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

Newman-Keuls Test and Tukey Test

Hervé Abdi · Lynne J. Williams

1 Pairwise Comparisons

An analysis of variance (ANOVA) indicates if several means come from the same population. Such a procedure is called an *omnibus* test, because it tests the whole set of means at once (*omnibus* means "for all" in latin). In an ANOVA omnibus test, a significant result indicates that at least two groups differ from each other but it does not identify the groups that differ. So an ANOVA is generally followed by an analysis whose goal is to identify the pattern of differences in the results. This analysis is often performed by evaluating all the pairs of means in order to decide which ones show a significant difference. In a general framework, this approach, called pairwise comparisons, is a specific case of "*a posteriori* contrast analysis" (see entry on contrast), but it is specific enough to be studied in itself. Two of the most common methods of pairwise comparisons are the Tukey test and the Newman-Keuls test. Both tests are based

Hervé Abdi
The University of Texas at Dallas

Lynne J. Williams
The University of Toronto Scarborough

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

on the "Studentized range" or "Student's q". They differ in that the Newman-Keuls test is a sequential test designed to have more power than the Tukey test.

Choosing between the Tukey and Newman-Keuls tests is not straightforward and there is no consensus on this issue. The Newman-Keuls test is most frequently used in psychology, while the Tukey test is most commonly used in other disciplines. An advantage of the Tukey test is to keep the level of the Type I error (i.e., finding a difference when none exists) equal to the chosen alpha level (e.g., $\alpha = .05$ or $\alpha = .01$). An additional advantage of the Tukey test is to allow the computation of confidence intervals for the differences between the means. Although the Newman-Keuls test has more power than the Tukey test, the exact value of the probability of making a Type I error of the Newman-Keuls test cannot be computed due to the sequential nature of this test. In addition, because the criterion changes for each level of the Newman-Keuls test, confidence intervals cannot be computed around the differences between means. Therefore, selecting whether to use the Tukey or Newman-Keuls test depends upon whether or not additional power is required to detect significant differences between means.

1.1 Studentized Range and Student's q

Both the Tukey and Newman-Keuls tests use a sampling distribution derived by Gosset (who was working for Guiness and decided to publish under the pseudonym of "Student" because of Guiness' confidentiality policy). This distribution, called the Studentized Range or Student's q, is similar to a t-distribution. It corresponds to the sampling distribution of the *largest* difference between two means coming from a set of A means (when $A = 2$ the q distribution corresponds to the usual Student's t).

In practice, one computes a criterion denoted q_{observed} which evaluates the difference between the means of two groups. This criterion is computed as:

$$q_{\text{observed}} = \frac{M_{i.} - M_{j.}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} \tag{1}$$

where $M_{i.}$ and $M_{j.}$ are the group means being compared, MS_{error} is the mean square error from the previously computed ANOVA (i.e., this is the mean square used for the denominator of the omnibus F ratio), and S is the number of observations per group (the groups are assumed to be of equal size).

Once the q_{observed} is computed, it is then compared with a q_{critical} value from a table of critical values (see appendix). The value of q_{critical} depends upon the α-level, the degrees of freedom $\nu = N - K$ where N is the total number of participants and K is the number of groups, and a parameter R, which is the number of means being tested. For example, in a group of $K = 5$ means ordered from smallest to largest,

$$M_{1.} < M_{2.} < M_{3.} < M_{4.} < M_{5.}$$

$R = 5$ when comparing $M_{5.}$ to $M_{1.}$; however, $R = 3$ when comparing $M_{3.}$ to $M_{1.}$.

1.1.1 F-range

Some statistics textbooks refer to a pseudo-F distribution called the "F-range" or "F_{range}", rather than the Studentized q distribution. The F_{range} can be easily computed from q using the following formula:

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

1.2 Tukey Test

For the Tukey test, q_{observed} (see Equation 1) is computed between any pair of means that need to be tested. Then, q_{critical} is determined using $R = \textit{total number of means}$. The q_{critical} is the same for all pairwise comparisons. Using the previous example, $R = 5$ for all comparisons.

1.3 Newman-Keuls Test

The Newman-Keuls test is similar to the Tukey test, except that the Newman-Keuls test is a sequential test in which q_{critical} depends

on the range of each pair of means. To facilitate the exposition, we suppose that the means are ordered from the smallest to the largest. Hence $M_{1.}$ is the smallest mean and $M_{A.}$ is the largest mean.

