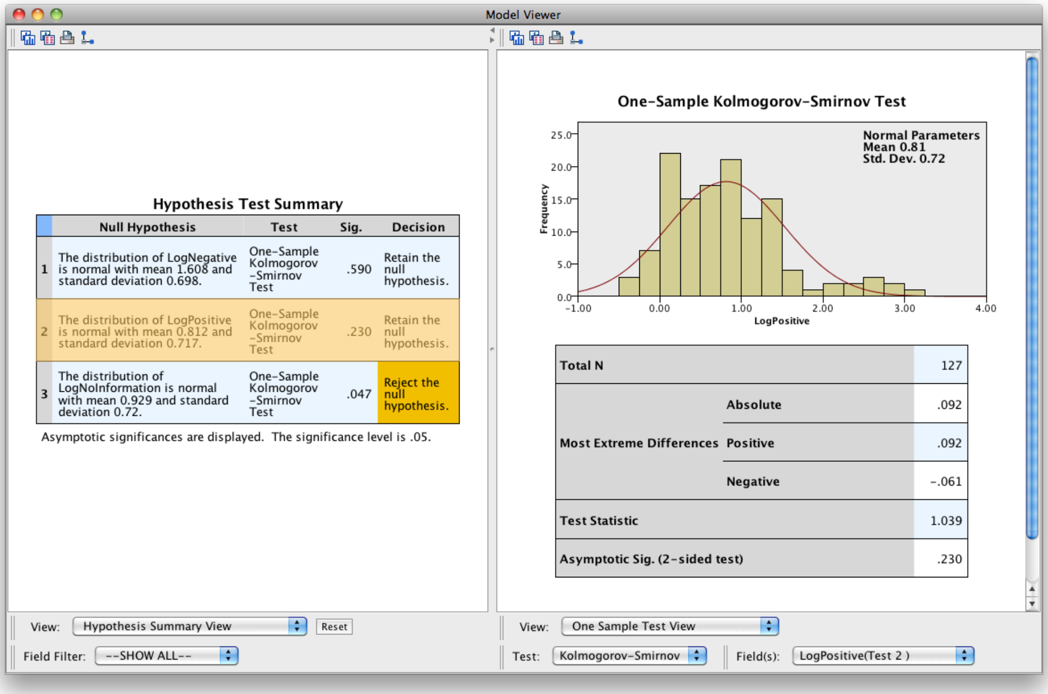
Total N		127
Most Extreme Differences	Absolute	.069
	Positive	.061
	Negative	-.069
Test Statistic		.772
Asymptotic Sig. (2-sided test)		.590

View: One Sample Test View

Test: Kolmogorov-Smirnov

Field(s): LogNegative(Test 1)

Figure 18



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of LogNegative is normal with mean 1.608 and standard deviation 0.698.	One-Sample Kolmogorov -Smirnov Test	.590	Retain the null hypothesis.
2	The distribution of LogPositive is normal with mean 0.812 and standard deviation 0.717.	One-Sample Kolmogorov -Smirnov Test	.230	Retain the null hypothesis.
3	The distribution of LogNoInformation is normal with mean 0.929 and standard deviation 0.72.	One-Sample Kolmogorov -Smirnov Test	.047	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N		127
Most Extreme Differences	Absolute	.092
	Positive	.092
	Negative	-.061
Test Statistic		1.039
Asymptotic Sig. (2-sided test)		.230

