Post Hoc Procedures in ANOVA & The General Linear Model

Eysenck Example:

Number of Words Recalled as a Function of the Level of Processing

	Rhyming	Imagery	Intentional	Total
	11	23	19	
	9	19	15	
	8	16	14	
	7	12	14	
	7	12	11	
	6	11	11	
	6	11	11	
	6	11	10	
	6	10	10	
	3	9	5	
Mean	6.90	13.40	12.00	10.77
s	2.13	4.50	3.74	4.49
s^2	4.54	20.27	14.00	20.12

The abbreviated data above are taken from Eysenck's study of the relationship between recall and level of processing. Eysenck randomly assigned 30 participants between the ages of 55 – 65 years to one of three groups. Participants were then asked to read through a list of 27 words three times and follow specific instructions when reading the words. Instructions for each group are as follows:

1. Rhyming: Participants are asked to think of a word that rhymes with each word.
2. Imagery: Participants are asked to form vivid images of each word.
3. Intentional: Participants are asked to memorize words for later recall.

Post Hoc Procedures

After running an ANOVA and finding a significant overall (omnibus *F* test) difference, we know that at least two groups differ significantly. However, we do not know which two groups differ significantly or if all groups differ significantly.

Why not just conduct several *t* tests comparing each pair of group means after a significant F test?

Types of Error Rates

When a single decision is being made (e.g., t test, overall F test), the probability of committing a Type I error is the chance of rejecting a true null hypothesis. When multiple groups are compared, there is more than one decision and more than one way to describe Type I error.

Error rate per contrast/comparison is the probability that a particular contrast will be falsely declared significant. In other words, if a contrast whose true population value is zero were to be tested over and over again in repeated studies, error rate per contrast is the proportion of times that the contrast would be found to be statistically significant.

Experimentwise error rate, is the probability that one or more comparisons/contrasts will be falsely declared significant in an experiment. In other words, if an experiment were to be conducted repeatedly, experimentwise error rate is the proportion of those experiments that would contain at least 1 type I error.

In designs with more than one factor, it is necessary to define yet another error rate, called the **familywise error rate**. In multifactor designs, significance tests involving different factors are usually regarded as constituting different families. For this reason, a single experiment may contain several families of tests, in which case familywise and experimentwise are different. However, in single factor designs, which is all that we have discussed until now, familywise and experimentwise are identical.

Error rate per experiment is the expected number of contrasts that will be falsely declared significant in a single experiment. Notice this is not a probability and can exceed a value of one in certain circumstances.

Let's look at an example with four groups and three contrasts to be tested. In this example, error rate per contrast/comparison is equal to .05 because each comparison was tested at an alpha level of .05. For any single comparison, there is a 5 percent chance of a Type I error. What is the value of the error rate per experiment? Error rate per experiment will equal .15 because the expected number of type I errors per contrast is .05 and there are three contrasts tested in the experiment. In general, with *c* contrasts each tested at alpha level per contrast, error rate per experiment = alpha**c*. Finally, experimentwise alpha is the probability of at least one Type I error being made in the experiment. You can consider $(1-\alpha)$ to be the probability of no Type I error for a particular comparison. If comparisons are independent, then the following is true:

Pr(at least one Type I error) = 1 – Pr(no Type I errors) = 1 – (1- alpha)^c

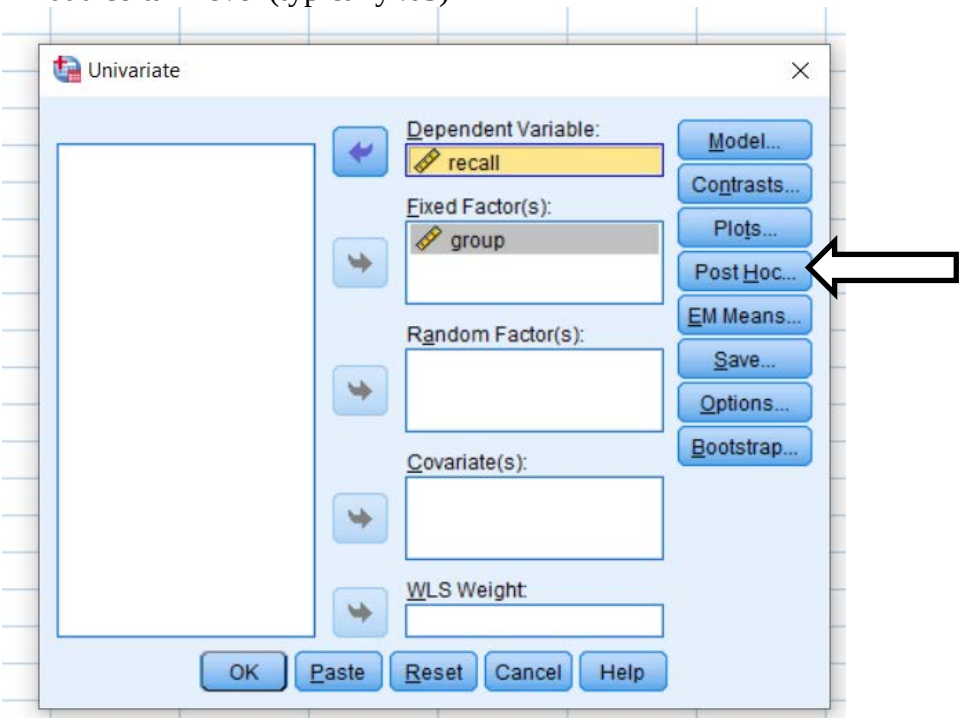
In our example then, experimentwise error rate = $1-(1-.05)^3 = 1-(.95)^3 = 1-.857 = .143$