The Newman-Keuls test starts exactly like the Tukey test. The largest difference between two means is selected. The range of this difference is $R = A$. A q_{observed} is computed using Equation 1 and that value is compared to the critical value, q_{critical}, in the critical values table using α, ν, and R. The null hypothesis can be rejected if q_{observed} is greater than q_{critical}. If the null hypothesis cannot be rejected, the test stops here, because not rejecting the null hypothesis for the largest difference implies not rejecting the null hypothesis for any other difference.

If the null hypothesis is rejected for the largest difference, the two differences with a range of $A-1$ are examined. These means will be tested with $R = A-1$. When the null hypothesis for a given pair of means cannot be rejected, none of the differences included in that difference will be tested. If the null hypothesis is rejected, then the procedure is reiterated for a range of $A-2$ (i.e., $R = A-2$). The procedure is reiterated until all means have been tested or have been declared non-significant by implication.

It takes some experience to determine which comparisons are implied by other comparisons. Figure 1 describes the structure of implication for a set of 5 means numbered from 1 (the smallest) to 5 (the largest). The pairwise comparisons implied by another comparison are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise comparison, then all the comparisons included in it are crossed out so that the are not tested.

2 An Example

An example will help describe the use of the Tukey and Newman-Keuls tests and Figure 1. We will use the results of a (fictitious) replication of a classic experiment on eyewitness testimony by Loftus and Palmer (1974). This experiment tested the influence of question wording on the answers given by eyewitnesses. The authors presented

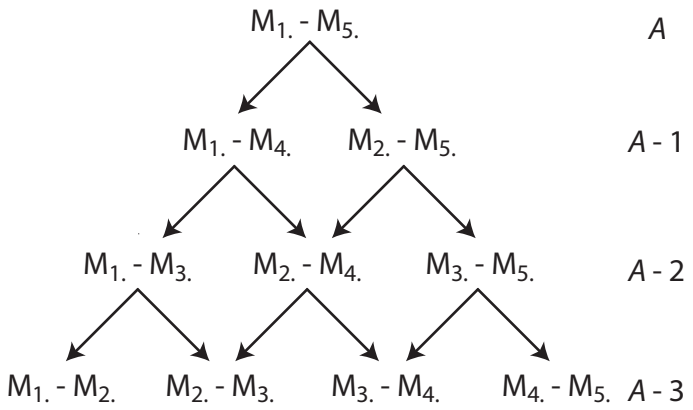

Figure 1 Structure of implication of the pairwise comparisons when $A = 5$ for the Newman-Keuls test. Means are numbered from 1 (the smallest) to 5 (the largest). The pairwise comparisons implied by another one are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise comparison, then all the comparisons included in it can be crossed out in order to omit them from testing.

a film of a multiple car accident to their participants. After seeing the film, participants were asked to answer a number of specific questions about the accident. Among the questions, one question about the speed of the car was presented in five different versions:

- HIT: About how fast were the cars going when they *hit* each other?
- SMASH: About how fast were the cars going when they *smashed* into each other?
- COLLIDE: About how fast were the cars going when they *collided* with each other?
- BUMP: About how fast were the cars going when they *bumped* into each other?
- CONTACT: About how fast were the cars going when they *contacted* each other?

In our replication we used 50 participants (10 in each group); their responses are given in Table 1.

2.1 Tukey test

For the Tukey test, the q_{observed} are computed between every pair of means using Equation 1. For example, taking into account that the

Table 1 A set of data to illustrate the Tukey and Newman-Keuls tests.

	Experimental Group				
	Contact	Hit	Bump	Collide	Smash
	21	23	35	44	39
	20	30	35	40	44
	26	34	52	33	51
	46	51	29	45	47
	35	20	54	45	50
	13	38	32	30	45
	41	34	30	46	39
	30	44	42	34	51
	42	41	50	49	39
	26	35	21	44	55
	$M_{1.}$	$M_{2.}$	${}_{M}3.$	${}_{M}4.$	$M_{5.}$
$M_{a.}$	30.00	35.00	38.00	41.00	46.00

$S = 10; \quad MS_{\text{error}} = 80.00$.

MS_{error} from the previously calculated ANOVA is 80.00, the value of q_{observed} for the difference between $M_{1.}$ and $M_{2.}$ (i.e., "contact" and "hit") is equal to:

$$q_{\text{observed}} = \frac{M_{1.} - M_{2.}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}}$$

$$= \frac{35.00 - 30.00}{\sqrt{80.00\left(\frac{1}{10}\right)}}$$

$$= \frac{5}{\sqrt{8}}$$

$$= 1.77$$

The values of q_{observed} are shown in Table 2. With Tukey's approach, each q_{observed} is declared significant at the $\alpha = .05$ level (or the

Table 2 Absolute values of q_{observed} for the data from Table 1. For the Tukey test, q_{observed} is significant at $\alpha = .05$ (or at the $\alpha = .01$ level), if q_{observed} is larger than $q_{\text{critical}} = 4.04$ ($q_{\text{critical}} = 4.93$).

