In Neil Salkind (Ed.), ***Encyclopedia of Research Design.***
Thousand Oaks, CA: Sage. 2010

Post-Hoc Comparisons

Lynne J. Williams · Hervé Abdi

1 Introduction

The F test used in analysis of variance (ANOVA) is called an *omnibus* test because it can detect only the presence or the absence of a global effect of the independent variable on the dependent variable. However, in general we want to draw *specific* conclusions from the results of an experiment. Specific conclusions are derived from *focused comparisons* which are, mostly, implemented as *contrasts* between experimental conditions. When these comparisons are decided *after* the data are collected, they are called *post-hoc* or *a posteriori* analyses. These comparisons are performed after an ANOVA has been performed on the experimental data of interest. In the ANOVA framework, post-hoc analyses take two general forms: *(1)* comparisons that involves all possible contrasts, and *(2)* comparisons which are restricted to comparing pairs of means (called *pairwise* comparisons).

Lynne J. Williams
The University of Toronto Scarborough

Hervé Abdi
The University of Texas at Dallas

Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/∼herve

2 Notations

The experimental design is a one factor ANOVA. The total number of observations is denoted N, with S denoting the number of observations per group. The number of groups is denoted A, a given group is labelled with the letter a, the group means are denoted M_{a+} and the grand mean is denoted M_{++}. A contrast is denoted ψ_a, the contrast coefficients (*aka* contrast weights) are denoted C_a. The α-level per comparison is denoted $\alpha[PC]$ and the α-level per family of comparisons is denoted $\alpha[PF]$.

3 Planned versus Post-hoc comparisons

Planned (or *a priori*) comparisons are selected before running the experiment. In general, they correspond to the research hypotheses. Because these comparisons are planned, they are usually few in number. In order to avoid an inflation of the Type I error (*i.e.,* declaring significant an effect when it is not), the p values of these comparisons are corrected with the standard Bonferonni or Šidàk approaches.

In contrast, post-hoc (or *a posteriori*) comparisons are decided after the experiment has been run and analyzed. The aim of post-hoc comparisons is to make sure that (unexpected) patterns seen in the results are reliable. This implies that the actual family of comparisons consists of all possible comparisons, even if they are not explicitly tested.

4 What is a contrast?

A contrast is a prediction precise enough to be translated into a set of numbers called contrast coefficients or contrast weights (denoted C_a) which express this prediction. For convenience, contrast coefficients have a mean equal to zero. The correlation between the contrast coefficients and the values of the group means quantifies the similarity between the prediction and the results.

All contrasts are evaluated using the same general procedure. A contrast is formalized as a set of contrast coefficients which represent the predicted pattern of experimental results. For example, if we have 4 groups and we predict that the first group should have better performance than the other 3 groups and that these remaining 3 groups are equivalent, we get the following contrast:

$$\begin{array}{rccccc} & C_1 & C_2 & C_3 & C_4 & \text{Mean} \\ \psi = & 3 & -1 & -1 & -1 & 0 \end{array} \tag{1}$$

When the data have been collected and the experimental means and mean squares have been computed, the next step it to evaluate the if the contrast correlates significantly with the mean values. In order to do so a specific F ratio (denoted F_ψ) is computed. Finally, the probability associated with F_ψ is evaluated. (see entry *Contrast Analysis*, this volume). If this probability is small enough, the contrast is considered “significant.”