Which error rate should be .05? When multiple contrasts are tested, it is impossible to achieve a .05 value for all three types of error. Instead, a decision must be made regarding which type of error is to be controlled at the 5 percent level.

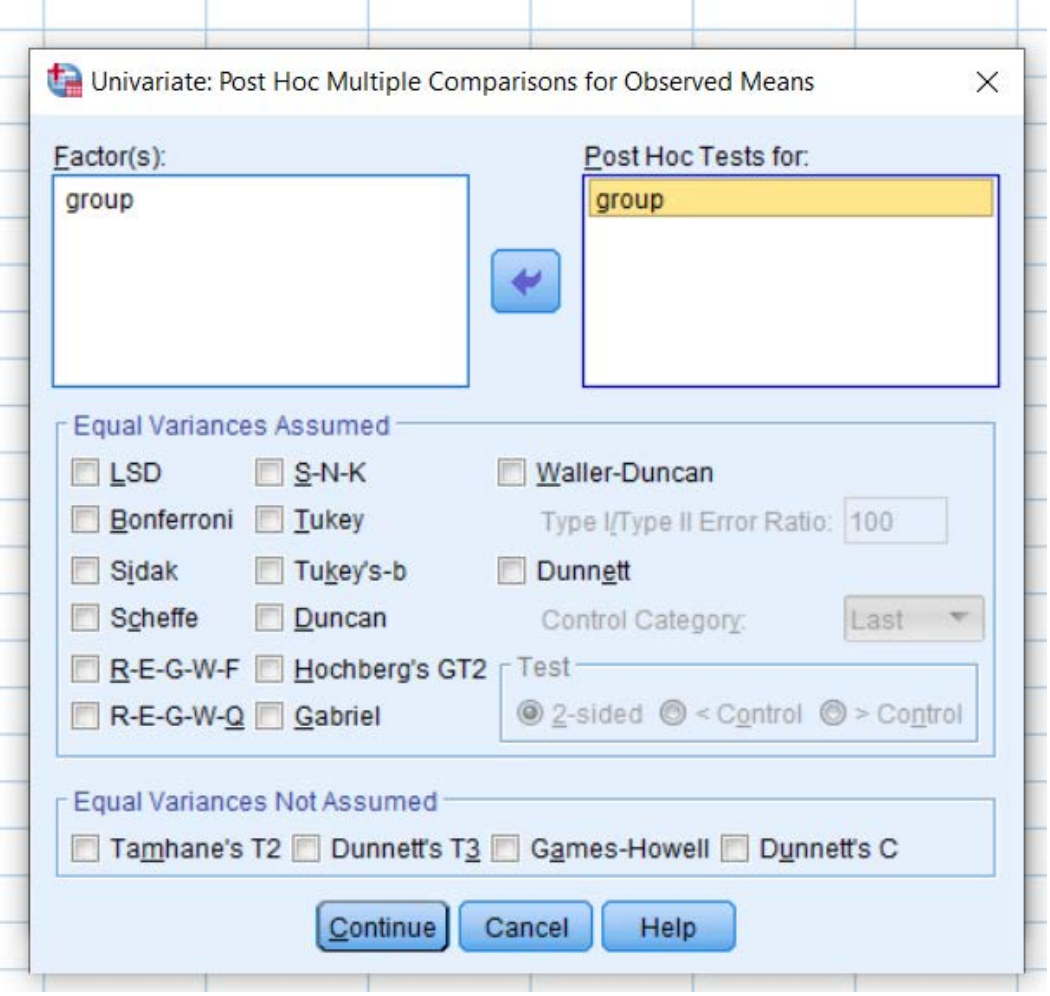
Usually the preference is to control the experimentwise error rate at .05. By keeping experimentwise error rate at .05, the probability of a Type I error occurring anywhere in a given experiment is at most .05. If, instead, error rate per comparison were controlled at .05, studies with multiple contrasts would have a higher Type I error rate than .05. In this situation, an experimenter could increase his or her chances of obtaining a statistically significant result simply by testing many contrasts.