View: Hypothesis Summary View

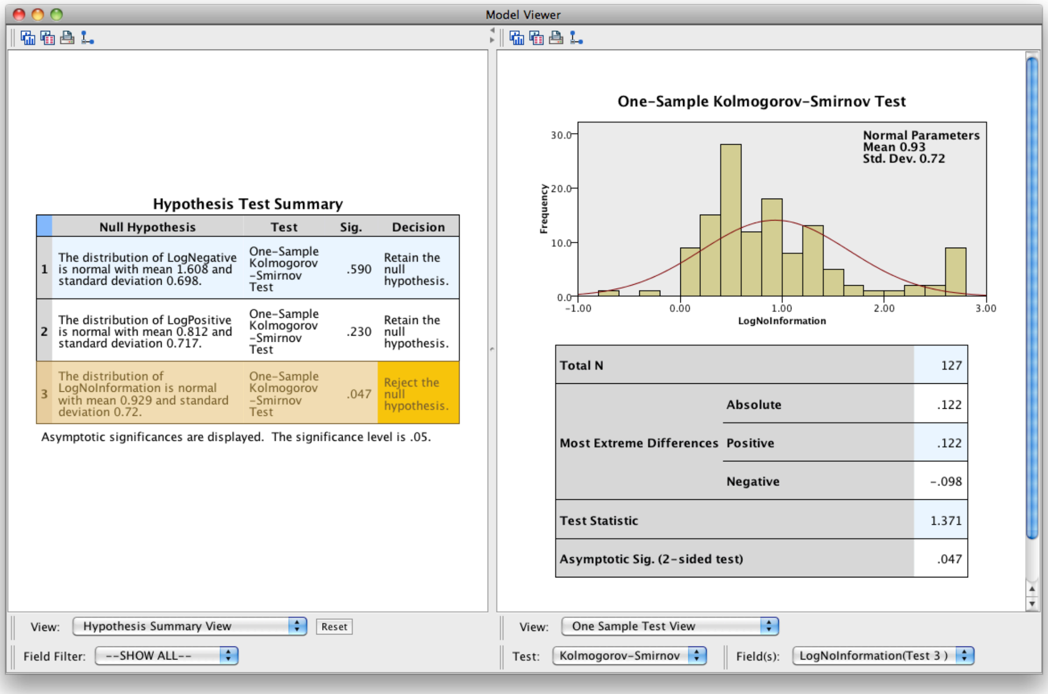
Field Filter: --SHOW ALL--

View: One Sample Test View

Test: Kolmogorov-Smirnov

Field(s): LogPositive(Test 2)

Figure 4


Model Viewer
One-Sample Kolmogorov-Smirnov Test
Normal Parameters
Mean 0.93
Std. Dev. 0.72
Frequency
30.0
20.0
10.0
0.0
-1.00
0.00
1.00
2.00
3.00
LogNoInformation
View: Hypothesis Summary View
Reset
Field Filter: --SHOW ALL--
View: One Sample Test View
Test: Kolmogorov-Smirnov
Field(s): LogNoInformation(Test 3)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of LogNegative is normal with mean 1.608 and standard deviation 0.698.	One-Sample Kolmogorov -Smirnov Test	.590	Retain the null hypothesis.
2	The distribution of LogPositive is normal with mean 0.812 and standard deviation 0.717.	One-Sample Kolmogorov -Smirnov Test	.230	Retain the null hypothesis.
3	The distribution of LogNoInformation is normal with mean 0.929 and standard deviation 0.72.	One-Sample Kolmogorov -Smirnov Test	.047	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

