R&D REPORT
NO. 118

Proficiency testing for sensory ranking tests: statistical guidelines - part 1

2000

Campden BRI

Campden BRI

Chipping Campden
Gloucestershire
GL55 6LD, UK

Tel: +44 (0)1386 842000
Fax: +44 (0)1386 842100
www.campden.co.uk

R&D Report No. 118

Proficiency testing for sensory ranking tests: statistical guidelines - part 1

JA McEwan

2000

EXECUTIVE SUMMARY

Proficiency testing in sensory analysis is an important step which aims to demonstrate that data obtained from trained sensory assessors are as reliable as one would expect from any other objective measurement tool. Sensory analysis is unique in that it uses human assessors to measure the perception of a wide range of stimuli, as detected by the senses of sight, sound, smell, taste and touch. Sensory measurements are perceptual translations of physical/chemical stimuli, and as such differ from other directly physical or chemical measurements.

The uniqueness of sensory analysis poses some specific problems for measuring the proficiency of the sensory panel. Cultural and psychological/physiological differences may give rise to different thresholds of perception, and the panel's product experiences may lead to differences in the ability to discriminate between samples. Such factors make the job of the statistician more demanding; defining the expected level of performance in terms of sample discrimination, for example, becomes difficult. Another issue is the definition of a 'true' value or expected result, which is not so clearly defined for sensory analysis.

There are a number of methods in the literature that could be used to evaluate the performance of sensory panels for ranking tests. These include correlation coefficients, Friedman test, the coefficient of concordance, the 'egg shell' procedure and the rank interaction test. These methods are investigated for their potential use in proficiency testing, and selected methods are explored further using data collected as part of two ring trials.

This first stage of research proposes a procedure for the establishment of performance criteria for future ring trials, and how panels can be assessed according to these measures. Moreover, the important issue of 'true value' in proficiency testing seems to have been resolved through the calculation of an 'expected result'.

ACKNOWLEDGEMENTS

The work reported is part of an EU funded project called ProfiSens (SMT-4-CL98-2227), which is running from September 1998 to August 2001. This project involves 17 partners, representing ten EU and one non-EU country. The participants are:

1 CCFRA, UK
2 VTT Biotechnology, Finland
3 Swedish Meat Research Institute, Sweden
4 Matforsk – Norwegian Food Research Institute, Norway
5 Polish Academy of Sciences, Poland
6 BioSS, UK
7 University College Cork, Ireland
8 TNO Nutrition and Food Research Institute, Netherlands
9 Unilever Research Colworth Laboratory, UK
10 Biotechnological Institute, Denmark
11 AINIA – Instituto Tecnologico Agroalimentario, Spain
12 Adriant, France
13 SIK – Swedish Institute for Food and Biotechnology, Sweden
14 Nestle R&D Centre Bjuv, Sweden
15 VALIO, Finland
16 INRAN - Instituto Nazionale di Ricerca per gli Alimenti e la Nutrizione, Italy
17 V&S VinSprit – Swedish Wine and Spirits Corporation, Sweden

This report is based on work undertaken by TG2 on Statistical Guidelines for Proficiency Testing. This group included CCFRA (Jean A. McEwan), BioSS (Tony Hunter), Matforsk (Per Lea) and TNO (Leo van Gemert). Particular thanks are given to the contribution of Jean McEwan and Tony Hunter, who undertook the bulk of the data analysis and report writing.

Thanks are also given to the other participants, particularly to those in TG3 undertaking the organisation and sensory evaluation with respect to the ranking tests. These data play an important role is developing the statistical guidelines.

CONTENTS

1. INTRODUCTION

1.1 Background to Proficiency Testing

Proficiency testing in sensory analysis is an important step to demonstrate that data obtained from human instruments are as reliable as one would expect from any objective measurement tool. Sensory analysis is unique in that it uses human assessors to measure the perception of a wide range of stimuli, as detected through the senses of sight, sound, smell, taste and touch. Sensory measurements are perceptual translations of physical/chemical stimuli, and as such differ from other direct physical or chemical measurements.

The uniqueness of sensory analysis poses some specific problems for measuring the proficiency of the instrument (panel) providing the data. Cultural and individual differences may give rise to different thresholds of perception, and product experience of the panel may lead to differences in the ability to discriminate between samples. Such factors make the job of the statistician more difficult, as defining the expected level of performance in terms of which samples are discriminated, for example, becomes difficult.

Another issue for the statistical evaluation of the data is the definition of a 'true' value, which is not so clearly defined for sensory analysis. In the case of ranking, the most logical definition is the rank order of the samples according to the way in which they were spiked. However, care must be taken to ensure that the spiking agent does not give rise to an unexpected interaction, thus rendering the supposed 'true' rank order incorrect. This issue is even more problematic for descriptive profile data (McEwan, 2000)

This document outlines approaches to the analysis of sensory ranking data, with the specific objective of monitoring the performance of the panel as part of a sensory proficiency testing scheme.

1.2 Methods Covered

There are a number of methods in the literature that could be used to evaluate the performance of sensory panels for ranking tests. The 'egg shell' plot is a graphical method that allows problem assessors to be easily identified (Appendix 6). While the 'egg shell' plot does not directly address the objective of measuring panel performance, it should not be forgotten as an aid to communication and as a diagnostic tool.

Formal methods include correlation coefficients, the coefficient of concordance, and the rank interaction test. The rank interaction test, while having an initial appeal was rejected as being rather complex, and questions were raised regarding the statistical distributions (Appendix 7).

The correlation coefficient is attractive in that it is simple to calculate, whilst the coefficient of concordance offers a tool to measure the agreement between assessors in a panel, as well as the concordance between the panel ranking and the expected ('true') rank. Appendix 5 explores the statistical properties of the coefficient of concordance.

1.3 Panel Performance or Assessor Performance

One important aspect to clarify at the outset, is the purpose of proficiency testing with respect to performance of panels or performance of assessors.

It is very clear, that whether in research or commercial projects, it is the panel result that is used to make decisions about the samples being evaluated. Therefore, proficiency testing is about measuring the performance of a panel, not individuals in the panel. If individual assessors perform poorly, then their data will bring down the overall performance of the panel, and consequently the panel will not have performed well.

However, concordance between members of the panel is of interest, as one measure of a panel's performance. It is measured by determining if each member of the panel provides the same information.

This document is, therefore, concerned mainly with the performance of panels.

1.4 Document Format

Chapter 2 addresses the important issue of experimental design, whilst Chapter 3 reports on the standard procedure for analysing rank data to determine if the samples are significantly different (Friedman rank test), and to determine which samples are significantly different (Studentised Range multiple comparison test).

Chapter 4 treats the important subject of expected ('true') result. Clearly, where samples are not spiked in a uni-dimensional way, calculation of the true or expected rank is fundamental to later assessment of panel performance. This chapter outlines the stages in establishing panel performance.

Chapters 5 and 6 provide an explanation of the correlation method and coefficient of concordance for measuring panel performance. Chapter 7 reports the results of the Friedman test to establish significant differences between samples and also looks at the calculation of the expected ranking. Chapter 8 works through some data for measuring panel performance. Chapters 9 summarise the results of two ring trials conducted in 1999. Finally Chapter 10 considers how to set the performance criteria for future trials.

2. EXPERIMENTAL DESIGN

2.1 Introduction

Statistically designed and analysed comparative experiments are of greatest value in those experimental circumstances where treatment (or Sample) effects are likely to be small compared to the underlying variation. Statistical analysis is carried out to estimate the effects of treatment and to assign well-founded estimates of variation to treatment effects. This in turn leads to either tests of significance or to confidence intervals.

In order to satisfy the assumptions of 'statistical' analysis there should be elements of randomisation in the design. To increase the precision of the experiment in estimating differences between treatments, it is usual to identify known sources of extraneous variation and to seek to 'block' by these factors i.e. incorporate them into the design of the experiment. For sensory profiling experiments these are Assessor, Order of Sample Presentation and the Effects of Previous Sample. The interpretation of treatment effects is greatly facilitated by the adoption of a factorial treatment structure for the Samples. The precision of an experiment can be improved by increasing the replication.

For sensory science, in general, there has been a lack of awareness of the advantages of carefully designing sensory experiments (Hunter, 1996; MacFie *et al.*, 1989) and of developments in the analysis of such data (Jones and Wang, 2000). Sensory profiling experiments, in particular, are very similar statistically to the 'cross over' designs used in pre-clinical and clinical medicine (Jones and Kenward, 1989). These designs were originally used for *in vivo* animal studies in the biological sciences.

Below the logic of proposed designs for sensory profiling is developed. It is recommended that such designs are also used in Ranking experiments. This is divided into three parts, first the Assessor design, second the design of the Samples and thirdly Replication. An alternative fuller account is given by Hunter (1996).

2.2 The Assessor Design

It is common to find that Assessors are the largest effect in the Analysis of Variance of the data for each attribute. This arises because Assessors use different parts of the scale. Nevertheless, Assessors provide useful information on the differences between Samples. Provided that each Assessor tests each Sample the same number of times, it is possible to estimate treatment effects entirely within Assessors. Assessors are thus a block factor.

For data from Sensory Profiling experiments, 'Order' effects are also known to be important (Muir and Hunter, 1991/2). A Sample tested first in a Session is usually rated differently from the same Sample tested later in a Session. Such is the magnitude of this effect that it is important to either randomise Order of presentation within a Session or alternatively design it into the experiment as a block effect in addition to Assessors.

Experience indicates that Assessor and Order effects can be effectively designed to be block effects using designs based on Latin Squares. Furthermore, if the cyclic Latin Squares due to Williams (1949) are used as a base for the design then protection is provided against interference effects from the Previous Sample (MacFie *et al.,* 1989 and Hunter, 1996). Although there is very little evidence to show that these interference effects are important (Muir and Hunter, 1991/2), it is prudent to design the experiment so that treatment estimates are protected. Hunter (1996) shows how Williams Latin Squares can be used to generate statistically efficient yet practical designs for nearly all profiling experiments.

2.3 Sample Design

For many experiments it is not possible to impose a factorial treatment structure on the Samples. However, when the Samples are from the laboratory it is usually possible to structure them in a factorial manner. Such structuring improves the ability to interpret the results of the experiment. A complete set of all factorial combinations is often used and can be very helpful in identifying 'active' factors. In those circumstances where the number of

factorial combinations is too numerous for one sensory experiment then fractional designs should be considered. It is possible to find designs from the literature which allow the main effects of three factors to be investigated with four Samples, seven factors with eight Samples and eleven factors with twelve Samples.

2.4 Replication

Replication is primarily used to increase the precision of the Sample estimates although it should not be used as a substitute for an undersized panel of Assessors. A secondary advantage is that Replication allows each Assessor's reliability (i.e. the agreement between different ratings of the same Sample by the same Assessor) to be determined.

The word Replication is used in a number of different circumstances in Sensory Science and does not appear to have the precise usage that is seen in the Biological Sciences. Below we explain our understanding of Replication and following it we explore an alternative idea about Replication.

First Scenario

In the first Replicate, the Assessors rate each Sample, if necessary spreading the assessments over a number of Sessions. In the second and subsequent Replicates, the Assessors rate the Samples again using a new randomisation that preserves the "blindness" of the trial. If Assessors are given Samples in the same order in each Replicate then they will eventually become aware of this fact and will anticipate the results thus nullifying the concept of independent ratings. Also, different randomisations for each Replicate allows each Assessor's data to be independently tested for Order and Session effects.