Post hoc procedures will allow us to:

a. find where the significant group differences exist
b. some procedures maintain overall alpha (called experimentwise or familywise error rate) at a certain level (typically .05)



There are several post hoc procedures:



Pairwise Multiple Comparisons

Designed to test differences between all pairs of means within an experiment. Assumes that all differences between means are important and of theoretical interest.

Fisher's LSD
Duncan's Multiple Range Test
Newman-Keuls (S-N-K in SPSS)
Tukey's HSD
Bonferroni
Scheffe

Fisher's Least Significant Difference (LSD) Test

Simultaneous Method
Uses the *t* statistic
Extremely Powerful
Does not control for experimentwise error rate
Do not use with more than three groups

Hypothesis: $H_0 : \mu_i = \mu_j$ $H_1 : \mu_i \neq \mu_j$

LSD Critical Test Statistic: LSD = $t_{\alpha\,;\,N-k}\sqrt{MS_w(\frac{1}{n_i}+\frac{1}{n_j})}$

Compare absolute mean difference ($\bar{x}_i - \bar{x}_j$) to LSD Critical (*t* statistic) calculated above. If the absolute mean difference is larger than the LSD Critical, Reject the Null. If the absolute mean difference is less than the LSD Critical, Fail to Reject the Null.

Confidence Intervals: $(\bar{x}_i - \bar{x}_j) \pm t_{\alpha\,;\,N-k}\sqrt{MS_w(\frac{1}{n_i}+\frac{1}{n_j})}$

Duncan's Multiple Range Test

Stepdown Method
Powerful
Does not control experimentwise error rate well
Uses the q statistic
Similar to Tukey and Newman-Keuls approaches with an adjustment to reduce the level of protection against Type I error

Newman-Keuls (S-N-K in SPSS)

Stepdown Method
Uses the q statistic
Controls for experimentwise error rate
Same calculations as with Tukey, but compares largest mean difference first, followed by the second largest mean difference until no longer significant

Tukey's HSD (Honestly Significant Difference; just called Tukey in SPSS) is typically preferred for paired comparisons:

Simultaneous Methods
Uses the *q* statistic
Controls experimentwise alpha
Fairly powerful at detecting difference

Hypothesis: $H_0 : \mu_i = \mu_j$ $H_1 : \mu_i \neq \mu_j$

Tukey's HSD Test Statistic: HSD = $\frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{MS_w}{n}}}$ where *n* is the sample size when groups are equal.

Compare HSD statistic to Critical value (q statistic). If the HSD statistic calculated above is larger than the Critical value, Reject the Null. If the calculated HSD statistic is less than the Critical value, Fail to Reject the Null.

The Critical Value is given by the studentized range statistic (q; see Table B.2 on pages 405-406 for critical values): $q_{\alpha\,;\,k,N-k}$

Confidence Intervals: $(\bar{x}_i - \bar{x}_j) \pm q_{\alpha\,;\,k,N-k}\sqrt{\frac{MS_w}{n}}$

In unbalanced designs (unequal group sizes), use the harmonic mean:$2n_in_j/(n_i + n_j)$. The harmonic mean is used by default in SPSS.

Bonferroni

Simultaneous Method
Uses the *t* statistic
Same as Fisher's LSD method with adjusted *p* values [alpha = experimentwise alpha (usually .05)/number of comparisons]
Controls for experimentwise error
Fairly conservative test, especially as number of groups increase

Scheffé Test

Simultaneous Method
Uses the *F* statistic
Controls experimentwise alpha
Conservative at detecting differences

Hypothesis: $H_0 : \mu_i = \mu_j \quad H_1 : \mu_i \neq \mu_j$

Scheffé Test Statistic: S = $\frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{MS_w}{n}}}$ where *n* is the sample size when groups are equal.

Critical Value:$\sqrt{(k-1)F_{\alpha\,;\;k-1,N-k}}$

Compare Scheffé statistic to Critical value. If the Scheffé test statistic calculated above is larger than the Critical value, Reject the Null. If the calculated Scheffé test statistic is less than the Critical value, Fail to Reject the Null.

Confidence Intervals: $(\bar{x}_i - \bar{x}_j) \pm \sqrt{(k-1)F_{\alpha\,;\;k-1,N-k}}\sqrt{\frac{MS_w}{n}}$

In unbalanced designs (unequal group sizes), use the harmonic mean:$2n_in_j/(n_i + n_j)$. The harmonic mean is used by default in SPSS.