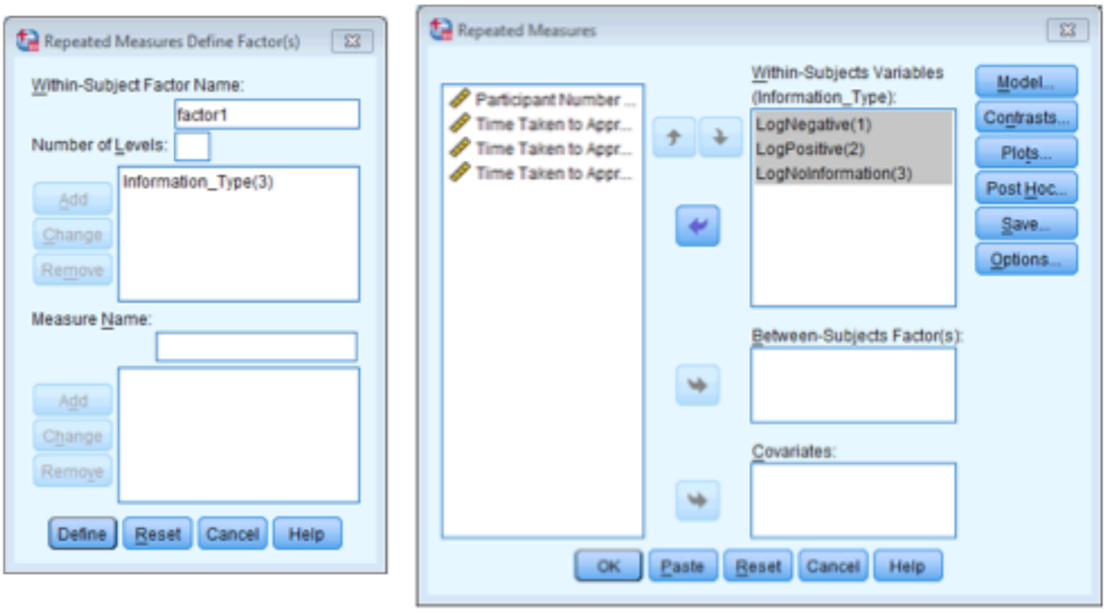
Total N		127
Most Extreme Differences	Absolute	.122
	Positive	.122
	Negative	-.098
Test Statistic		1.371
Asymptotic Sig. (2-sided test)		.047

Figure 20

Task 9:

Conduct a one-way ANOVA on the log-transformed scores in Task 8. Do children take longer to put their hands in a box that they believe contains an animal about which they have been told nasty things?

To do the ANOVA we have to define a variable called **Information_Type** and then specify the three logged variables (Figure 21).



Repeated Measures Define Factor(s)

Within-Subject Factor Name: factor1

Number of Levels:

Information_Type(3)

Measure Name:

Repeated Measures

Participant Number ...
Time Taken to Appr...
Time Taken to Appr...
Time Taken to Appr...

Within-Subjects Variables (Information_Type):

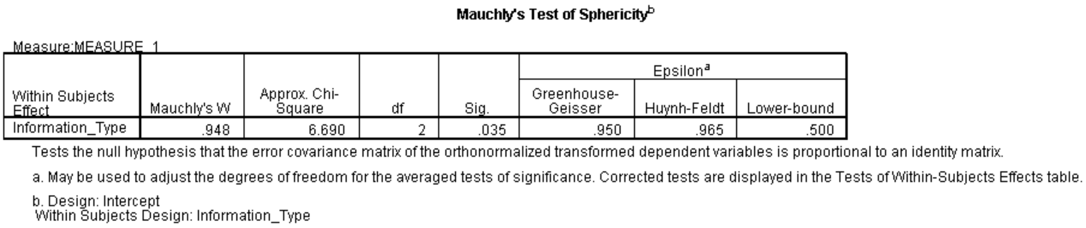
LogNegative(1)
LogPositive(2)
LogNoInformation(3)

Between-Subjects Factor(s):

Covariates:

Figure 5

You can specify some simple contrasts (comparing everything to the last category (no information) or *post hoc* tests. I actually did something slightly different because I wanted to get precise Bonferroni-corrected confidence intervals for my *post hoc* comparisons, but if you ask for some *post hoc* tests you will get the same profile of results that I did.



Mauchly's Test of Sphericity[b]

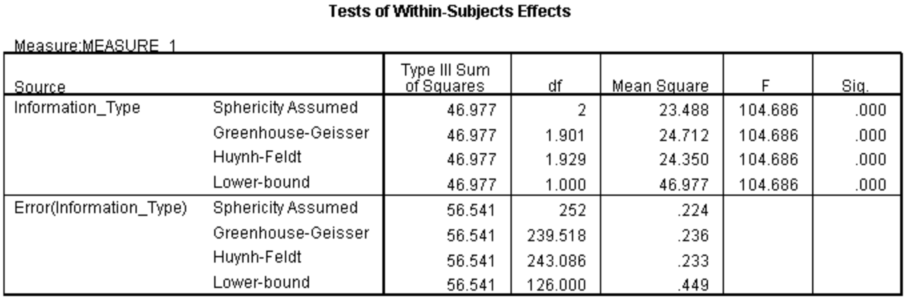
Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Information_Type	.948	6.690	2	.035	.950	.965	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