Where the number of Samples require more than one Session per Replicate for assessment, then it is desirable, on statistical grounds, that in each Replicate the Samples are divided into

the Sessions differently for each Assessor. If this is not possible (i.e. when hot Samples are being tested) then the Samples should be divided into Sessions differently in each Replicate.

Second Scenario

When more than one Session per Replicate is required the alternative to the above is to randomly allocate each Replicate of each Sample to a session with the restriction that Replicates of the same Sample cannot occur in the same Session. An incomplete block design (Fully Balanced or Partially Balanced Incomplete Block Design) might be used for this purpose. If at all possible, different randomisations of the design should be used for each Assessor.

Overall, the *"Second Scenario"* offers no advantage and suffers from the disadvantage of requiring the design to be completed in order to yield data that can be easily analysed. The *"First Strategy"* can be recommended as it allows the experiment to be completed one Replicate at a time and because it allows learning effects and/or changes in the Samples to be easily monitored.

Finally, it is desirable (although infrequently realised) that for Replicate assessments new Samples are drawn for each Replicate. This then allows variability between Samples of the same product or formulation to be incorporated into the experiment. Otherwise specific Samples are being compared without allowing for sampling or manufacturing variation.

2.5 Conclusions

Careful design of sensory experiments, using well established techniques freely available in the literature, allows the maximum amount of information to be derived from the work of the Sensory Laboratory.

3. ANALYSIS OF RANKING DATA

3.1 Introduction

This chapter concentrates on the analysis of ranking data for a designated attribute as normally undertaken by the sensory analyst to establish if there are differences between samples. The method chosen is the Friedman rank test to establish if there are differences between the samples. To identify if two samples are different, two methods were investigated: a Studentised Range test (Hochberg and Tamhane, 1987) and a method proposed by Conover (1999). The latter method was finally chosen as the multiple comparison value calculation is based on an analysis of variance of the rank table.

Most statistical packages will allow the Friedman rank test to be undertaken. However, it is less common to find a package that offers multiple comparison tests on non-parametric data. For this reason the calculations are presented to allow the user to undertake the multiple comparison test by hand (Section 3.4).

3.2 Tabulating the Data

A useful starting point for rank data is to produce a table of results (see Table 3.1) for one replicate assessment to rank five samples on sweetness. A table is produced for each replicate, and each replicate should be analysed separately. Table 3.1 shows the ranking provided by each assessor, and from these data the Rank Sum of each sample can be calculated. This gives a first impression of whether there are likely to be differences between the samples. If the Rank Sums are similar, then this would indicate that the panel would not have be able to discriminate between the samples. Conversely, the greater the difference between the Rank Sums, the more likely that differences will be detected between samples.

Table 3.1: Example of a set of rank data where 5 apple juice samples were ranked according to sweetness intensity.

	Samples				
Assessor	**1**	**2**	**3**	**4**	**5**
1	5	2	1	3	4
2	5	1	2	3	4
3	5	2	3	1	4
4	5	1	2	3	4
5	5	1	4	3	2
6	5	1	2	3	4
7	5	2	3	1	4
8	5	1	3	2	4
Rank Sum (ΣRi)	40	11	20	19	30
ΣRi^2	1600	121	400	361	900
Panel Rank	5	1	3	2	4
Panel Mean Rank	5.0	1.4	2.5	2.4	3.8
Expected Rank	5	1	2	3	4
Expected Rank Sum	40	8	16	24	32

The final two rows of Table 3.1 show the 'expected' rank order and 'expected' rank sum. Using this information provides some initial indication as to whether the panel provided a 'correct' answer.

3.3 The Friedman Rank Test

Undertaking a Friedman rank test on Minitab (Version 12), a Friedman statistic (S) of 25.1 was obtained. Minitab also provides the p-value (significance level), which was p = 0.000. This result implied that there was one or more significant differences between the samples at more that the 0.1% level of significance.

3.4 Identifying Sample Differences

The calculations for the Conover method are illustrated below. The first step is to work out the value of A, using the following formula.

$$A = \frac{JI(I+1)(2I+1)}{6}$$

$$= \frac{(8x5)(5+1)(10+1)}{6}$$

$$= 440$$

I = number of samples
J = number of assessors

Next the actual multiple comparison value is worked out.

$$|R_i - R_j| \geq t_{1-\alpha/2} \left[\frac{2(JA - \Sigma R_j^2)}{(J-1)(I-1)} \right]^{1/2}$$

$$\geq 2.048 \left[\frac{2(8x440 - 3382)}{7x4} \right]^{1/2}$$

$$\geq 2.048 \left[\frac{2(3520 - 3382)}{28} \right]^{1/2}$$

$$\geq 2.048 \quad \text{Sqrt}(9.857)$$

$$\geq 6.4$$

I = number of samples
J = number of assessors
R_i = rank sum of Sample i
R_j = rank sum of Sample j
R_j^2 = rank sum squared of Sample j

$t_{1-\alpha/2}$ is obtained from the Student Tables at the desired level of Significance (α = 5% here)
Degrees of freedom are (J-1)(I-1)

The comparison value of 6.4 is used to compare sample rank sums. In order to compare panel mean ranks, this figure is divided by the number of assessors in the panel, in this case 8, to give a value of 0.8.

The use of mean ranks is justified, as the means help establish later a more meaningful correlation between the 'expected rank' and the panel ranking results.

Table 3.2 highlights which pairs of samples were different. If a sample shares the same letter (3rd column), then they are not significantly different, at the 5% level of significance.

Table 3.2: Representation of sample differences, where samples with different letters are significantly difference (5% significance).

		Difference	
Sample	**Mean Rank**	**5%**	**1%**
2	1.4	a	a
4	2.4	b	ab
3	2.5	b	ab
5	3.8	c	bc
1	5.0	d	c
	Conover -MC	0.8	1.4

3.5 Consistent Results

While the above tests demonstrate whether the panel perceived differences between the samples, and where the differences were, this does not necessarily guarantee results consistent with sample 'spiking' or with the results of other panels.

As part of a proficiency testing procedure, two ranking evaluations are undertaken. A consistent panel should produce the same result on both occasions. However, the level of consistency needs to be defined (see Chapters 5 and 6).

4. THE EXPECTED RESULT

4.1 What is the True Value

The true value is a term used in other proficiency schemes, and is defined to be the result that is expected based on prior knowledge of the samples. This is a relatively simple concept in analytical tests where the result is directly related to how the samples have been modified.

In the case of ranking, the true (or expected) ranking may be defined logically when samples have been spiked in a uni-dimensional way, for example, increase in sucrose leads to an increase in perceived sweetness. However, the spiking ingredient could interact with the other ingredients in the product, thus distorting this ordering.

In addition, ranking is undertaken on more complex foods, where a number of ingredients may have been altered, as in the case of a mixture experiment. It is then very difficult to establish the true ranking in advance of the experiment. A good set of sensory data is required to make an accurate estimate of the true ranking. Without this, panel performance could not be measured.

As well as defining the expected ranking, it is also important for sensory analysis to state the level of discrimination expected both in terms of overall significance and the specific samples that are expected to be different.

4.2 Establishing the Expected Rank Order

The expected rank order for samples may be defined as the logical perception order according to how samples were spiked. However, in the case of uni-dimensional spiking and in other more complex situations, the expected ranking should be calculated or confirmed by examining the data from all panels in a ring trial, if this has not been previously undertaken.

Calculating the expected ranking from, say, 10 panels is simply achieved by working out the overall mean rank for each sample, based on the data from all assessors in the 10 sensory panels. It is desirable that the expected rank is based on data from sensory panels known to have good ability with both the method and product category.

However, it is possible for a panel to achieve the expected rank order, yet only find one sample different from one other sample. Such a panel would not perform well in its ability to discriminate between samples that are known to be different. Further, a panel order may switch the ranking of 2 samples simply because there is no perceptible difference between them. In this case, deviance from the expected rank order does not necessarily indicate poor performance. This problem can be minimised by working on mean panel ranks, rather than ranking the panel rank sums for each sample.

Thus, it is important to evaluate a panel's performance, not only on its ability to produce the expected ranking, but also on its ability to discriminate between the samples in terms of the appropriate statistical test showing a specified level of significance. Moreover, the panel would be expected to find specified pairs of samples different, at a given level of statistical significance.

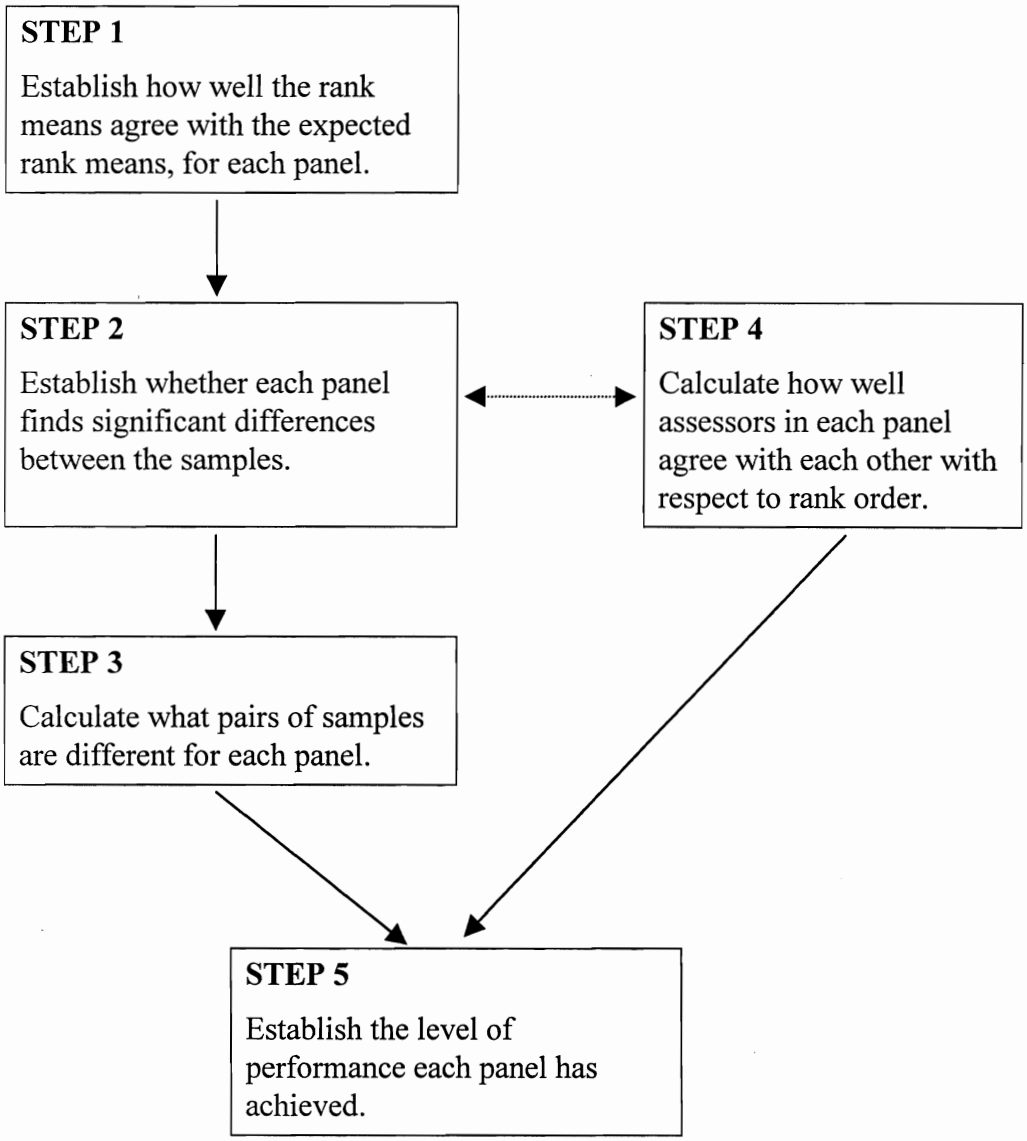
Finally, a good panel is one whose assessors agree well on the expected rank order, and therefore some measure of panel agreement may enhance overall evaluation of panel performance.