Dunnett's t

Compares a control group to each treatment group
Controls for experimentwise error rate
Critical Values in Table B.3 on page 407

Eysenck Example:

Between-Subjects Factors

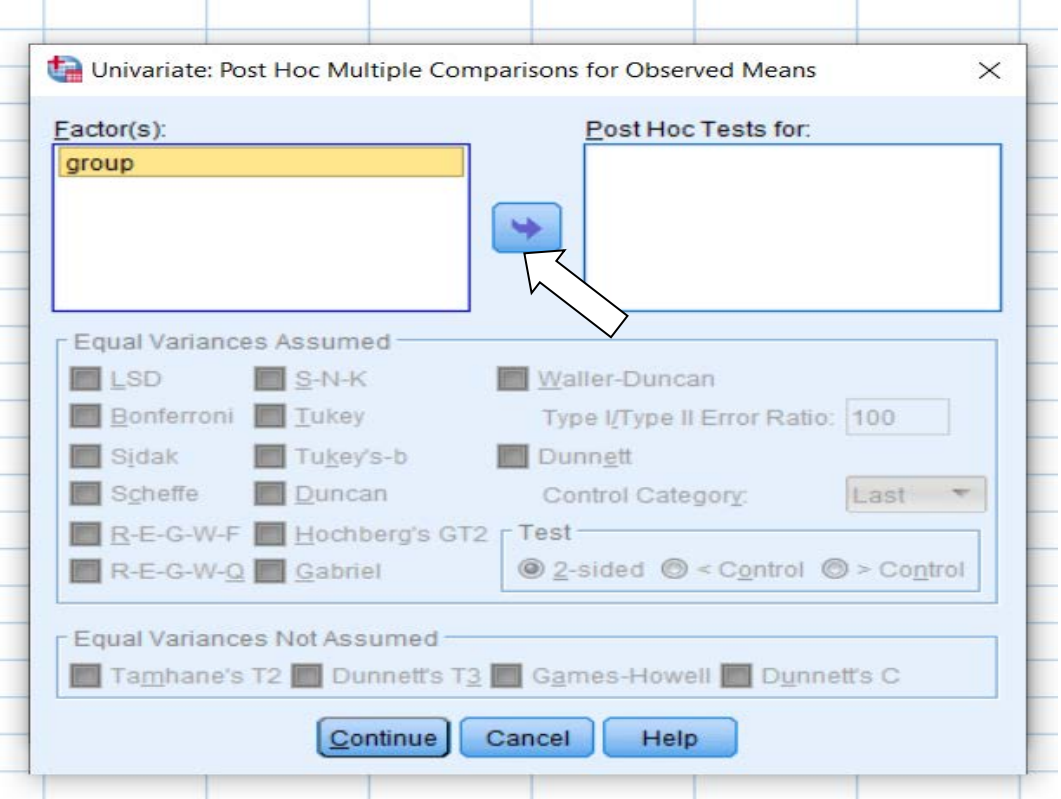
		Value Label	N
group	1.00	rhyming	10
	2.00	imagery	10
	3.00	intentional	10