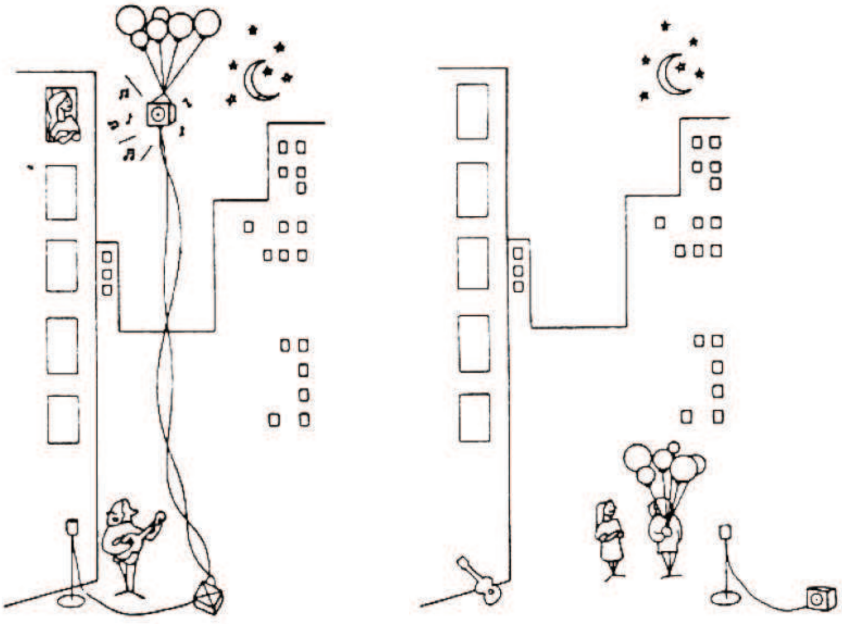
5 An example

This example is a fictitious replication of Bransford and Johnson’s (1972) study examining the effect of context on memory. In this study, Bransford and Johnson read the following paragraph to their participants:

> If the balloons popped, the sound would not be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.

To show the importance of context on memory for texts, the authors used the following four experimental conditions:

1) NO CONTEXT: Participants listened to the passage and tried to remember it,
2) APPROPRIATE CONTEXT BEFORE: Participants were provided with an appropriate context (see Figure 1-a) in the form of a picture and then listened to the passage,



(a) (b)

Figure 1: (a) Appropriate context and (b) partial context for Bransford and Johnson's (1972) experiment

3) APPROPRIATE CONTEXT AFTER: Participants first listened to the passage and then were provided with an appropriate context (see Figure 1-a) in the form of a picture, and
4) PARTIAL CONTEXT: Participants were provided with a context that did not allow them to make sense of the text when they listened to it (see Figure 1-b).

The dependent variable is the number of "ideas" recalled. The data are shown in Table 1 and the results of the ANOVA are shown in Table 2.

Table 1: Number of ideas recalled for 20 subjects in the fictitious replication of Bransford and Johnson (1972). The experimental conditions are ordered from the smallest to the largest mean.

	Context before	Partial context	Context after	No context
	5	5	2	3
	9	4	4	3
	8	3	5	2
	4	5	4	4
	9	4	1	3
$\sum$	35	21	16	15
M_{a+}	7	4.2	3.2	3
$M_{++} = 4.35$				
$S = 5$				

Table 2: ANOVA results for the Bransford and Johnson (1972) experiment.

Source	df	SS	MS	F	$p(F)$
Between	3	50.90	16.97	7.22	.00288
Error	16	37.60	2.35		
Total	19	88.50			

6 Post-hoc comparisons

6.1 Scheffé test: All possible contrasts

When a post-hoc comparison is performed, the family of comparisons consists of all possible comparisons. This number is, in general, very large even when dealing with a small number of experimental conditions, and this large number practically precludes using a Bonferonni approach to correct for multiple testings because this large number of comparisons would make the correction much too conservative. The Scheffé test is also conservative, but less than the Bonferonni approach. The idea behind the Scheffé's test begins with the omnibus F testing the null hypothesis that all population means are equal, and this null hypothesis in turn implies that all possible contrasts are also zero. Note that testing the contrast with the largest sum is equivalent to testing all possible contrasts at once (because a failure to reject the null hypothesis with the largest contrast implies a failure to reject the null hypothesis for any smaller contrast).

Table 3: Post-hoc or *a posteriori* contrasts for the Bransford and Johnson (1972) experiment.