4.3 Establishing the Expected Discrimination Between Samples

In order to determine the expected level of discrimination between samples, some advanced data are required to enable expected differences to be defined. This can be achieved in two ways. Firstly, the panel ranks from all panels in a ring trial can be used. These panels ranks would be submitted to the Friedman and Conover tests, and through examination of the data, expected levels of performance set. However, this method suffers from being data dependent.

Ideally, the expected results should be set in advance of the ring trial. This approach was adopted later in the project where a number of pre-tests were undertaken using trained and untrained assessors (McEwan, 2001). The trained assessors of known ability are used to establish what a 'good' panel could achieve, whilst the untrained assessors reflect a 'poorer' performance.

4.4 Stages in Establishing Panel Performance



5. CORRELATION METHODS FOR PERFORMANCE

5.1 Principle

One of the simplest ways to measure whether the ranking provided by an assessor or a panel corresponds to the expected result is to calculate the rank correlation. The correlation measures the strength of the relationship between the observed ranking and the expected ('true') result. If the correlation is one, then the assessor or panel has performed perfectly. A correlation of zero would imply no relationship, whilst a negative correlation would suggest either the ranking was done in reverse (the panel totally misunderstood the test instructions), the panel did not perform well, or that there were no perceivable differences between the samples. The latter should not be the case for proficiency testing scheme samples.

5.2 Procedure

The first step is to produce a table of the rank data, together with the expected or true ranking.

Table 5.1: The ranking results provided as a panel result, together with the 'expected' ranking and deviation (d) from this.

Sample	Expected Rank	Panel Rank	Difference d	d^2
1	5	5.0	0	0
2	1	1.4	0.4	0.16
3	3	2.5	0.5	0.25
4	2	2.4	1.4	1.96
5	4	3.8	0.2	0.04
				Sum = 2.41

The Spearman's rank correlation (ρ) can then calculated between the true rank order and the panel ranking using the information in Table 4.1 and the formula below.

$$\rho = 1 - \frac{6\Sigma(d^2)}{n^3 - n}$$

d = difference between true rank and panel rank

n = number of samples ranked

$$= 1 - \frac{6*2.41}{216 - 6}$$

$$0.931$$

However, as the panel mean ranks were used, it is more useful to use the Pearson's correlation, particularly as the 'expected rank' in the ring trials may not be whole numbers, as it will be based on the results of several panels. The Pearson correlation (r) is worked out by most statistical software, and so the formula is not reported here (see O'Mahoney, 1986).

For the data in Table 5.1, a correlation coefficient of 0.973 was calculated. Statistical tables will reveal that a correlation of 0.685 is required to achieve a 10% level of significance, 0.803 is required to achieve a 5% level of significance, and a correlation of 0.933 for a 1% significance. This test is one-sided, as the interest is in securing a correlation as high as possible.

6. COEFFICIENT OF CONCORDANCE

6.1 Principle

The procedure reported in Chapter 5 essentially dealt with the case of two rankings, and the strength of the relationship between them. However, the problem of a sensory panel is that there are several assessors each providing a ranking, and that the reliability of the panel rank is, in part, dependent on the performance of the panel. Therefore, the concordance between the ranks provided by each assessor in the panel is of interest. However, a panel can show a high concordance, yet not perform well in relation to the 'expected' ranking.

It should be noted that the coefficient of concordance (W) is essentially the same test as the Friedman test, but looks at the data from a different angle. In other words, Steps 2 and 4 of Section 4.4, take a different perspective of the data, but are based on the same underlying statistical test.

6.2 Procedure

Kendall and Gibbons (1990) proposed a solution to this problem of calculating the agreement between more than two rankings, by proposing the Coefficient of Concordance (W). W is based partly on the deviations between the rankings as shown in Table 5.1. However, in this case it is necessary to calculate the deviations from all assessors. The coefficient of concordance (W) is then written as follows:

$$W = \frac{12\Sigma S}{m^2(n^3 - n)}$$

S = sum of the squared deviations between sample rank totals around their mean
n = number of samples ranked
m = number of assessors

The value W ranges from 0 to 1, where 0 would imply no agreement among the rankings (by assessors) and 1 would imply perfect agreement (concordance). In other words, as W increases from 0 to 1 the deviations become larger, and by implication a greater agreement between the rankings.

Appendix 4 provides some critical values of W, which are based on similar calculations to the Friedman test.

7. FRIEDMAN RESULTS AND TRUE RANK CALCULATIONS

7.1 Apple Juice

Sample Information

Five samples of apple juice were made using a blend of glucose and fructose, where each mixture comprised 50 ml of apple juice, 50 ml of water and 6.5 g of the sugar blend. The sugar blends are shown in Table 7.1. In this case, the 'logical' rank was not known in advance.

Table 7.1: Sugar blends and 3-digit numbers used to code the products for 2 replicate assessments.

	Sugar Blend		Code	Code
Sample	**Glucose**	**Fructose**	**Rep 1**	**Rep 2**
1	100%	-	966	166
2	-	100%	551	352
3	25%	75%	962	173
4	50%	50%	439	826
5	75%	25%	985	387

Calculation of Expected Rank and Sample Discrimination

For evaluation, assessors in each sensory panel were asked to rank the samples from least to most sweet, or vice-versa. However, all laboratories were asked to ensure that the data sent to the data co-ordinator were coded from '5' for 'least sweet' to '1' for 'most sweet'.

Table 7.2 shows the mean rank across all panels and assessors in the apple juice evaluation. There is a clear expected rank order as both replicates provide the same result.

Table 7.2: Average rank for the apple juice samples, calculated to one decimal place.

Rep 1	Mean 1	Rep 2	Mean 2	Expected	
966	4.9	166	4.9	4.9 (5)	Least sweet
985	3.8	387	3.9	3.8 (4)	
439	2.9	826	2.9	2.9 (3)	
962	2.0	173	2.0	2.0 (2)	
551	1.3	352	1.3	1.3 (1)	Most sweet

Table 7.3 shows the average rank for each panel. In order to determine which samples are significantly different a Friedman test was undertaken.

Table 7.3: Average rank for the apple juice samples, calculated to one decimal place, over both replicates for each panel.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
966/166	5	4.9	4.8	5.0	4.9	4.9	5.0	5.0	4.9	5.0	4.7	4.8
985/387	4	3.9	3.8	3.6	4.0	4.0	4.1	3.8	3.8	3.9	4.0	3.7
439/826	3	2.9	3.0	3.1	2.7	3.1	2.8	2.8	2.9	2.8	2.8	3.0
962/173	2	1.9	1.9	1.9	2.1	2.0	2.2	2.3	2.1	1.9	2.0	2.0
551/352	1	1.5	1.5	1.3	1.3	1.2	1.0	1.1	1.2	1.4	1.4	1.5

The Friedman statistics was 44, with $p < 0.00001$. Table 7.4 shows the rank sums from the Friedman analysis, and the differences identified between samples based on the Conover multiple comparison test.

These results suggest that panels should be able to at least discriminate between all samples. Thus, this ranking was very easy! In addition, the coefficient of concordance for these panel mean ranks was 0.82, suggesting good agreement between the eleven panels.

Table 7.4: Panel analysis rank means, together with the significant differences based on Conover comparison values of 0.14 and 0.24 at the 5% and 1% level.

			Differences	
Sample	**Rank Sum**	**Mean Rank**	**5%**	**1%**
966/166	53.9	4.9	a	a
985/387	42.5	3.9	b	b
439/826	31.9	2.9	c	c
962/173	22.3	2.0	d	d
551/352	14.4	1.3	e	e

Summary of Friedman Results and Multiple Comparisons

Tables 7.5 and 7.6 show the rank means, and multiple comparison value (MC at 5% significance) for each panel, for the 2 replicate assessments. If the difference between two rank means are greater than the multiple comparison value, then the samples were significantly different at the 5% level of significance.

Table 7.5: Rank means, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: apple juice, replicate 1.

	True	Panel										
Sample	**Value**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**K**	**L**	**M**
966	5	4.8	4.8	5.0	4.9	5.0	5.0	5.0	5.0	5.0	4.5	4.7
985	4	4.2	3.6	3.8	3.9	3.9	4.0	3.7	3.8	3.8	4.0	3.8
439	3	2.7	3.1	3.1	2.7	3.1	2.7	2.9	3.0	3.0	3.1	2.9
962	2	2.2	1.8	1.9	2.2	2.0	2.3	2.2	2.1	1.8	1.7	2.3
551	1	1.1	1.8	1.3	1.3	1.0	1.0	1.2	1.1	1.4	1.7	1.3
MC-5%	--	0.5	0.7	0.6	0.6	0.2	0.3	0.6	0.6	0.5	0.7	0.8
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n		10	17	8	12	10	10	11	8	12	13	12

Table 7.6: Rank means, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: apple juice, replicate 2.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
166	5	5.0	4.9	5.0	4.9	4.7	4.9	5.0	4.9	4.9	4.8	4.8
387	4	3.6	4.1	3.5	4.1	4	4.1	3.8	3.8	4.1	4.1	3.5
826	3	3.0	2.9	3.1	2.7	3	2.8	2.7	2.9	2.7	2.5	3.3
173	2	1.6	1.9	2.0	2.1	2	2.2	2.4	2.1	2.0	2.4	1.7
352	1	1.8	1.2	1.4	1.3	1.3	1	1.1	1.4	1.3	1.2	1.8
MC-5%	--	0.7	0.4	0.8	0.5	0.71	0.34	0.5	0.9	0.5	0.5	0.7
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n		10	17	8	12	10	10	11	8	12	13	12

7.2 Tomato Soup

Sample Information

Five samples of tomato soup were made using different levels of starch. Thus, the logical rank order was known in advance (Table 7.7).

Table 7.7: Soup samples and 3-digit numbers used to code the products for 2 replicate assessments.

Tomato Soup Formulation	Cornflour Supplement	Product Code Rep1	Product Code Rep 2	Logical Rank
TS1	Nil	681	254	5
TS2		753	310	4
TS3		272	679	3
TS4		449	239	2
TS5	Most	308	792	1

Calculation of Expected Rank and Sample Discrimination

For evaluation, assessors in each sensory panel were asked to rank the samples from least to most thick, or vice-versa. However, all laboratories were asked to ensure that the data sent to the data co-ordinator were coded from '5' for 'least thick' to '1' for 'most thick'.

Table 7.8 shows the mean rank across all panels and assessors in the tomato evaluation. There is a clear 'true' rank order as both replicates provide the same result. Table 7.9 shows the average rank for each panel, and in order to determine which samples are significantly different, the data were converted to ranks, and a Friedman test undertaken. The Friedman statistics was 33.7, with $p < 0.00001$. Table 7.10 shows the rank means from the Friedman analysis, and the differences identified between samples based on the Conover multiple comparison test.

Table 7.8: Average rank for the tomato soup samples, calculated to one decimal place.