	Experimental Group				
	$M_{1.}$ Contact 30	$M_{2.}$ Hit 1 35	$M_{3.}$ Bump 38	$M_{4.}$ Collide 41	$M_{5.}$ Smash 46
$M_{1.} = 30$ Contact	0	$1.77ns$	$2.83ns$	$3.89ns$	5.66^{**}
$M_{2.} = 35$ Hit		0	$1.06ns$	$2.12ns$	$3.89ns$
$M_{3.} = 38$ Bump			0	$1.06ns$	$2.83ns$
$M_{4.} = 41$ Collide				0	$1.77ns$
$M_{5.} = 46$ Smash					0

$*p < .05, \quad ** \; p < .01$

Table 3 Presentation of the results of the Tukey test for the data from Table 1.

	Experimental Group				
	$M_{1.}$ Contact 30	$M_{2.}$ Hit 1 35	$M_{3.}$ Bump 38	$M_{4.}$ Collide 41	$M_{5.}$ Smash 46
$M_{1.} = 30$ Contact	0	5.00 *ns*	8.00 *ns*	11.00 *ns*	16.00^{**}
$M_{2.} = 35$ Hit		0	3.00 *ns*	6.00 *ns*	11.00 *ns*
$M_{3.} = 38$ Bump			0	3.00 *ns*	8.00 *ns*
$M_{4.} = 41$ Collide				0	5.00 *ns*
$M_{5.} = 46$ Smash					0

$*p < .05, \quad ** \; p < .01$

$\alpha = .01$ level) if it is larger than the critical value obtained for this alpha level from the table with $R = 5$ and $\nu = N - K = 45$ degrees of freedom (45 is not in the table so 40 is used instead). The $q_{\text{critical (5), }\alpha=.05}$ is equal to 4.04 and the $q_{\text{critical (5), }\alpha=.01}$ is equal to 4.93.

When performing pairwise comparisons, it is customary to report the table of differences between means with an indication of their significance (e.g., one star meaning significant at the .05 level, and two stars meaning significant at the .01 level). This is shown in Table 3.

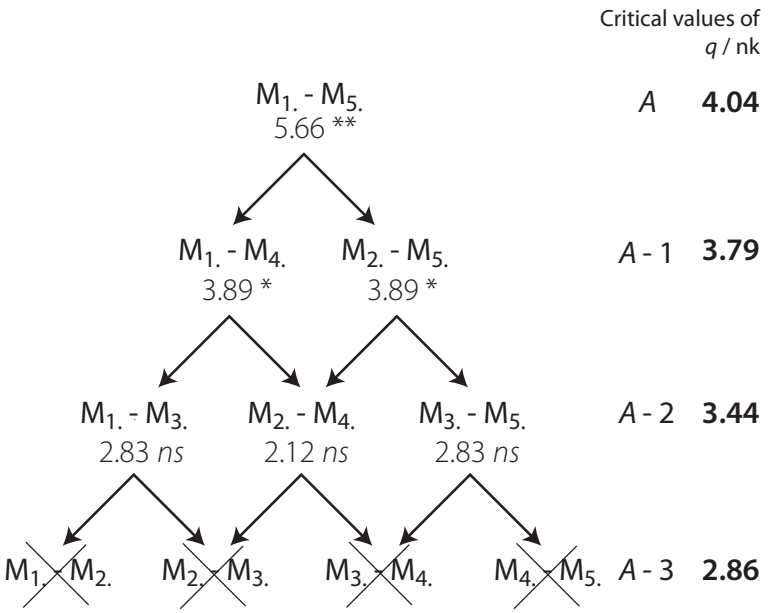

Figure 2 Newman-Keuls test for the data from a replication of Loftus & Palmer (1974). The number below each range is the q_{observed} for that range.