	Context before	Partial context	Context after	No context
ψ_1	1	1	1	-3
ψ_2	0	0	1	-1
ψ_3	3	-1	-1	-1
ψ_4	1	-1	0	0

Suppose that Bransford and Johnson wanted to test the following four contrasts *after* having collected their data: *(1)* no context *vs.* all other conditions, *(2)* context after *vs.* no context, *(3)* context before *vs.* all other conditions, and *(4)* context before *vs.* partial context. The contrast weights (*i.e., C_a*-s) are shown in Table 3. The F_ψ ratio for the maximum contrast ψ_a is equal to

$$F_\psi = \frac{SS_\psi}{MS_{\text{error}}} = \frac{SS_{\text{between}}}{MS_{\text{error}}} = \frac{(A-1)MS_{\text{between}}}{MS_{\text{error}}} = (A-1)F_{\text{omnibus}} \; . \quad (2)$$

To have the Sheffé test contrast equivalent to the omnibus test, we need to reject the null hypothesis under the same conditions as the omnibus test. To reject the omnibus null hypothesis, F_{omnibus} must be greater than or equal to $F_{\text{critical, omnibus}}$. Because F_ψ is equal to $(A-1)F_{\text{omnibus}}$, then, we reject the null hypothesis when

$$(A-1)F_{\text{omnibus}} > (A-1)F_{\text{critical, omnibus}} \; , \quad (3)$$

and, therefore,

$$F_\psi > (A-1)F_{\text{critical, omnibus}} \; . \quad (4)$$

Consequently, the critical value to test all possible contrasts is

$$F_{\text{critical, Scheffé}} = (A-1)F_{\text{critical, omnibus}} \quad (5)$$

with

$$\nu_1 = A-1 \quad \text{and} \quad \nu_2 = A(S-1) \quad (6)$$

degrees of freedom. For our example, $F_{\text{critical, Scheffé}}$ is equal to

$$F_{\text{critical, Scheffé}} = (A-1)F_{\text{critical, omnibus}} = (4-1)\times 3.24 = 9.72 \quad (7)$$

Table 4: The results of the Scheffé test for the contrasts in Table 3.

	SS_{ψ}	F_{ψ}	$p(F_{\text{Scheffé}})$	Decision
ψ_1	12.15	5.17	.2016	*ns*
ψ_2	0.10	0.04	$F < 1$	*ns*
ψ_3	46.82	19.92	.0040	reject H_0
ψ_4	19.60	8.34	.0748	*ns*

with $\nu_1 = A - 1 = 4 - 1 = 3$ and $\nu_2 = A(S - 1) = 4(5 - 1) = 16$. An alternative approach is to correct the value of F_{ψ} by dividing it by $(A - 1)$ and evaluating its probability according to a Fisher distribution with $(A - 1)$ and $A(S - 1)$ degrees of freedom (*i.e.,* the number of degrees of freedom of the omnibus test).

The results of the Scheffé test for the contrasts in Table 3 are shown in Table 4. The third contrast, ψ_3, context before *vs.* all other contexts is the only significant contrast ($F_{\psi_3} = 19.92$, $p < .01$). This shows that memorization is facilitated only when the context information is presented before learning.

6.2 Pairwise Comparisons

6.2.1 Honestly Significant Difference (HSD) test

The HSD test is a conservative test that uses Student's q statistics and the Studentized range distribution. The range is the number of groups falling between groups (including the groups under consideration). For example, in the Bransford and Johnson experiment, the range of means between the largest and the smallest means is equal to 4 because there are four means going from 7.0 to 3.0 (*i.e.,* 7.0, 4.2, 3.2, and 3.0).

To begin the HSD test, all pairwise differences are computed. These differences are shown in Table 5 for the Bransford and Johnson example. If the difference between two means is greater than the honestly significant difference, then the two conditions are significantly different at the

Table 5: HSD. Pairwise differences between means for the Bransford and Johnson example. Differences greater than 2.37 are significant at the $p = .05$ level and are indicated by *. Differences greater than 3.59 are significant at the $p = .01$ level and are indicated by **.