Rep 1	Mean 1	Rep 2	Mean 2	Expected	
681	4.5	254	4.7	4.6 (5)	Least thick
753	4.2	310	3.9	4.1 (4)	
272	3.0	679	3.2	3.1(3)	
449	1.9	239	1.9	1.9 (2)	
308	1.5	792	1.4	1.5 (1)	Most thick

Table 7.9: Average rank for the tomato soup samples, calculated to one decimal place, over both replicates for each panel.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
681/254	5	4.1	4.6	4.9	5.0	4.6	4.4	4.5	3.7	4.9	3.6	4.7
753/310	4	4.3	4.3	3.8	3.6	4.3	4.4	4.3	2.8	4.0	3.9	4.0
272/679	3	3.5	3.0	3.1	3.3	2.9	3.3	2.6	2.4	3.1	4.0	3.1
449/239	2	2.1	2.1	2.3	1.7	2.1	1.7	1.6	2.8	1.6	2.6	2.0
308/792	1	1.1	1.0	1.0	1.4	1.2	1.3	2.0	3.4	1.4	1.2	1.3

Table 7.10: Panel analysis rank means, together with the significant differences based on Conover comparison values of 0.48 and 0.85 at the 5% and 1% level.

			Differences	
Sample	**Rank Sum**	**Mean Rank**	**5%**	**1%**
681/254	49.0	4.5	a	a
753/310	43.2	3.9	b	ab
272/679	34.3	3.2	c	b
449/239	22.6	2.1	d	c
308/792	16.3	1.5	e	c

From Table 7.10, it could be concluded that a very good panel should be able to discriminate all samples from each other. A good panel should discriminate the samples shown under 1%, but may not find differences between Samples 681/254 and 753/310 at the 5% level or differences between Samples 449/239 and 308/792.

The coefficient of concordance was 0.63, and therefore there was not such a good agreement between the panels for tomato soup.

Summary of Friedman Results and Multiple Comparisons

Tables 7.11 and 7.12 show the rank sums, and multiple comparison value (MC at 5%) for each panel, for the 2 replicate assessments. If the difference between two rank means are greater than the multiple comparison value, then the samples were significantly different at the 5% level of significance.

Table 7.11: Rank means, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: tomato soup, replicate 1.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
681	5	4.1	4.2	4.8	5.0	4.2	3.8	4.5	4.5	4.9	3.7	4.9
753	4	4.3	4.8	3.5	3.8	4.7	5.0	4.2	2.7	4.1	3.1	3.6
272	3	3.4	2.9	3.6	3.1	2.6	3.2	2.2	2.0	3.0	2.9	3.4
449	2	2.1	2.1	2.1	2.1	2.5	1.3	1.4	3.2	1.3	2.9	1.8
308	1	1.1	1.0	1.0	1.0	1.0	1.7	2.8	2.7	1.7	2.4	1.3
MC-5%	--	0.8	0.3	0.7	0.3	0.5	0.4	0.8	1.7	0.3	0.8	0.5
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.078	0.000	0.000	0.000
N		10	17	8	12	10	10	11	6	11	12	11

Table 7.12: Rank means, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: tomato soup, replicate 2.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
254	5	4.1	4.9	5.0	5.0	4.9	4.5	4.5	2.8	4.9	2.9	4.5
310	4	4.2	3.8	4.0	3.3	3.9	4.4	4.4	2.8	3.8	3.3	4.3
679	3	3.5	3.2	2.6	3.6	3.2	3.1	3.1	2.8	3.3	3.1	2.8
239	2	2.0	2.1	2.4	1.3	1.7	1.8	1.8	2.3	2.0	3.2	2.1
792	1	1.2	1.0	1.0	1.8	1.3	1.2	1.2	4.2	1.0	2.5	1.3
MC-5%	--	0.8	0.3	0.4	0.5	0.5	0.3	0.5	1.9	0.4	0.8	0.7
p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.339	0.000	0.000	0.000
n		10	17	8	12	10	11	11	6	11	12	11

8. SELECTED WORKED EXAMPLES: PANEL PERFORMANCE

8.1 Correlation Method

Table 8.1 shows the mean rank results for Panel L, together with the expected rank mean. The correlation coefficients between the expected rank and panel ranks were 0.884 and 0.417, respectively.

Table 8.1: Overall rank for tomato soup samples, as calculated from Panel L.

Sample	Expected Rank	Panel	
		Rep 1	Rep 2
681/254	4.6	3.7	2.9
753/310	4.1	3.1	3.3
272/679	3.1	2.9	3.1
449/239	1.9	2.9	3.2
308/792	1.5	2.4	2.5

For replicate 1, the correlation of 0.884 is significant at the 1% level, and therefore the panel rank mean was in good agreement with the expected rank. However, a non-significant result was recorded for replicate 2, and so the panel did not perform well.

8.2 Coefficient of Concordance – Panel Concordance

Table 8.2 shows the raw data from Panel A, which is used to calculate the coefficient of concordance. The first step is to produce Table 8.3.

Table 8.2: The first replicate ranking results for the 10 assessors in Panel A, together with the 'expected' ranking.

Sample	True Value	Assessors									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
681/254	5	3	1	5	4	5	5	4	5	5	4
753/310	4	5	4	4	5	3	4	5	4	4	5
272/679	3	4	5	3	3	4	3	3	3	3	3
449/239	2	2	3	2	2	2	2	2	2	2	2
308/792	1	1	2	1	1	1	1	1	1	1	1

Table 8.3: Calculation of squared deviations – Panel A, Replicate 1.

Sample	Row Sum	Mean	Deviation	Squared Deviation (S)
681/254	41	30	11	121
753/310	43	30	13	169
272/679	34	30	4	16
449/239	21	30	9	81
308/792	11	30	19	361
				748

The row sum is obtained by summing the values across assessors for each sample separately.

The mean is calculated as the total of all ranks divided by the number of samples. In other words: $[(5+4+3+2+1)*ass]/5 = (15*10)/5 = 30$.

The deviation is the difference between the row sum and the mean.

The calculation of W can now take place, by substituting the appropriate values into the formula.

$$W = \frac{12\Sigma S}{m^2(n^3 - n)}$$

$$W = \frac{12 \times 748}{10^2 \times (5^3 - 5)}$$

$$W = \frac{8976}{12000}$$

$$W = 0.748$$

S = sum of the squared deviations between sample rank totals around their mean

n = number of samples ranked
m = number of assessors

As the concordance (W) is 0.748, this illustrates that the members of the panel were not in total agreement with each other. Ideally, a good panel should have a coefficient of concordance of greater than 0.8, whilst a very good panel could achieve a value of greater than 0.9. However, the levels need to be set according to the difficulty of the task (see Chapter 10), and based on statistical criteria.

8.3 Concordance of Panel with the Expected Value

It is also possible to undertake the same calculation for the concordance between the true ranking and the panel ranking. The data for Panel A are provided in Table 8.4, whilst Table 8.5 provides the values that will be substituted into the formula for W.

The value of W = 0.950 is near 1.0, and therefore it can be concluded that Panel A is in good agreement with the 'true' ranking for both assessments.

Table 8.4 Average ranking for Panel A together with the 'expected' ranking'.

Sample	Expected Rank	Replicate 1	Replicate 2
681/254	4.6	4.1	4.1
753/310	4.1	4.3	4.2
272/679	3.1	3.4	3.5
449/239	1.9	2.1	2.0
308/792	1.5	1.1	1.2

Table 8.5: Calculation of squared deviations for two replicate ranking assessments from Panel A.

		Replicate 1			Replicate 2		
Sample	Mean	Sum	Deviation	S	Sum	Deviation	S
1	6	8.7	2.7	7.29	8.7	2.7	7.29
2	6	8.4	2.4	5.76	8.3	2.3	5.29
3	6	6.5	0.5	0.25	6.6	0.6	0.36
4	6	4.0	2.0	4.00	3.9	2.1	4.41
5	6	2.6	3.4	11.56	2.7	3.3	10.89
				28.86			28.24

The mean is calculated as the total of all ranks divided by the number of samples. In other words: [(5+4+3+2+1)*2]/5 = 30/5 = 6.

The next page demonstrates the calculation of W for both replicates, resulting in values of 0.722 and 0.706, respectively. This result implied that the panel performed reasonably well, but there is room for improvement.

$$W = \frac{12\Sigma S}{m^2(n^3 - n)}$$

S = sum of the squared deviations between sample rank totals around their mean

$$W = \frac{12 \times 28.86}{2^2 \times (5^3 - 5)}$$

n = number of samples ranked
m = number of assessors

$$W = \frac{346.32}{480}$$

$$W = 0.722$$

$$W = \frac{12\Sigma S}{m^2(n^3 - n)}$$

S = sum of the squared deviations between sample rank totals around their mean

$$W = \frac{12 \times 28.24}{2^2 \times (5^3 - 5)}$$

n = number of samples ranked
m = number of assessors

$$W = \frac{338.88}{480}$$

$$W = 0.706$$

9. SUMMARY OF PERFORMANCE

9.1 Apple Juice – Correlation Method

Tables 9.1 and 9.2 show the panel rank for the 11 panels, and for both replicates. These data are correlated with the true rank to give the results shown in Table 9.3.

Table 9.1: The panel mean ranks for Replicate 1 of the apple juice ranking.

Sample	Expected Rank	Panel										
		A	B	C	D	E	F	G	H	K	L	M
966	4.9	4.8	4.8	5.0	4.9	5.0	5.0	5.0	5.0	5.0	4.5	4.7
985	3.9	4.2	3.6	3.8	3.9	3.9	4.0	3.7	3.8	3.8	4.0	3.8
439	2.9	2.7	3.1	3.1	2.7	3.1	2.7	2.9	3.0	3.0	3.1	2.9
962	2.0	2.2	1.8	1.9	2.2	2.0	2.3	2.2	2.1	1.8	1.7	2.3
551	1.3	1.1	1.8	1.3	1.3	1.0	1.0	1.2	1.1	1.4	1.7	1.3

Table 9.2: The panel mean ranks for Replicate 2 of the apple juice ranking.

Sample	Expected Rank	Panel										
		A	B	C	D	E	F	G	H	K	L	M
166	4.9	5.0	4.9	5.0	4.9	4.7	4.9	5.0	4.9	4.9	4.8	4.8
387	3.9	3.6	4.1	3.5	4.1	4	4.1	3.8	3.8	4.1	4.1	3.5
826	2.9	3.0	2.9	3.1	2.7	3	2.8	2.7	2.9	2.7	2.5	3.3
173	2.0	1.6	1.9	2.0	2.1	2	2.2	2.4	2.1	2.0	2.4	1.7
352	1.3	1.8	1.2	1.4	1.3	1.3	1	1.1	1.4	1.3	1.2	1.8

Clearly, Table 9.3 reveals that all panels have performed well, in that the panel rank corresponds to the true rank on most occasions.

Table 9.3: Correlations between each panel and the 'true' rank order for apple juice.

Panel	Replicate	Correlation	Significance
A	1	0.988	1%
	2	0.968	1%
B	1	0.979	1%
	2	0.998	1%
C	1	0.996	1%
	2	0.987	1%
D	1	0.995	1%
	2	0.995	1%
E	1	0.996	1%
	2	0.997	1%
F	1	0.989	1%
	2	0.992	1%
G	1	0.994	1%
	2	0.985	1%
H	1	0.996	1%
	2	0.999	1%
K	1	0.995	1%
	2	0.996	1%
L	1	0.975	1%
	2	0.977	1%
M	1	0.995	1%
	2	0.961	1%

9.2 Tomato Soup – Correlation Method

Tables 9.4 and 9.5 show the panel rank means for the 11 panels, and for both replicates. These data are correlated with the true rank to give the results shown in Table 9.6.