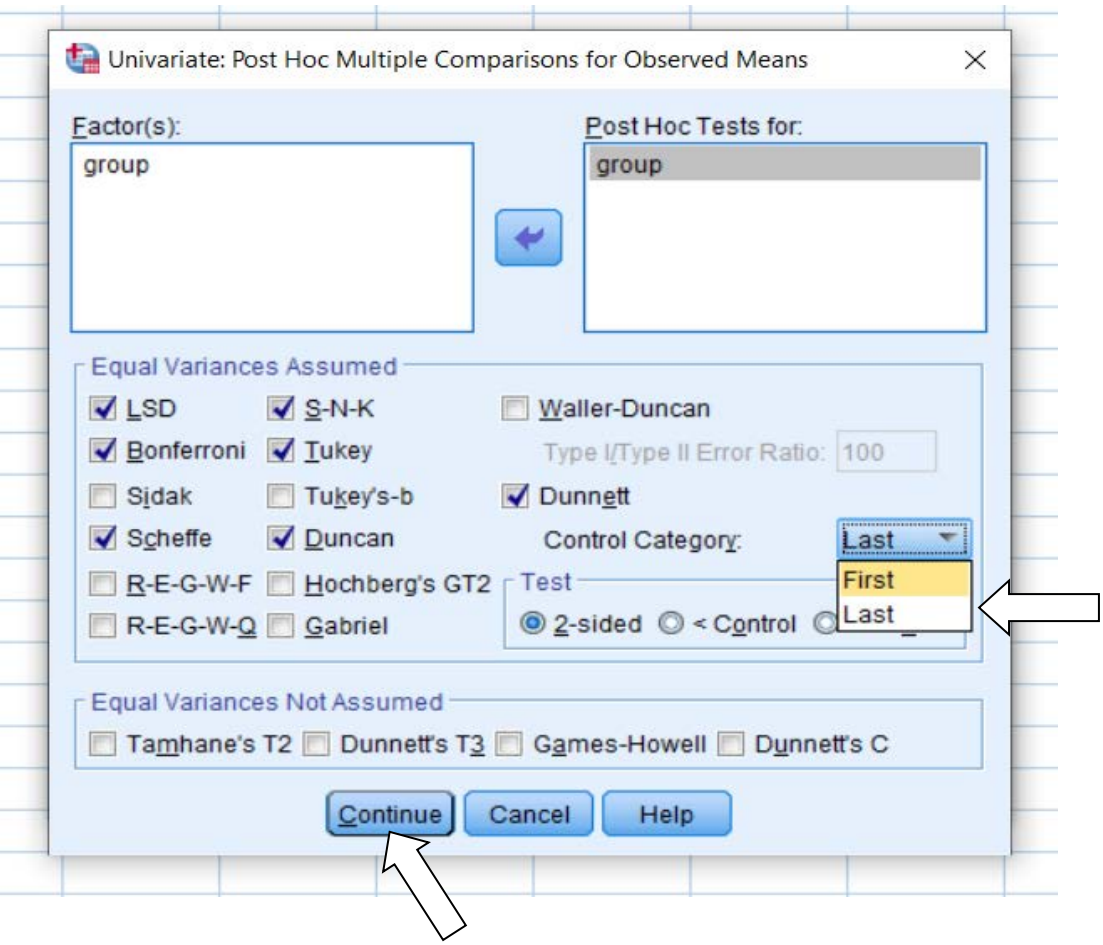
Tests of Between-Subjects Effects

Dependent Variable: recall

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	234.067[a]	2	117.033	9.046	.001
Intercept	3477.633	1	3477.633	268.812	.000
group	234.067	2	117.033	9.046	.001
Error	349.300	27	12.937		
Total	4061.000	30			
Corrected Total	583.367	29			

a. R Squared = .401 (Adjusted R Squared = .357)


Univariate: Post Hoc Multiple Comparisons for Observed Means
Factor(s):
group
Post Hoc Tests for:
Equal Variances Assumed
LSD
S-N-K
Waller-Duncan
Bonferroni
Tukey
Type I/Type II Error Ratio: 100
Sidak
Tukey's-b
Dunnett
Scheffe
Duncan
Control Category: Last
R-E-G-W-F
Hochberg's GT2
Test
R-E-G-W-Q
Gabriel
2-sided
< Control
> Control
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Continue
Cancel
Help


Univariate: Post Hoc Multiple Comparisons for Observed Means
Factor(s):
group
Post Hoc Tests for:
group
Equal Variances Assumed
LSD
S-N-K
Waller-Duncan
Bonferroni
Tukey
Type I/Type II Error Ratio: 100
Sidak
Tukey's-b
Dunnett
Scheffe
Duncan
Control Category: Last
First
Last
R-E-G-W-F
Hochberg's GT2
Test
R-E-G-W-Q
Gabriel
2-sided
< Control
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Continue
Cancel
Help

Click OK

Multiple Comparisons

Dependent Variable: recall

	(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	rhyming	imagery	-6.5000*	1.60854	.001	-10.4882	-2.5118
		intentional	-5.1000*	1.60854	.010	-9.0882	-1.1118
	imagery	rhyming	6.5000*	1.60854	.001	2.5118	10.4882
		intentional	1.4000	1.60854	.663	-2.5882	5.3882
	intentional	rhyming	5.1000*	1.60854	.010	1.1118	9.0882
		imagery	-1.4000	1.60854	.663	-5.3882	2.5882
Scheffe	rhyming	imagery	-6.5000*	1.60854	.002	-10.6662	-2.3338
		intentional	-5.1000*	1.60854	.014	-9.2662	-.9338
	imagery	rhyming	6.5000*	1.60854	.002	2.3338	10.6662
		intentional	1.4000	1.60854	.688	-2.7662	5.5662
	intentional	rhyming	5.1000*	1.60854	.014	.9338	9.2662
		imagery	-1.4000	1.60854	.688	-5.5662	2.7662
LSD	rhyming	imagery	-6.5000*	1.60854	.000	-9.8005	-3.1995
		intentional	-5.1000*	1.60854	.004	-8.4005	-1.7995
	imagery	rhyming	6.5000*	1.60854	.000	3.1995	9.8005
		intentional	1.4000	1.60854	.392	-1.9005	4.7005
	intentional	rhyming	5.1000*	1.60854	.004	1.7995	8.4005
		imagery	-1.4000	1.60854	.392	-4.7005	1.9005
Bonferroni	rhyming	imagery	-6.5000*	1.60854	.001	-10.6057	-2.3943
		intentional	-5.1000*	1.60854	.011	-9.2057	-.9943
	imagery	rhyming	6.5000*	1.60854	.001	2.3943	10.6057
		intentional	1.4000	1.60854	1.000	-2.7057	5.5057
	intentional	rhyming	5.1000*	1.60854	.011	.9943	9.2057
		imagery	-1.4000	1.60854	1.000	-5.5057	2.7057
Dunnett t (2-sided)[b]	imagery	rhyming	6.5000*	1.60854	.001	2.7466	10.2534
	intentional	rhyming	5.1000*	1.60854	.007	1.3466	8.8534