2.2 Newman-Keuls test

Note that for the Newman-Keuls test, the group means are ordered from the smallest to the largest. The test starts by evaluating the largest difference which corresponds to the difference between $M_{1.}$ and $M_{5.}$ (i.e., "contact" and "smash"). For $\alpha = .05$, $R = 5$ and $\nu = N - K = 45$ degrees of freedom, the critical value of q is 4.04 (using the ν value of 40 in the table). This value is denoted as $q_{\text{critical (5)}} = 4.04$. The q_{observed} is computed from Equation 1 (see also Table 2) as:

$$q_{\text{observed}} = \frac{M_{5.} - M_{1.}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} = 5.66 \tag{3}$$

The q_{observed} is greater than $q_{\text{critical (5)}}$ and H_0 is rejected for the largest pair.

Now we proceed to test the means with a range of 4, namely the differences $(M_{4.} - M_{1.})$ and $(M_{5.} - M_{2.})$. With $\alpha = .05$, $R = 4$ and 45 degrees of freedom, $q_{\text{critical (4)}} = 3.79$. Both differences are declared

Table 4 Presentation of the results of the Newman-Keuls test for the data from Table 1.

	Experimental Group				
	$M_{1.}$ Contact 30	$M_{2.}$ Hit 1 35	$M_{3.}$ Bump 38	$M_{4.}$ Collide 41	$M_{5.}$ Smash 46
$M_{1.} = 30$ Contact	0	5.00 *ns*	8.00 *ns*	11.00*	16.00**
$M_{2.} = 35$ Hit		0	3.00 *ns*	6.00 *ns*	11.00*
$M_{3.} = 38$ Bump			0	3.00 *ns*	8.00 *ns*
$M_{4.} = 41$ Collide				0	5.00 *ns*
$M_{5.} = 46$ Smash					0

$*p < .05, \quad ** \; p < .01$

significant at the .05 level ($q_{\text{observed (4)}} = 3.89$ in both cases). We then proceed to test the comparisons with a range of 3. The value of q_{critical} is now 3.44. The differences $(M_{3.} - M_{1.})$ and $(M_{5.} - M_{3.})$, both with a q_{observed} of 2.83 are declared non-significant. Further, the difference $(M_{4.} - M_{2.})$, with a q_{observed} of 2.12, is also declared non-significant. Hence, the null hypothesis for these differences cannot be rejected and all comparisons implied by these differences should be crossed out. That is, we do not test any difference with a range of $A - 3$ $[(M_{2.} - M_{1.})$, $(M_{3.} - M_{2.})$, $(M_{4.} - M_{3.})$, and $(M_{5.} - M_{4.})]$. Because the comparisons with a range of 3 have already been tested and found to be non-significant, any comparisons with a range of 2 will consequently be declared non-significant as they are implied or included in the range of 3 (i.e., the test has been performed implicitly).

As for the Tukey test, the results of the Newman-Keuls tests are often presented with the values of the pairwise differences between the means and with stars indicating the significance level (see Table 4). The comparison of Table 4 and Table 3 confirms that the Newman-Keuls test is more powerful than the Tukey test.

Related entries

Analysis of variance,Bonferroni procedure, Holms' sequential Bonferroni procedure, Honestly significant difference (HSD) test, multiple comparison test, Pairwise comparisons, Post-hoc comparisons, Scheffé's test.

Further readings

1. Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental Design and Analysis for Psychology.* Oxford: Oxford University Press.
2. Dudoit S., van der Laan, M. (2008). *Multiple Testing Procedures with Applicatoins to Genomics*. New York: Springer.
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.

Appendix: Table of Critical Values of the Studentized Range q

Studentized Range q Distribution. Table of Critical Values for $\alpha = .05$ $\boldsymbol{\alpha = .01}$