Table 9.4: The panel rank means for Replicate 1 of the tomato soup ranking.

Sample	Expected Rank	Panel										
		A	B	C	D	E	F	G	H	K	L	M
681	4.5	4.1	4.2	4.8	5.0	4.2	3.8	4.5	4.5	4.9	3.7	4.9
753	3.9	4.3	4.8	3.5	3.8	4.7	5.0	4.2	2.7	4.1	3.1	3.6
272	3.2	3.4	2.9	3.6	3.1	2.6	3.2	2.2	2.0	3.0	2.9	3.4
449	2.1	2.1	2.1	2.1	2.1	2.5	1.3	1.4	3.2	1.3	2.9	1.8
308	1.5	1.1	1.0	1.0	1.0	1.0	1.7	2.8	2.7	1.7	2.4	1.3

Table 9.5: The panel rank means for Replicate 2 of the tomato soup ranking.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
254	4.5	4.1	4.9	5.0	5.0	4.9	4.5	4.5	2.8	4.9	2.9	4.5
310	3.9	4.2	3.8	4.0	3.3	3.9	4.4	4.4	2.8	3.8	3.3	4.3
679	3.2	3.5	3.2	2.6	3.6	3.2	3.1	3.1	2.8	3.3	3.1	2.8
239	2.1	2.0	2.1	2.4	1.3	1.7	1.8	1.8	2.3	2.0	3.2	2.1
792	1.5	1.2	1.0	1.0	1.8	1.3	1.2	1.2	4.2	1.0	2.5	1.3

For the tomato soup data, Panel M correlated best with the expected rank on both replicates. The performance of Panel H was very poor. The results of Panel L were poor on the second replicate.

One point worth noting, is that these results should also be taken in conjunction with performance of expected sample discrimination. This is because deviations from perfection may be due to random results where no difference existed between two samples.

Table 9.6: Correlations between each panel and the 'true/expected' rank order for tomato soup.

Panel	Replicate	Correlation	Significance
A	1	0.967	1%
	2	0.970	1%
B	1	0.943	1%
	2	0.992	1%
C	1	0.970	1%
	2	0.962	1%
D	1	0.990	1%
	2	0.914	5%
E	1	0.914	5%
	2	0.995	1%
F	1	0.864	5%
	2	0.988	1%
G	1	0.751	10%
	2	0.988	1%
H	1	0.443	ns
	2	-0.475	ns
K	1	0.963	1%
	2	0.993	1%
L	1	0.907	5%
	2	0.453	ns
M	1	0.984	1%
	2	0.981	1%

9.3 Apple Juice – Coefficient of Concordance

The same tables of panel ranks correlations (Tables 9.1 and 9.2) were used to calculate the Pearson, and the coefficient of concordance for each panel (Table 9.7).

Table 9.7 shows the concordance between assessors in each panel, and it can be seen that only Panel F performed well over both replicates. The assessors in Panel M were not in good agreement, though this was not poor enough to affect the overall panel ranking (Table 9.8).

Table 9.7: Coefficient of concordance for the panel – apple juice.

Panel	Replicate 1	Replicate 2
A	0.902	0.776
B	0.653	0.891
C	0.891	0.791
D	0.810	0.886
E	0.982	0.778
F	0.958	0.950
G	0.851	0.879
H	0.884	0.750
K	0.843	0.874
L	0.892	0.857
M	0.699	0.701

The agreement between the panel rank and the expected rank was not calculated as this only reflects the information from the correlation coefficients, calculated previously.

9.4 Tomato Soup – Coefficient of Concordance

The same tables of panel ranks (Tables 9.3 and 9.4) were used to calculate the Pearson correlations, and the coefficient of concordance for each panel (Tables 9.8).

The agreement between assessors in a panel (Table 9.8) is generally poorer for tomato soup, than for the apple juice. The results for Panel H demonstrate that the assessors did not agree with each other, and this was a very poor and unacceptable result. However, on the positive side, the assessors in Panels B, F and K showed good agreement for both replicate assessments.

Table 9.8: Coefficient of concordance for the panel – tomato soup.

Panel	Replicate 1	Replicate 2
A	0.748	0.714
B	0.943	0.936
C	0.847	0.953
D	0.954	0.879
E	0.874	0.904
F	0.926	0.958
G	0.689	0.896
H	0.378	0.189
K	0.944	0.939
L	0.656	0.688
M	0.856	0.785

10. PROCEDURE FOR PERFORMANCE CRITERIA

10.1 Introduction

This chapter outlines a first attempt at establishing a scheme to evaluate the performance of panels in a proficiency testing scheme of the sensory ranking test. In proposing this scheme, it should be highlighted that the actual measurement criteria are at this stage illustrative, as these will be set by the Proficiency Testing Provider based on the results of screening work undertaken prior to a ring trial.

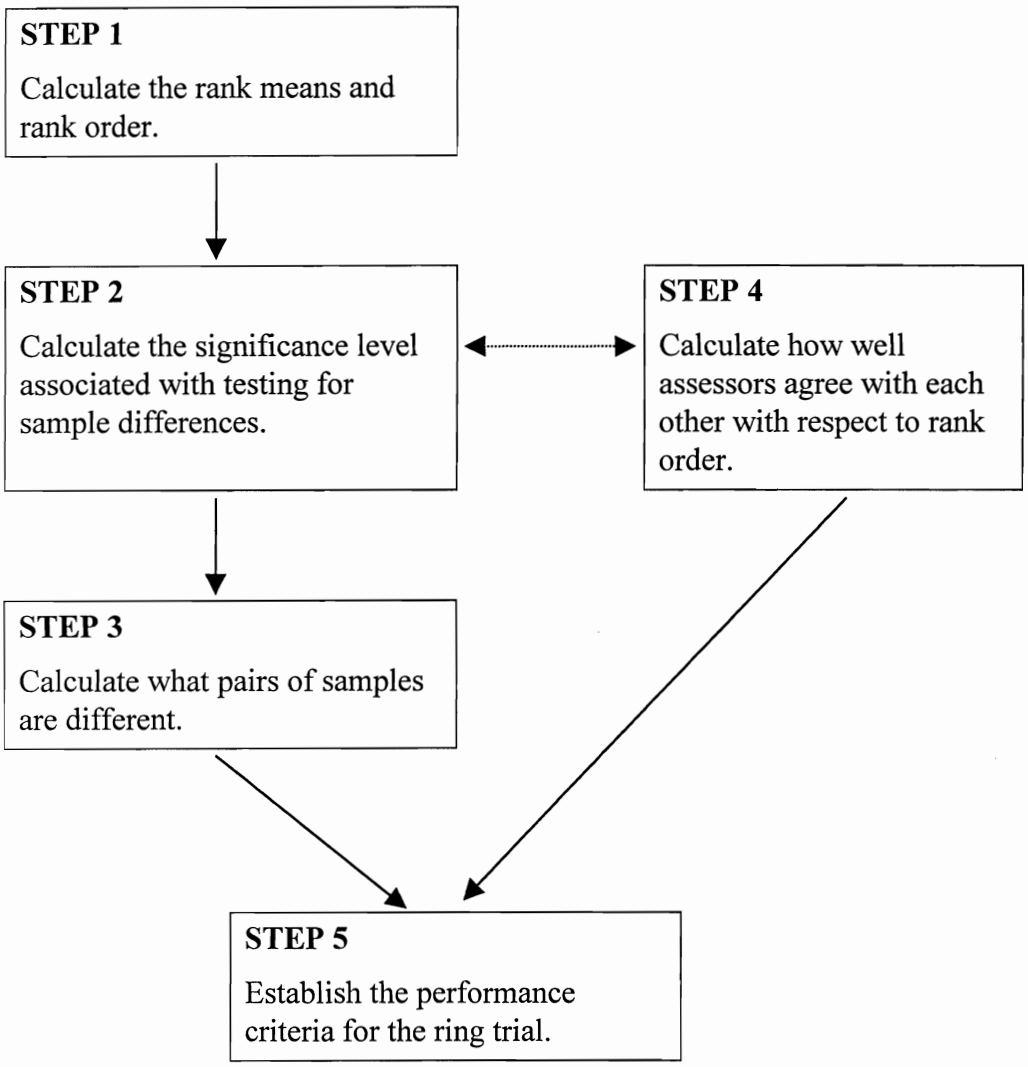
10.2 Establishing the Expected Result

The suggested stepwise procedure for establishing the expected result is demonstrated below through 5 key stages (see overleaf).

Step 1 – Calculate the Rank Means and Rank Order

For each panel in the pre-test, tabulate the rank data and work out the panel rank mean for each sample. If all pre-test panels agreed in their rank means, then the average over all pre-test panels can be set as the 'expected rank means'. If there is some disagreement, then steps 2 and 3 will help to establish if this is because samples were 'switched' in the ranking by assessors because there was no perceptible difference between them. Determine the Pearson correlation coefficient between the 'expected rank means' and the actual panel rank means at the 10% level of significance. This level of significance is chosen to eliminate the possibility of downgrading a panel because two or more samples were not perceptibly different.

Possible performance criteria: 0 if $p > 0.10$ or correlation is negative; 1 if $p \leq 0.10$.



Step 2 – Calculate the significance level associated with testing for sample differences

To establish how well each panel of assessors discriminated between the samples, a Friedman rank test should be undertaken and the level of significance recorded. If all panels performed well (i.e. $p \leq 0.01$ (1%)), then the results from an untrained panel may be required to help establish whether the pre-test was too easy (which would be the case if the task could be performed easily and accurately by an untrained panel), or whether the pre-test panels were just very good. If all pre-test panels perform poorly (i.e. $p > 0.10$ (10%)), then the nature of the samples may have made the ranking test too difficult (for example, the method of

preparation and serving may lead to sample inconsistencies). The Co-ordinator should be confident that the decisions based on the pre-test results will allow some panels in the main test to perform better than the expected result and still detect panels who perform worse than the expected result (see example), before deciding the 'expected significance level'

Possible performance criteria:

0 if $p > 0.10$	(10%)
1 if $p \leq 0.10$	(10%)
2 if $p \leq 0.05$	(5%)
3 if $p \leq 0.01$	(1%)
4 if $p \leq 0.001$	(0.1%)

Step 3 – Calculate which pairs of samples are different

Having established an expected significance level, the next step is to determine which pairs of samples are different at a specified level of significance (for example 1%, 5% and 10% significance). This can be achieved through the use of a suitable multiple comparison test, for example Conover's method. From these results the 'expected sample differences' can be set. At this point, the provider can confirm that the selected 'expected rank means' is satisfactory.

Possible performance criteria:

0 if no significant differences at 5% level
1 if 1 pair significantly different at 5% level
2 if 2 pairs significantly different at 5% level
3 if 3 pairs significantly different at 5% level
4 if 4 pairs significantly different at 5% level
5 if 1 pair significantly different at 1% level
6 if 2 or more pairs significantly different at 1% level

Step 4 – Calculate how well assessors agree with each other with respect to rank order

The Coefficient of Concordance is used to measure the agreement between assessors in a panel. Generally, a lack of agreement between assessors is reflected by a poor result in Steps 2 and 3. However, this method provides a single measure of how well the panel agrees to produce a given level of performance. The 'expected concordance level' is then set.

Possible performance criteria: 0 if $W < 0.70$
1 if $W \geq 0.70$
2 if $W \geq 0.80$
3 if $W \geq 0.90$
4 if $W \geq 0.95$

Step 5 – Setting the Performance Criteria

Finally, the information gathered in Steps 1-4 should be collated, and rules amended to define the level of performance linked to different categories of performance. For example: very good, good, average, poor, very poor. By allocating a score to each of these categories for steps 1 to 4, an overall performance criteria can be specified.