Based on observed means.
The error term is Mean Square(Error) = 12.937.

*. The mean difference is significant at the 0.05 level.

b. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Homogeneous Subsets

recall

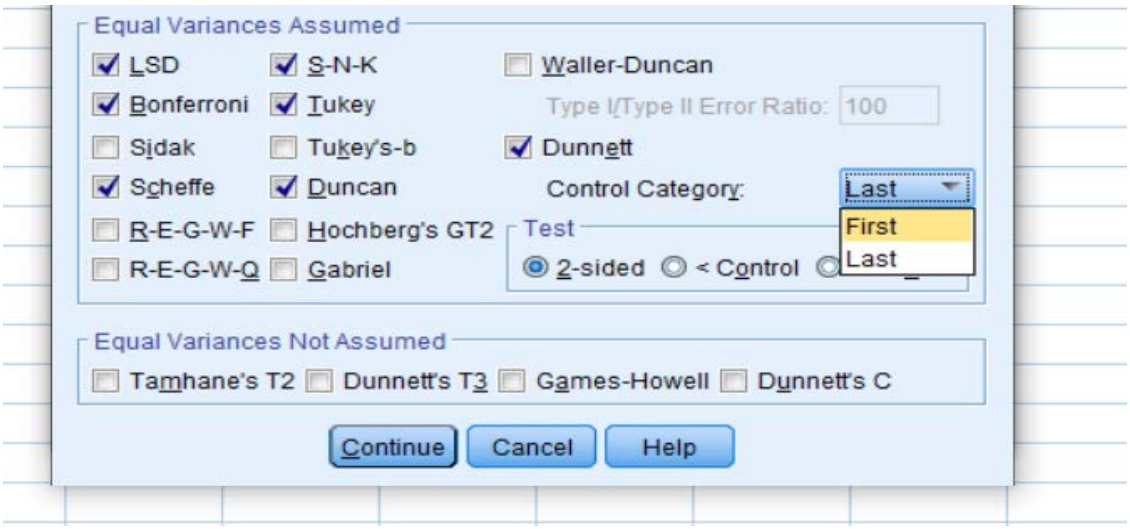
	group	N	Subset 1	Subset 2
Student-Newman-Keuls[a,b]	rhyming	10	6.9000	
	intentional	10		12.0000
	imagery	10		13.4000
	Sig.		1.000	.392
Tukey HSD[a,b]	rhyming	10	6.9000	
	intentional	10		12.0000
	imagery	10		13.4000
	Sig.		1.000	.663
Duncan[a,b]	rhyming	10	6.9000	
	intentional	10		12.0000
	imagery	10		13.4000
	Sig.		1.000	.392
Scheffe[a,b]	rhyming	10	6.9000	
	intentional	10		12.0000
	imagery	10		13.4000
	Sig.		1.000	.688

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 12.937.

a. Uses Harmonic Mean Sample Size = 10.000.

b. Alpha = 0.05.

Violation of Homogeneity of Variance: Several post hoc procedures assume homogeneity of variance. When this assumption is violated, post hoc procedures that do not assume homogeneity of variance must be used.



Games-Howell is designed for unequal variances and unequal sample sizes. It can be liberal when sample size is small and is recommended only when group sample sizes are greater than 5. It is recommended over Dunnett's C or T3 because it is only slightly liberal and more powerful. Tamhane's T2 is a conservative test.