	Context before	Partial context	Context after	No context
	$M_{1+} = 7.0$	$M_{2+} = 4.2$	$M_{3+} = 3.2$	$M_{4+} = 3.0$
$M_{1+} = 7.0$		2.8**	3.8**	4.0**
$M_{2+} = 4.2$			1.0	1.2
$M_{3+} = 3.2$				0.2

chosen alpha level. The HSD is computed as

$$|M_{a+} - M_{a'+}| > \text{HSD} = q_{A,\alpha}\sqrt{\frac{1}{2}MS_{\text{error}}\left(\frac{1}{S_a} + \frac{1}{S_{a'}}\right)} \tag{8}$$

with a range equal to A and $\nu = N - A$ degrees of freedom (for more details see entry *Tukey's Honestly Significant Difference*, a table of critical q values is provided in the *Newman-Keuls* entry). For the Bransford and Johnson example, the HSD values are:

$$\text{HSD}_{A,\alpha} = \text{HSD}_{4,\alpha=.05} = 2.37$$

for $p = .05$, and

$$\text{HSD}_{A,\alpha} = \text{HSD}_{4,\alpha=.01} = 3.59$$

for $p = .01$ with a range equal to $A = 4$ and $\nu = N - A = 20 - 4 = 16$ degrees of freedom. Using HSD, there are significant pairwise difference between the "context before" condition and the "partial context," "context after", and "no context" conditions.

6.2.2 Newman-Keuls

The Newman-Keuls test is a sequential test in which q_{critical} depends on the range of each pair of means. The Newman-Keuls test is the most popular *a posteriori* pairwise comparison test.

The Newman-Keuls test starts by computing the value of Student's q_{observed} for the largest pairwise difference.

$$q_{\text{observed}} = \frac{M_{a+} - M_{a'+}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} \tag{9}$$

(where a and a' are respectively the smallest and the largest means). This value is then evaluated with a Studentized distribution with a range of A and with $\nu = N - K$.

In or example, for the greatest pairwise difference, the q_{observed} is equal to (*cf.* Equation 9):

$$\begin{aligned} q_{\text{observed}} &= \frac{M_{1+} - M_{4+}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} \\ &= \frac{4.0}{\sqrt{\frac{2.35}{5}}} \; . \\ &= 5.83 \; . \end{aligned} \tag{10}$$

This value of $q_{\text{observed}} = 5.83$ needs to be compared to a value of q_{critical} for a range of 4 and for $\nu = N - A = 20 - 4 = 16$ degrees of freedom. From the table of critical values for the Studentized range we find that $q_{\text{critical}} = 4.05$ for $p = .05$ and 5.19 for $p = .01$, because $q_{\text{observed}} = 5.83$ is larger than $q_{\text{critical}} = 5.59$ we conclude that the "context before" and the "no context" conditions are significantly different at $p < .01$.

If the null hypothesis cannot be rejected for the largest difference, the test stops here. If the null hypothesis is rejected for the largest difference, the two differences with a range of $A - 1$ are examined. If the differences are non-significant, all other differences contained in that difference are not tested. If the differences are significant, then the procedure is reiterated until all means have been tested or have been declared non-significant by implication. This is shown for the Bransford and Johnson example by following the arrows in Figure 2. The results of the Newman-Keuls test are shown in Table 6. Again, we see significant differences be-

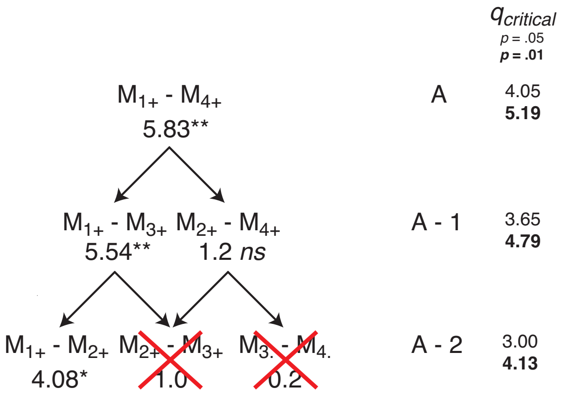


Figure 2: Structure of implication of the pairwise comparisons when $A = 4$ for the Newman-Keuls test. Means are numbered from 1 (the largest) to 4 (the smallest). The pairwise comparisons implied by another one are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise compairison, then all the compairisons included in it are omitted from testing.