For example:	score = 13-15	very good	
	score = 10-12	good	
	score = 7-9	average	
	score = 4-6	poor	(unacceptable)
	score ≤ 3	very poor	(unacceptable)

Comments on Scheme

In general, if the pre-test results are 'good' or better, the pre-test laboratories can discriminate between the samples, can rank the samples in the right order, can detect differences between the specified samples and the assessors within the panel agree with each other. In this case, the Co-ordinator should go ahead with the main trial. If the pre-test results are 'poor' or worse, the selection of samples should be rethought and a repeat pre-test organised with new samples. If the results are 'average' the data should be carefully considered again to be confident that the panels in the main inter-comparison will be able to discriminate between the samples, before recommending that the main trial goes ahead.

Having set the performance criteria, and having made the decision to carry on with the main trial, one level of performance should be designated as the 'expected result'. In practice, this

will usually be the 'good' or 'average' category (the decision should be based on knowledge of the samples and of the panels participating in the pre-test, as well as on the statistical analyses of the pre-test data). It is important that the expected result is achievable. For example, if the expected result is set too high (e.g. 'very good'), then it is likely that few panels may be as good as 'expected' in the main inter-comparison. For this reason, in coming to the decision on the 'expected result', it is also important to consider what might reasonably be 'expected' of a trained sensory panel in which one would normally have confidence in its ability to perform sensory ranking tests.

10.3 Determining the Actual Panel Performance

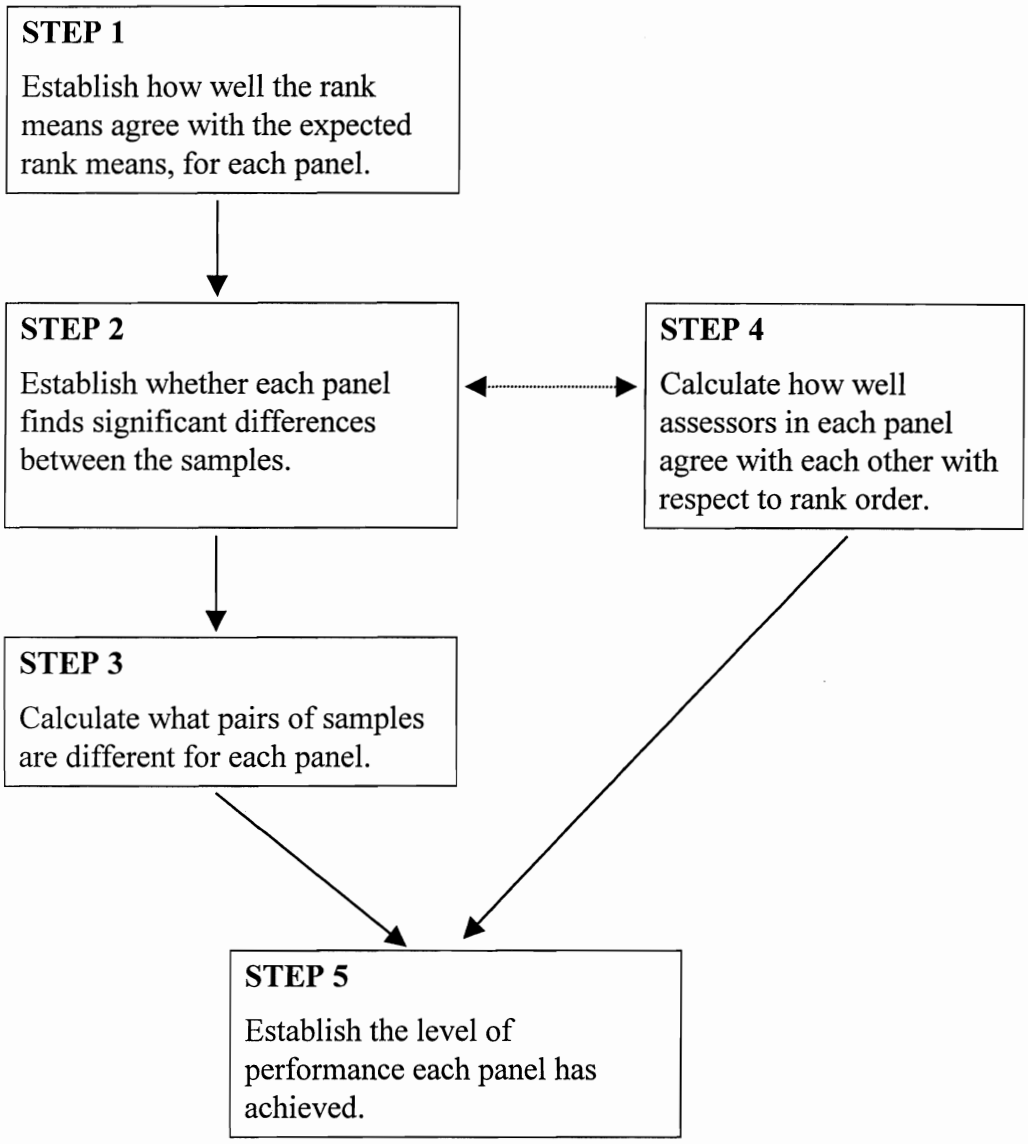
The stepwise procedure for establishing the actual performance of participants for ranking tests in relation to the 'expected result' determined from the pre-test is according to the following scheme (see overleaf).

Step 1 – Establish how well a panel's rank means agree with the expected rank means

For each participant in the main test, tabulate the data for each panel and calculate the panel rank for each panel. Calculate the Pearson correlation coefficient between the 'expected rank means' (from the pre-test) and the actual rank means, to establish how well they agree.

Possible performance criteria:

'Score' 0 if the correlation is negative
'Score' 0 if the significance level (p-value) associated with the correlation > 0.10
'Score' 1 if the significance level (p-value) associated with the correlation ≤ 0.10



Step 2 - Establish whether each panel finds significant differences between the samples

To establish how well each panel of assessors discriminated between the samples, perform the Friedman test on the data for each panel, note the level of significance achieved for sample discrimination, and record the performance score achieved.

Possible performance criteria:

'Score' 0 if $p > 0.10$	(10%)
'Score' 1 if $p \leq 0.10$	(10%)
'Score' 2 if $p \leq 0.05$	(5%)
'Score' 3 if $p \leq 0.01$	(1%)
'Score' 4 if $p \leq 0.001$	(0.1%)

Step 3 - Calculate what pairs of samples are different for each panel

Perform the multiple comparison test, for example Conover's method, using the multiple comparison value determined at the specified levels of significance in the pre-test, note which samples are different at each level for each panel, and record the performance score.

Possible performance criteria:

'Score' 0 if no significant differences at 5% level
'Score' 1 if 1 pair significantly different at 5% level
'Score' 2 if 2 pairs significantly different at 5% level
'Score' 3 if 3 pairs significantly different at 5% level
'Score' 4 if 4 pairs significantly different at 5% level
'Score' 5 if 1 pair significantly different at 1% level
'Score' 6 if 2 or more pairs significantly different at 1% level

It should be noted that the actual pairs should be specified for the actual proficiency scheme.

Step 4 - Calculate how well assessors in each panel agree with each other

Calculate the Coefficient of Concordance for each panel using the same procedure as used in the pre-test, and record the performance score.

Possible performance criteria:

'Score' 0 if $W < 0.70$
'Score' 1 if $W \geq 0.70$
'Score' 2 if $W \geq 0.80$
'Score' 3 if $W \geq 0.90$
'Score' 4 if $W \geq 0.95$

Step 5 - Establish the level of performance each panel has achieved

The data for each panel can now be compared to the 'expected results', and a score given to the performance in each of the 4 evaluation steps. A final overall score is then allocated and the performance level recorded.

10.4 Testing the Performance Criteria

The data from these trials did not lend itself to fully testing the proposed performance criteria. Therefore, it was necessary to undertake further trials to test the proposed performance scheme. However, considerable progress was made, and this report forms a sound basis from which to progress.

A further report (McEwan, 2001) will consider in more detail the establishment of expected results, and examine more closely the most appropriate way to establish the final performance of panels participating in proficiency tests.

REFERENCES

Conover, W.J. (1999). Practical Nonparametric Statistics. Third Edition. New York: John Wiley & Sons.

de Kroon, J. and van der Laan, P. (1981). Distrubution-free Test Procedures in Two-way Layouts: A Concept of Rank-Interaction. Statistica Neerlandica, **35**, 189-231.

Hirst, D and Næs, T. (1994). A Graphical Technique for Assessing Differences among a Set of Rankings. Journal of Chemometrics, **8**, 81-93.

Hochberg, Y. and Tamhane, A.C. (1987). Multiple Comparison Procedures. New York: John Wiley & Sons.

Hunter E.A. (1996). Experimental design. In: Næs, T. and Risvik, E. (Eds.). Multivariate Analysis of Data in Sensory Science, p. 37-69. Amsterdam: Elsevier.

Jones, B. and Kenward, M.G. (1989). Design and Analysis of Cross-over Trials. Chapman and Hall, London.

Jones, B. and Wang, J. (2000). The Analysis of Repeated Measurements in Sensory and Consumer Studies. Food Quality and Preference **11**, 35-41.

Kendall, M. and Gibbons, J.D. (1990). Rank Correlation Methods (5th Edition). London: Edward Arnold.

MacFie, H.J.H., Greenhoff, K., Bratchell, N. and Vallis, L. (1989). Designs to Balance the Effect of Order of Presentation and First-order Carry-over Effects in Hall Tests. Journal of Sensory Studies **4**, 129-148.

McEwan, J.A. (2000). Proficiency Testing for Sensory Profile Tests: Statistical Guidelines. Part 1. R&D Report No. 119. CCFRA.

McEwan, J.A.. (2001). Proficiency Testing for Sensory Ranking Tests: Statistical Guidelines. Part 2. R&D Report. CCFRA. In preparation.

Muir, D.D. and Hunter, E.A. (1991/2). Sensory Evaluation of Cheddar Cheese: Order of Tasting and Carryover Effects. Food Quality and Preference **3**, 141-145.

Næs, T., Hirst, D.and Baardseth, P. (1994). Using Cumulative Ranks to Detect Individual Differences in Sensory Profiling. Journal of Sensory Studies, **9**, 87-99.

O'Mahony, M. (1986). Sensory Evaluation of Food: Statistical Methods and Procedures. New York: Marcel Dekker, Inc.

Sprent, P. (1993). Applied Nonparametric Statistical Methods. London: Chapman & Hall.

Williams, E.J. (1949). Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. Australian Journal of Scientific Research, Series A, 2, 149-168.

APPENDIX 1: PANEL RANK SUMS FROM RING TRIALS

Apple Juice

Tables 1 and 2 show the rank sums, and multiple comparison value for each panel, for the 2 replicate assessments. If the difference between two rank sums are different, then the samples were significantly different at the 5% level of significance.