Multiple Comparisons

Dependent Variable: recall

	(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	rhyming	imagery	-6.5000*	1.60854	.001	-10.4882	-2.5118
		intentional	-5.1000*	1.60854	.010	-9.0882	-1.1118
	imagery	rhyming	6.5000*	1.60854	.001	2.5118	10.4882
		intentional	1.4000	1.60854	.663	-2.5882	5.3882
	intentional	rhyming	5.1000*	1.60854	.010	1.1118	9.0882
		imagery	-1.4000	1.60854	.663	-5.3882	2.5882
Games-Howell	rhyming	imagery	-6.5000*	1.57515	.003	-10.6654	-2.3346
		intentional	-5.1000*	1.36178	.006	-8.6560	-1.5440
	imagery	rhyming	6.5000*	1.57515	.003	2.3346	10.6654
		intentional	1.4000	1.85113	.734	-3.3382	6.1382
	intentional	rhyming	5.1000*	1.36178	.006	1.5440	8.6560
		imagery	-1.4000	1.85113	.734	-6.1382	3.3382

Based on observed means.
The error term is Mean Square(Error) = 12.937.

*. The mean difference is significant at the 0.05 level.

Eysenck Results Example - One-Way ANOVA

The purpose of the study is to determine if level of processing affects recall of verbal material. A one-way analysis of variance (ANOVA) was conducted on the number of words correctly recalled with type of processing (rhyming, imagery, and intentional) as the independent variable. Prior to conducting the formal analysis of variance procedures, I examined the data to ensure that there were no influential observations and that the ANOVA assumptions seemed plausible. The standardized residuals were first examined in order to identify possibly outlying values on the number of words recalled. One observation in the data set had a standardized residual value larger than 2.5. However, a sensitivity study showed that deleting this score did not change important study results.[1] Inspection of the study data and Kolmogorov-Smirnov tests did not indicate any serious violations of the normality assumption. Levene's test supported the assumption of homogeneity of variance, F (2, 27) = 2.4, $p > .05$. Finally, the independence assumption seems reasonable, since the treatments were individually administered.

The results of the ANOVA are shown in Table 1. The effect of processing on recall is statistically significant, F (2, 27) = 9.0, $p < .01$. Processing accounted for approximately 40% of the variance in recall. In addition, the overall strength of relationship is substantial, as the partial eta squared of .40 is greater than the .14 cutoff (Cohen, 1977).

Table 1

Analysis of Variance for Processing

Source	*SS*	*df*	*MS*	*F*	Partial η^2	*p*
Processing	234.07	2	117.03	9.05	.401	< .01
Error	349.30	27	12.94			
Total	583.37	29				

[1] The ANOVA with the outlying case removed indicated a significant effect of processing level, F (2, 26) = 9.56, $p < .01$, accounting for approximately 42% of the variance in words recalled.

Comparisons of means using the Tukey HSD approach are summarized in Table 2. These comparisons indicate that average recall is similar in the imagery and intentional conditions. However, average recall in both imagery and intentional groups is significantly higher than average recall in the rhyming group (see Table 3).

Table 2

Differences in Means for the Number of Words Correctly Recalled

Contrast	Mean Difference	*SE*	95% Confidence Interval
Imagery versus rhyming	6.50*	1.61	2.51, 10.49
Intentional versus rhyming	5.10*	1.61	1.11, 9.09
Imagery versus intentional	1.40	1.61	-2.59, 5.39

* Differences are significant at the .05 level using the Tukey HSD procedure.

Table 3

Number of Words Recalled as a Function of Processing Level

Processing Level	*n*	*M*	*SD*
Rhyming	10	6.90	2.13
Imagery	10	13.40	4.50
Intentional	10	12.00	3.74

Note. The maximum possible score = 27.

General Linear Model

Fixed Effects One-Way ANOVA can be presented as a linear model.

General Linear Model (Fixed Effects One-Way ANOVA): $y_{ik} = \mu + \alpha_k + \varepsilon_{ik}$,

where y_{ik} is the score for the *i*th participant in the *k*th group; μ is the grand mean for all participants; $\alpha_k = \mu_k - \mu$ is the treatment effect for the *k*th treatment/group; and ε_{ik} is random error.

The linear model indicates that an individual's score is conceived as a linear composite of three components:

1. Grand Mean (μ) – it is a constant and reflects the overall average of scores.
2. α_k is the effect of treatment *k* – it is a constant for all scores in group *k* – it reflects the increase or decrease in these scores that is associated with treatment *k* (i.e., $\alpha_k = \mu_k - \mu$).
3. ε_{ik} is the error or residual for the score y_{ik} - it is the residual of the score y_{ik} when predicted from μ and α_k. If you rearrange the linear model: $\varepsilon_{ik} = y_{ik} - \mu - \alpha_k$. Because $\alpha_k = \mu_k - \mu$: $\varepsilon_{ik} = y_{ik} - \mu - (\mu_k - \mu)$. Thus, $\varepsilon_{ik} = y_{ik} - \mu_k$. ε_{ik} is normally distributed within each group, independent, and have same variance for each group treatment.