Table 6: Newman-Keuls. Pairwise differences between means for the Bransford and Johnson example. Differences significant at the $p = .05$ level and are indicated by *. Differences significant at the $p = .01$ level and are indicated by **.

	Context before $M_{1+} = 7.0$	Partial context $M_{2+} = 4.2$	Context after $M_{3+} = 3.2$	No context $M_{4+} = 3.0$
$M_{1+} = 7.0$		2.8*	3.8**	4.0**
$M_{2+} = 4.2$			1.0	1.2
$M_{3+} = 3.2$				0.2

tween the “context before” condition and all other conditions. (For more information see entry *Newman-Keuls* , this volume).

6.2.3 Duncan test

The Duncan test follows the same general sequential pattern as the Newman-Keuls test. The difference is that the critical values for the test come from a different table. The only difference between the two tests is that Duncan test uses the Fisher F distribution with a Šidàk correction. The Šidàk correction is computed using the Šidàk inequality

$$\alpha[PC] = 1 - (1 - \alpha[PF])^{\frac{1}{\text{Range}-1}} \qquad (11)$$

Table 7: Duncan test. F values for the Bransford and Johnson example. Significant F values at the $\alpha[PF] = .05$ level are indicated by *, and at the $\alpha[PF] = .01$ level by **.

	Context before $M_{1+} = 7.0$	Partial context $M_{2+} = 4.2$	Context after $M_{3+} = 3.2$	No context $M_{4+} = 3.0$
$M_{1+} = 7.0$		8.3403*	15.3619**	17.0213**
$M_{2+} = 4.2$			1.0638	1.5320
$M_{3+} = 3.2$				0.0425

where $\alpha[PC]$ is the α-level per comparison and $\alpha[PF]$ is the α-level per family of comparisons. For the Bransford and Johnson example with $A = 4$ conditions,

$$\begin{aligned}\alpha[PC] &= 1 - (1 - \alpha[PF])^{\frac{1}{\text{Range}-1}} \\ &= 1 - (1 - .05)^{\frac{1}{3}} \\ &= 0.0170 \ . \end{aligned} \tag{12}$$

The q-values computed for the Newman-Keuls test can be changed into F-values using the following formula:

$$F_{\text{range}} = \frac{q^2}{2} \ . \tag{13}$$

For the Bransford and Johnson example, this results in the F values shown in Table 7.

For the first step, the critical F-values are 7.10 for an $\alpha[PF]$ of .05 and 12.45 for an $\alpha[PF]$ of .01 with Range $= A = 4$ and $\nu_2 = N - A = 16$ degrees of freedom. Then, the same recursive procedure as the Newman-Keuls test is then followed and results in the pattern of results shown in Table 7. (see entry *Duncan's multiple range test* for more information).

7 Related entries

Analysis of variance, Bonferroni procedure, Contrasts, Duncan's multiple range test, Fisher's least significant difference (LSD) test, Multiple comparisons, Newman-Keuls test, Pairwise comparisons, Post-hoc comparisons, Scheffe's test, Tukey's Honestly Significant Difference (HSD) Test

Further readings

1. Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N.J. Salkind (Ed.): *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): Sage. pp. 103–107.
2. Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental Design and Analysis for Psychology*. Oxford: Oxford University Press.
3. Hochberg, Y., & Tamhane, A.C. (1987). *Multiple comparison procedures.* New York: Wiley.
4. Jaccard, J., Becker, M.A., Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 94,* 589–596.
5. Miller, R.G. (1981). *Simultaneous statistical inferences.* Berlin: Springer Verlag.
6. Rosenthal, R., Rosnow, R.L., Rubin, D.B. (2000). *Contrasts and effect sizes in behavioral research: a correlational approach*. Cambridge: Cambridge Universit Press.