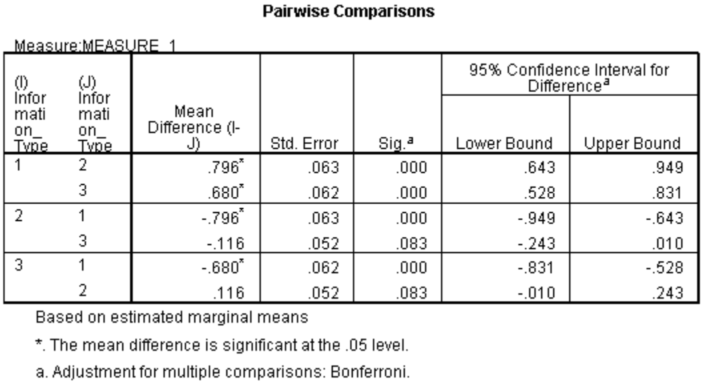
b. Design: Intercept
Within Subjects Design: Information_Type



Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Information_Type	Sphericity Assumed	46.977	2	23.488	104.686	.000
	Greenhouse-Geisser	46.977	1.901	24.712	104.686	.000
	Huynh-Feldt	46.977	1.929	24.350	104.686	.000
	Lower-bound	46.977	1.000	46.977	104.686	.000
Error(Information_Type)	Sphericity Assumed	56.541	252	.224		
	Greenhouse-Geisser	56.541	239.518	.236		
	Huynh-Feldt	56.541	243.086	.233		
	Lower-bound	56.541	126.000	.449		



Pairwise Comparisons

Measure:MEASURE_1

(I) Information_Type	(J) Information_Type	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.796*	.063	.000	.643	.949
	3	.680*	.062	.000	.528	.831
2	1	-.796*	.063	.000	-.949	-.643
	3	-.116	.052	.083	-.243	.010
3	1	-.680*	.062	.000	-.831	-.528
	2	.116	.052	.083	-.010	.243

Based on estimated marginal means

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

a. Adjustment for multiple comparisons: Bonferroni.

Output 18

Note from Output 18, first of all, that the sphericity test is significant. Therefore, in Field (2006) I reported Greenhouse–Geisser-corrected degrees of freedom and significance. The main ANOVA shows that the type of information significantly affected how long the children took to place their hands in the boxes. The *post hoc* tests and the graph tell us that a child took longer to place their hand in the box that they believed contained an animal about which they had heard bad things compared to the boxes that they believed contained animals that they had heard positive information or no information about. There was not a significant difference between the approach times for the 'positive information' and 'no information' boxes.

You could report these results as follows:

> The latencies to approach the boxes were positively skewed (Kolmogorov–Smirnov *D* = 1.89, 2.82, 3.09 for the threat, positive and no information boxes, respectively) and so were transformed using the natural log of the score. The

resulting distributions were not significantly different from normal (Kolmogorov–Smirnov D = 0.77, 1.04 and 1.37 for the threat, positive and no information boxes, respectively). A one-way repeated-measures ANOVA revealed a significant main effect of the type of box,[2] $F(1.90, 239.52) = 104.69$, $p < .001$. Bonferroni-corrected *post hoc* tests revealed a significant difference between the threat information box and the positive information box, $p < .001$; the threat information box and the no information box, $p < .001$; but not the positive information box and the no information box, $p > .05$.

[2] An analysis of the untransformed scores using a non-parametric test (Friedman's ANOVA) also revealed significant differences between approach times to the boxes, $\chi^2(2) = 140.36$, $p < .001$.