Table 1: Rank sums, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: apple juice, replicate 1.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
966	5	48	81	40	59	50	50	55	40	60	59	56
985	4	42	61	30	47	39	40	41	30	45	52	46
439	3	27	53	25	32	31	27	32	24	36	40	35
962	2	22	30	15	26	20	23	24	17	22	22	28
551	1	11	30	10	16	10	10	13	9	17	22	15
MC-5%	--	4.7	11.2	4.6	7.1	2.0	3.1	6.1	4.7	6.5	9.5	9.0
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2: Rank sums, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: apple juice, replicate 2.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
166	5	50	83	40	59	47	49	55	39	59	63	58
387	4	36	69	28	49	40	41	42	30	49	53	42
826	3	30	49	25	32	30	28	30	23	32	33	39
173	2	16	33	16	25	20	22	26	17	24	31	20
352	1	18	21	11	15	13	10	12	11	16	15	21
MC-5%	--	7.2	6.3	6.3	5.5	7.1	3.4	5.5	6.9	5.8	6.4	8.9
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Tomato Soup

Tables 3 and 4 show the rank sums, and multiple comparison value for each panel, for the 2 replicate assessments. If t the difference between two rank sums are different, then the samples were significantly different at the 5% level of significance.

Table 3: Rank sums, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: tomato soup, replicate 1.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
681	5	41	72	38	60	42	38	49	27	54	44	54
753	4	43	81	28	46	47	50	46	16	45	37	40
272	3	34	49	29	37	26	32	24	12	33	35	37
449	2	21	36	17	25	25	13	15	19	14	35	20
308	1	11	17	8	12	10	17	31	16	19	29	14
MC-5%	--	7.6	4.5	5.4	3.5	5.4	4.1	8.8	10.1	3.7	9.6	6.0
p-value	--	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.078	0.000	0.000	0.000

Table 4: Rank sums, Friedman rank test results (p-value) and multiple comparison (MC) value to compare samples: tomato soup, replicate 2.

Sample	True Value	Panel										
		A	B	C	D	E	F	G	H	K	L	M
254	5	41	84	40	60	49	50	50	17	54	35	50
310	4	42	65	32	40	39	48	48	17	42	40	47
679	3	35	54	21	43	32	34	34	17	36	37	31
239	2	20	35	19	16	17	20	20	14	22	38	23
792	1	12	17	8	21	13	13	13	25	11	30	14
MC-5%	--	8.1	4.8	3.0	5.7	4.7	3.1	5.1	11.3	3.9	9.1	7.3
p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.339	0.000	0.000	0.000

APPENDIX 2: MULTIPLE COMPARISON VALUES

The table below provides the Studentised range multiple comparison values to use when wishing to determine sample differences when ranking 5 samples. All values are appropriate for the 5% level of significance. These are provided for information, though the Conover method reported is the one chosen for this project (Section 3.4).

N-Assessor	N-Sample	Normal[1]	Chi-squared[1]	Studentised Range[1]
3	5	7.6	11.9	10.6
4	5	8.8	13.8	12.2
5	5	9.8	15.4	13.6
6	5	10.7	16.9	14.9
7	5	11.6	18.2	16.1
8	5	12.4	19.5	17.3
9	5	13.1	20.7	18.3
10	5	13.9	21.8	19.3
11	5	14.5	22.8	20.2
12	5	15.2	23.9	21.1
13	5	15.8	24.8	22.0
14	5	16.4	25.8	22.8
15	5	17.0	26.7	23.6
16	5	17.5	27.6	24.4
17	5	18.1	28.4	25.1
18	5	18.6	29.2	25.9
19	5	19.1	30.0	26.6
20	5	19.6	30.8	27.3

[1] Hochberg and Tamhane (1987)

To undertake the comparisons using mean ranks, the numbers in the above table should be divided by the appropriate number of assessors.

APPENDIX 3: CRITICAL VALUES FOR THE CORRELATION COEFFICIENT

Spearman

The table below provides the rank correlation coefficient values required to achieve different levels of significance relating two rankings with 5 samples. Note that this is based on a one-sided test.

Significance level	Correlation coefficient
1%	0.900
5%	0.800
10%	0.700
15%	0.600
20%	0.500

Pearson

The table below provides the Pearson product moment correlation coefficient values required to achieve different levels of significance relating two rankings with 5 samples. Note that this is based on a one-sided test.

Significance level	Correlation coefficient
1%	0.9325
5%	0.8031
10%	0.6854
15%	0.5839
20%	0.4926

APPENDIX 4: CRITICAL VALUES FOR THE COEFFICIENT OF CONCORDANCE

n = 5 samples

Assessors (m)	Significance Level	W ≥ ?
6	1%	0.489
6	5%	0.372
6	10%	0.317
6	15%	0.278
6	20%	0.250
7	1%	0.433
7	5%	0.327
7	10%	0.273
7	15%	0.237
7	20%	0.216
8	1%	0.384
8	5%	0.288
8	10%	0.241
8	15%	0.209
8	20%	0.188
9	1%	0.341
9	5%	0.254
9	10%	0.210
9	15%	0.185
9	20%	0.165
10	1%	0.310
10	5%	0.230
10	10%	0.190
10	15%	0.166
10	20%	0.150

Assessors (m)	Significance Level	W ≥ ?
11	1%	0.284
11	5%	0.210
11	10%	0.175
11	15%	0.152
11	20%	0.137
12	1%	0.264
12	5%	0.193
12	10%	0.160
12	15%	0.140
12	20%	0.125
13	1%	0.241
13	5%	0.178
13	10%	0.148
13	15%	0.129
13	20%	0.115
14	1%	0.228
14	5%	0.166
14	10%	0.138
14	15%	0.119
14	20%	0.107
15	1%	0.213
15	5%	0.156
15	10%	0.130
15	15%	0.111
15	20%	0.100

Assessors (m)	Significance Level	W ≥ ?
16	1%	0.201
16	5%	0.146
16	10%	0.121
16	15%	0.105
16	20%	0.094
17	1%	0.188
17	5%	0.138
17	10%	0.113
17	15%	0.099
17	20%	0.088
18	1%	0.180
18	5%	0.130
18	10%	0.107
18	15%	0.094
18	20%	0.083
19	1%	0.170
19	5%	0.124
19	10%	0.101
19	15%	0.089
19	20%	0.079
20	1%	0.161
20	5%	0.117
20	10%	0.096
20	15%	0.084
20	20%	0.075

APPENDIX 5: BACKGROUND TO W CRITICAL VALUES

Friedman's Statistic and Kendall's Coefficient of Concordance W

The tests both assume that the data consists of m assessors who rank n samples (without ties in this exposition). The data can be envisaged as a matrix of m rows (assessors) and n columns (samples). The entries, r_{ij}, are the ranks assigned by the i^{th} assessor to the j^{th} sample. The column sums, R_j, are used in calculating both these statistics.

$$R_j = \sum_{i=1}^{m} r_{ij} \qquad \text{j = 1....n}$$

Friedman's statistic is often designated T_1 and is computed using the corrected sum of square of ranks S where

$$S = \sum_{j=1}^{n}\left(R_j - \frac{m(n+1)}{2}\right)^2 = \sum_{j=1}^{n} R_j^2 - nm^2(n+1)^2/4$$

and $$T_1 = \frac{12S}{mn(n+1)}$$

This statistic is asymptotically distributed as a χ^2 with n-1 degrees of freedom. We later show that this approximation is not sufficiently accurate to be useful in the particular circumstances of the PROFISENS project.

For estimating statistical significance from critical values, a transformation of T_1 (i.e. T_2) is to be preferred. T_2 is defined as:

$$T_2 = \frac{(m-1)T_1}{m(n-1) - T_1} \quad \text{and conversely} \quad T_1 = \frac{m(n-1)T_2}{(m-1) + T_2}$$

It is later shown that this is a more acceptable approximation.

Kendall's Coefficient of Concordance (W) was developed independently of Friedman's statistic with a different purpose in mind. It is defined as:

$$W = \frac{12S}{m^2(n^3 - n)} = \frac{12S}{m^2 n(n+1)(n-1)}$$

It is bounded by 0 and 1 for all values of m and n and thus offers the advantage of allowing comparisons between panels with different numbers of assessors. W can be expressed in term of T_1 by:

$$W = \frac{T_1}{m(n-1)}$$

and in terms of T_2

$$W = \frac{T_2}{m-1+T_2}$$

Thus, the Friedman's statistic and Kendall's Coefficient of Concordance yield the same level of significance when applied to the same data. As noted above there can be advantages in computing W in order to allow comparison of panels with different numbers of assessors.

A set of tables (Tables 1-8) has been produced to allow the study of critical values for 5 samples (n) and 2-20 assessors (m) so that well founded recommendations can be made.

Tables 1 – 3 contain the results of a simulation study (1,000,000 simulations) for 2 –20 assessors (m). The results are expressed for Friedman's statistic (T_1) Table 1, Coefficient of Concordance (W) – Table 2 and for the corrected sum of squared rank sums (S) – Table 3. Tables 1-3 contain the same information but expressed in terms of different statistics.

Table 4 contains the critical values using the χ^2 approximation T_1, the traditional Friedman's statistic. Note that this approximation is independent of the value of assessors (m) and depends only on the number of samples (n). Table 5 contains the critical values using the F approximation to T_2 but the critical values are expressed in terms of T_1 the traditional Friedman's statistic. The results are in good (but not excellent) agreement with those of Table 1. Tables 6-8 give published values of Friedman (T_1), Kendall (W) and the corrected sum of squares of rank totals (S) from the literature. It is notable that critical values of these statistics are not readily available for 5 samples (n) and 2-20 assessors (m). The values available have been tabulated to validate the simulation process and largely do so. These values appear to have been smoothed and hence do not fully reflect the integer nature of the data.

It is our opinion that the simulation values are the most trustworthy estimates of the critical values. The χ^2 approximation is too imprecise to be useful and so the F approximation is to be preferred in the absence of values from a simulation study.

Table 1 - Friedman Statistics (5 samples) – Critical Values

The critical values are derived in parallel with Kendall's Coefficient of Concordance and S. Based on 1 million simulations.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2	4.00	5.20	6.00	6.40	6.80	7.20	7.60	8.00
3	3.47	5.07	6.13	6.67	7.20	8.27	9.87	10.93
4	3.40	5.00	6.00	6.60	7.40	8.60	11.00	13.00
5	3.52	4.96	5.92	6.72	7.52	8.80	11.52	14.24
6	3.47	4.93	6.00	6.67	7.60	8.93	11.73	15.07
7	3.54	4.91	6.06	6.63	7.66	9.03	12.00	15.43
8	3.50	4.90	6.00	6.70	7.60	9.10	12.20	15.80
9	3.38	4.89	5.96	6.67	7.64	9.16	12.36	16.27
10	3.44	4.96	6.00	6.64	7.68	9.20	12.40	16.40
11	3.42	4.87	6.04	6.69	7.71	9.24	12.51	16.65
12	3.47	4.87	6.00	6.73	7.67	9.27	12.60	16.80
13	3.38	4.92	5.97	6.71	7.69	9.29	12.68	17.05
14	3.43	4.91	6.00	6.69	7.71	9.31	12.69	17.03
15	3.41	4.91	6.03	6.72	7.73	9.33	12.75	17.17
16	3.40	4.90	6.00	6.75	7.75	9.35	12.80	17.20
17	3.44	4.89	5.98	6.73	7.72	9.32	12.80	17.27
18	3.42	4.89	6.00	6.71	7.69	9.33	12.84	17.38
19	3.41	4.88	6.02	6.74	7.71	9.35	12.84	17.39
20	3.40	4.92	5.96	6.72	7.72	9.36	12.88	17.48

Table 2 -Kendall's Coefficient of Concordance W (5 Samples) – Critical Values

The critical values are derived in parallel with Friedman's Statistic and S.