	R = Range (Number of Groups)													
ν_2	2	3	4	5	6	7	8	9	10	12	14	16	18	20
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.79	7.03	7.24	7.43	7.59
	5.24	**6.33**	**7.03**	**7.56**	**7.97**	**8.32**	**8.61**	**8.87**	**9.10**	**9.10**	**9.48**	**10.08**	**10.32**	**10.54**
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.43	6.66	6.85	7.02	7.17
	4.95	**5.92**	**6.54**	**7.01**	**7.37**	**7.68**	**7.94**	**8.10**	**8.37**	**8.71**	**9.00**	**9.24**	**9.46**	**9.65**
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.18	6.39	6.57	6.73	6.87
	4.75	**5.64**	**6.20**	**6.62**	**6.96**	**7.24**	**7.47**	**7.68**	**7.86**	**8.18**	**8.44**	**8.66**	**8.85**	**9.03**
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.98	6.19	6.36	6.51	6.64
	4.60	**5.43**	**5.96**	**6.35**	**6.66**	**6.91**	**7.13**	**7.33**	**7.49**	**7.78**	**8.03**	**8.23**	**8.41**	**8.57**
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.83	6.03	6.19	6.34	6.47
	4.48	**5.27**	**5.77**	**6.14**	**6.43**	**6.67**	**6.87**	**7.05**	**7.21**	**7.49**	**7.71**	**7.91**	**8.08**	**8.23**
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.71	5.90	6.06	6.20	6.33
	4.39	**5.15**	**5.62**	**5.97**	**6.25**	**6.48**	**6.67**	**6.84**	**6.99**	**7.25**	**7.47**	**7.65**	**7.81**	**7.95**
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.62	5.80	5.95	6.09	6.21
	4.32	**5.05**	**5.50**	**5.84**	**6.10**	**6.32**	**6.51**	**6.67**	**6.81**	**7.06**	**7.27**	**7.44**	**7.59**	**7.73**
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.53	5.71	5.86	6.00	6.11
	4.26	**4.96**	**5.40**	**5.73**	**5.98**	**6.19**	**6.37**	**6.53**	**6.67**	**6.90**	**7.10**	**7.27**	**7.42**	**7.55**
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.46	5.64	5.79	5.92	6.03
	4.21	**4.89**	**5.32**	**5.63**	**5.88**	**6.08**	**6.26**	**6.41**	**6.54**	**6.77**	**6.96**	**7.13**	**7.27**	**7.40**
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.40	5.57	5.72	5.85	5.96
	4.17	**4.84**	**5.25**	**5.56**	**5.80**	**5.99**	**6.16**	**6.31**	**6.44**	**6.66**	**6.85**	**7.00**	**7.14**	**7.26**
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.35	5.52	5.66	5.79	5.90
	4.13	**4.79**	**5.19**	**5.49**	**5.72**	**5.92**	**6.08**	**6.22**	**6.35**	**6.56**	**6.74**	**6.90**	**7.03**	**7.15**
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.31	5.47	5.61	5.73	5.84
	4.10	**4.74**	**5.14**	**5.43**	**5.66**	**5.85**	**6.01**	**6.15**	**6.27**	**6.48**	**6.66**	**6.81**	**6.94**	**7.05**
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.27	5.43	5.57	5.69	5.79
	4.07	**4.70**	**5.09**	**5.38**	**5.60**	**5.79**	**5.94**	**6.08**	**6.20**	**6.41**	**6.58**	**6.73**	**6.85**	**6.97**
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.23	5.39	5.53	5.65	5.75
	4.05	**4.67**	**5.05**	**5.33**	**5.55**	**5.73**	**5.89**	**6.02**	**6.14**	**6.34**	**6.51**	**6.65**	**6.78**	**6.89**
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.20	5.36	5.49	5.61	5.71
	4.02	**4.64**	**5.02**	**5.29**	**5.51**	**5.69**	**5.84**	**5.97**	**6.09**	**6.29**	**6.45**	**6.59**	**6.71**	**6.82**
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.10	5.25	5.38	5.44	5.59
	3.96	**4.55**	**4.91**	**5.17**	**5.37**	**5.54**	**5.69**	**5.81**	**5.92**	**6.11**	**6.26**	**6.39**	**6.51**	**6.61**
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	5.00	5.15	5.27	5.38	5.48
	3.89	**4.45**	**4.80**	**5.05**	**5.24**	**5.40**	**5.54**	**5.65**	**5.76**	**5.93**	**6.08**	**6.20**	**6.31**	**6.41**
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.90	5.04	5.16	5.27	5.36
	3.82	**4.37**	**4.70**	**4.93**	**5.11**	**5.26**	**5.39**	**5.50**	**5.60**	**5.76**	**5.90**	**6.02**	**6.12**	**6.21**
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.81	4.94	5.06	5.15	5.24
	3.76	**4.28**	**4.59**	**4.82**	**4.99**	**5.13**	**5.25**	**5.36**	**5.45**	**5.60**	**5.73**	**5.84**	**5.93**	**6.02**
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.71	4.84	4.95	5.04	5.13
	3.70	**4.20**	**4.50**	**4.71**	**4.87**	**5.01**	**5.12**	**5.21**	**5.30**	**5.44**	**5.56**	**5.66**	**5.75**	**5.83**
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.62	4.74	4.85	4.93	5.01
	3.64	**4.12**	**4.40**	**4.60**	**4.76**	**4.88**	**4.99**	**5.08**	**5.16**	**5.29**	**5.40**	**5.49**	**5.57**	**5.65**