The grand mean is typically not of interest in ANOVA. The main interest is in the effects (α_k).

The above population parameters must be estimated:

μ: $\hat{\mu} = \bar{x}$

$\alpha_k = \mu_k - \mu$: $\hat{\alpha}_k = \bar{x}_k - \bar{x}$

[Recall: $SS_b = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$]

$\varepsilon_{ik} = y_{ik} - \mu_k$: $\hat{\varepsilon}_{ik} = y_{ik} - \bar{x}_k$

[Recall: $SS_w = \sum_1(x_{i1} - \bar{x}_1)^2 + \sum_2(x_{i2} - \bar{x}_2)^2 + \cdots + \sum_k(x_{ik} - \bar{x}_k)^2$]

Example 1: No treatment effect and no random error

T_1	T_2	T_3
20	20	20
20	20	20
20	20	20
20	20	20

$\hat{\mu} = \bar{x}$ $\quad$ $\hat{\alpha}_k = \bar{x}_k - \bar{x}$ $\quad$ $\hat{\varepsilon}_{ik} = y_{ik} - \bar{x}_k$

$\hat{\mu} = 20$ $\quad$ $\hat{\alpha}_1 = 20 - 20 = 0$ $\quad$ $\hat{\varepsilon}_{11} = 20 - 20 = 0$ [for all in group 1]

$\hat{\alpha}_2 = 20 - 20 = 0$ $\quad$ $\hat{\varepsilon}_{12} = 20 - 20 = 0$ [for all in group 2]

$\hat{\alpha}_3 = 20 - 20 = 0$ $\quad$ $\hat{\varepsilon}_{13} = 20 - 20 = 0$ [for all in group 3]

$y_{11} = 20 + 0 + 0,$
$y_{11} = 20$ [for every participant]
$y_{ik} = \mu$

Example 2: Treatment effect but no random error

T_1	T_2	T_3
19	17	24
19	17	24
19	17	24
19	17	24

$\hat{\mu} = \bar{x}$ $\quad$ $\hat{\alpha}_k = \bar{x}_k - \bar{x}$ $\quad$ $\hat{\varepsilon}_{ik} = y_{ik} - \bar{x}_k$

$\hat{\mu} = 20$ $\quad$ $\hat{\alpha}_1 = 19 - 20 = -1$ $\quad$ $\hat{\varepsilon}_{11} = 19 - 19 = 0$ [for all in group 1]

$\hat{\alpha}_2 = 17 - 20 = -3$ $\quad$ $\hat{\varepsilon}_{12} = 17 - 17 = 0$ [for all in group 2]

$\hat{\alpha}_3 = 24 - 20 = 4$ $\quad$ $\hat{\varepsilon}_{13} = 24 - 24 = 0$ [for all in group 3]

$y_{11} = 20 - 1 + 0,$
$y_{11} = 20 - 1 = 19$ [for all in group 1]
$y_{12} = 20 - 3 + 0,$
$y_{12} = 20 - 3 = 17$ [for all in group 2]
$y_{13} = 20 + 4 + 0,$
$y_{13} = 20 + 4 = 24$ [for all in group 3]

$y_{ik} = \mu + \alpha_k$

Example 3: Treatment effect and random error (values slightly different from the book)

T_1	T_2	T_3
18	18	23
24	19	22
18	15	27
16	16	24

$\hat{\mu} = \bar{x}$ $\quad \hat{\alpha}_k = \bar{x}_k - \bar{x}$ $\quad \hat{\varepsilon}_{ik} = y_{ik} - \bar{x}_k$

$\hat{\mu} = 20$ $\quad \hat{\alpha}_1 = 19 - 20 = -1$ $\quad \hat{\varepsilon}_{11} = 18 - 19 = -1,\ \hat{\varepsilon}_{21} = 24 - 19 = 5$, etc.

$\hat{\alpha}_2 = 17 - 20 = -3$ $\quad \hat{\varepsilon}_{12} = 18 - 17 = 1\ ,\ \hat{\varepsilon}_{22} = 19 - 17 = 2$, etc.

$\hat{\alpha}_3 = 24 - 20 = 4$ $\quad \hat{\varepsilon}_{13} = 23 - 24 = -1,\ \hat{\varepsilon}_{23} = 22 - 24 = -2$, etc.

$y_{11} = 20 - 1 - 1 = 18$, $y_{21} = 20 - 1 + 5 = 24$, etc.
$y_{12} = 20 - 3 + 1 = 18$, $y_{22} = 20 - 3 + 2 = 19$, etc.
$y_{13} = 20 + 4 - 1 = 23$, $y_{23} = 20 + 4 - 2 = 22$, etc.

$y_{ik} = \mu + \alpha_k + e_{ik}$