Based on 1 million simulations.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2	0.50	0.65	0.75	0.80	0.85	0.09	0.95	1.00
3	0.29	0.42	0.51	0.56	0.60	0.69	0.82	0.91
4	0.21	0.31	0.38	0.41	0.46	0.54	0.69	0.81
5	0.18	0.25	0.30	0.34	0.38	0.44	0.58	0.71
6	0.14	0.21	0.25	0.28	0.32	0.37	0.49	0.63
7	0.13	0.18	0.22	0.24	0.27	0.32	0.43	0.55
8	0.11	0.15	0.19	0.21	0.24	0.28	0.38	0.49
9	0.09	0.14	0.17	0.19	0.21	0.25	0.34	0.45
10	0.09	0.12	0.15	0.17	0.19	0.23	0.31	0.41
11	0.08	0.11	0.14	0.15	0.18	0.21	0.28	0.38
12	0.07	0.10	0.13	0.14	0.16	0.19	0.26	0.35
13	0.07	0.09	0.11	0.13	0.15	0.18	0.24	0.33
14	0.06	0.09	0.11	0.12	0.14	0.17	0.23	0.30
15	0.06	0.08	0.10	0.11	0.13	0.16	0.21	0.29
16	0.05	0.08	0.09	0.11	0.12	0.15	0.20	0.27
17	0.05	0.07	0.09	0.10	0.11	0.14	0.19	0.25
18	0.05	0.07	0.08	0.09	0.11	0.13	0.18	0.24
19	0.04	0.06	0.08	0.09	0.10	0.12	0.17	0.23
20	0.04	0.06	0.07	0.08	0.10	0.12	0.16	0.22

Table 3 – S (5 samples) – Critical Values

The critical values are derived in parallel with Friedman's Statistic and Kendall's coefficient of concordance. Based on 1 million simulations.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2	20	26	30	32	34	36	38	40
3	26	38	46	50	54	62	74	82
4	34	50	60	66	74	86	110	130
5	44	62	74	84	94	110	144	178
6	52	74	90	100	114	134	176	226
7	62	86	106	116	134	158	210	270
8	70	98	120	134	152	182	244	316
9	76	110	134	150	172	206	278	366
10	86	124	150	166	192	230	310	410
11	94	134	166	184	212	254	344	458
12	104	146	180	202	230	278	378	504
13	110	160	194	218	250	302	412	554
14	120	172	210	234	270	326	444	596
15	128	184	226	252	290	350	478	644
16	136	196	240	270	310	374	512	688
17	146	208	254	286	328	396	544	734
18	154	220	270	302	346	420	578	782
19	162	232	286	320	366	444	610	826
20	170	246	298	336	386	468	644	874

Table 4 – Friedman's Statistic (5 Samples) – Critical Values

The critical values are derived using χ^2 approximation.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
3	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
4	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
5	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
6	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
7	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
8	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
9	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
10	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
11	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
12	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
13	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
14	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
15	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
16	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
17	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
18	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
19	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47
20	3.36	4.88	5.99	6.75	7.78	9.49	13.28	18.47

From the table it can be seen that the values for each significance level are independent of the number of assessors.

Table 5 – Friedman's Statistic (5 Samples) Critical Values

The critical values are derived using variance ratio (F) approximation.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2	4.00	5.09	5.70	6.05	6.43	6.92	7.53	7.85
3	3.77	5.06	5.88	6.39	7.01	7.89	9.34	10.54
4	3.66	5.03	5.93	6.51	7.24	8.33	10.29	12.20
5	3.59	5.00	5.96	6.57	7.37	8.50	10.88	13.30
6	3.55	4.98	5.97	6.61	7.45	8.75	11.28	14.08
7	3.52	4.97	5.97	6.63	7.50	8.86	11.56	14.66
8	3.50	4.97	5.98	6.65	7.54	8.94	11.77	15.10
9	3.48	4.95	5.98	6.66	7.57	9.00	11.94	15.45
10	3.47	4.94	5.98	6.67	7.59	9.06	12.07	15.74
11	3.46	4.94	5.98	6.68	7.61	9.10	12.18	15.97
12	3.45	4.93	5.98	6.69	7.62	9.13	12.27	16.18
13	3.44	4.93	5.99	6.69	7.64	9.16	12.35	16.34
14	3.44	4.93	5.99	6.69	7.65	9.18	12.42	16.49
15	3.43	4.92	5.99	6.70	7.66	9.20	12.47	16.61
16	3.43	4.92	5.99	6.70	7.67	9.22	12.52	16.73
17	3.42	4.92	5.99	6.70	7.67	9.24	12.57	16.82
18	3.42	4.92	5.99	6.71	7.68	9.25	12.61	16.91
19	3.42	4.91	5.99	6.71	7.68	9.27	12.64	16.99
20	3.41	4.91	5.99	6.71	7.69	9.28	12.67	17.06

Table 6 – Friedman Statistics (5 Samples) – Critical Values

This table is from published information, and blanks indicate unavailable data.

Assessor	Probability (%)							
	50	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2						7.600[b]	8.000[b]	
3					7.467[a]	8.533 [ab]	10.13 [ab]	11.47 [a]
4					7.600 [a]	8.000 [ab]	11.20 [ab]	13.20 [a]
5					7.680 [a]	8.960 [ab]	11.68 [ab]	14.40 [a]
6	3.600[c]	5.067[c]	6.133[c]	6.800[c]	7.733 [ac]	9.067 [abc]	11.87 [abc]	15.20 [ac]
7	3.657[c]	5.029[c]	6.171[c]	6.743[c]	7.771 [ac]	9.143 [abc]	12.11 [abc]	15.66 [ac]
8	3.600[c]	5.000[c]	6.100[c]	6.800[c]	7.700 [ac]	9.200 [abc]	12.30 [abc]	16.00 [ac]
9					7.733 [a]	9.244 [ab]	12.44 [ab]	16.36 [a]
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								

[a] Table 24 – Lindley, D.W. and Scott, W.F. (1984). New Cambridge Statistical Tables.

[b] Table 4.3 – Neave, H.R. (1988). Statistics Tables (2nd Edition), Allan and Unwin, London

[c] Odeh, R.E. (1997). Extended tables of the distribution of Friedman's S-statistic in the two-way layout. Communications in Statistics and Simulation Computations, B6(1), 29-48.

Table 7 – Coefficient of Concordance W (5 Samples) – Critical Values

No published data were found.

Assessor	Probability (%)							
	50	30	20	15	10	5	1	0.1
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								

Table 8 – S (5 Samples) – Critical Values

This table is from published information, and blanks indicate unavailable data.

Assessor	Probability (%)							
	50	30	20	15	10	5	1	0.1
2								
3						64.4 [a]	75.6 [a]	
4						88.4 [a]	109.3 [a]	
5						112.3 [a]	142.8 [a]	
6						136.1 [a]	176.1 [a]	
7								
8						183.7 [a]	242.7 [a]	
9								
10						231.2 [a]	309.1 [a]	
11								
12								
13								
14								
15						349.8 [a]	475.2 [a]	
16								
17								
18								
19								
20						468.5 [a]	641.2 [a]	

[a] Table R – Siegal, S. (1956). Non-parametric Statistics for the Behavioural Sciences.

Table 9 – S (5 Samples) – Critical Values

This table is based on a personal communication (Tony Hunter, BioSS) from Mark A. van de Wiel. Not yet published. Blanks indicate no available data found.

	Probability (%)							
Assessor	**50**	**30**	**20**	**15**	**10**	**5**	**1**	**0.1**
2					36	38	40	*
3					56	64	76	86
4					76	88	112	132
5					96	112	146	180
6					116	136	178	228
7					136	160	212	274
8					154	184	246	320
9					174	208	280	368
10					194	232	312	414
11					214	256	346	460
12					232	280	378	506
13								
14								
15								
16								
17								
18								
19								
20								

APPENDIX 6: THE 'EGG SHELL' PLOT PROCEDURE ADAPTED

Principal

The 'egg shell' plot (Næs *et at.*, 1994) is based on using cumulative ranks, and judging each assessor against a baseline. It was first used to evaluate assessor performance in descriptive analysis, but may have potential for ranked data. There are a number of features of 'egg shell' plots that can be used to determine performance, and these will become apparent in the following sections.

Procedure

This method offers a way of achieving a consensus ranking for a set of samples free of arbitrary decisions.

The first step is to tabulate the rankings for each assessor, together with the defined 'true' ranking. Næs ***et al.*** (1994) reports that if a consensus rank is required (no true rank available), then the consensus is obtained by ranking the scores of the first principal component, after performing principal component analysis where assessors are the variables. However, for this application, the 'true' ranking will be used.

The next step is to replace each assessor's ranks by cumulative ranks, cumulated in the true rank order (or consensus order). This can be achieved by ensuring the table of ranks is sorted by the 'true' rank order.

The cumulative rank is then calculated for an assessor who ranks all the samples the same, in other words finds no difference between them. This can be done using the following formula.

$(1 + n)/2$; $[(1 + n)/2] * 2$; …….. n = number of samples

Finally, the cumulative scores are usually plotted against the true rank, and for each assessor the cumulative scores are joined together to give one line per assessor.

Interpretation of Egg Shell Plots

The area between the lower curve, or the shell itself, and the curve corresponding to a given assessor, is a measure of this particular assessor's agreement with the consensus of the panel (or the with the true value, if such a value exists) see Figure 1. Agreement is here defined as ranking the samples in the same order as the consensus/true value. The extreme disagreement with the consensus is an assessor who has ranked the samples in exactly the opposite order as the consensus. This assessor's contribution to the eggshell plot would be a whole egg, Figure 2. Note, however, that the likeness to a whole egg is not too obvious with only 5 samples being ranked.

Figure 1

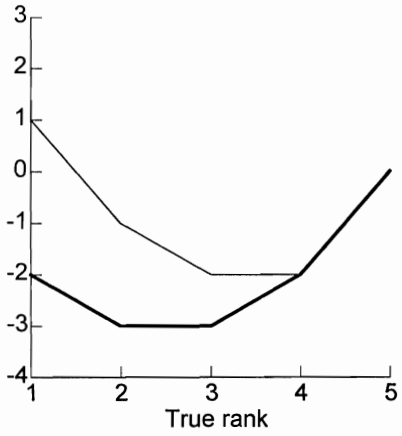
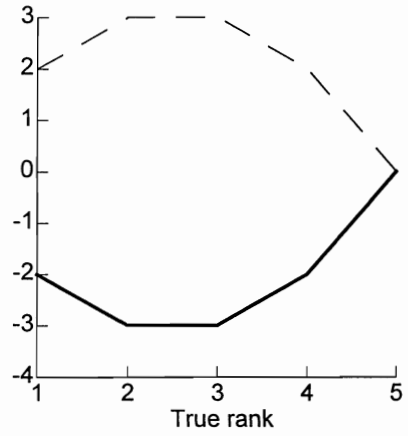


Figure 2: Total disagreement



To distinguish all the assessors in a panel is often difficult, or downright impossible. Therefore, one often finds the eggshell plots divided into two parts: one plot showing all the assessors in the panel, and one part consisting of small plots each representing a single assessor, Figures 3 and 4.

Figure 3 - All assessors

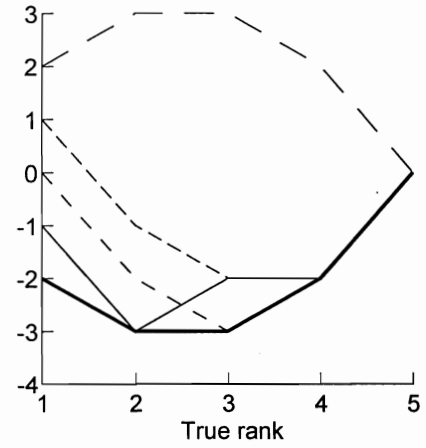
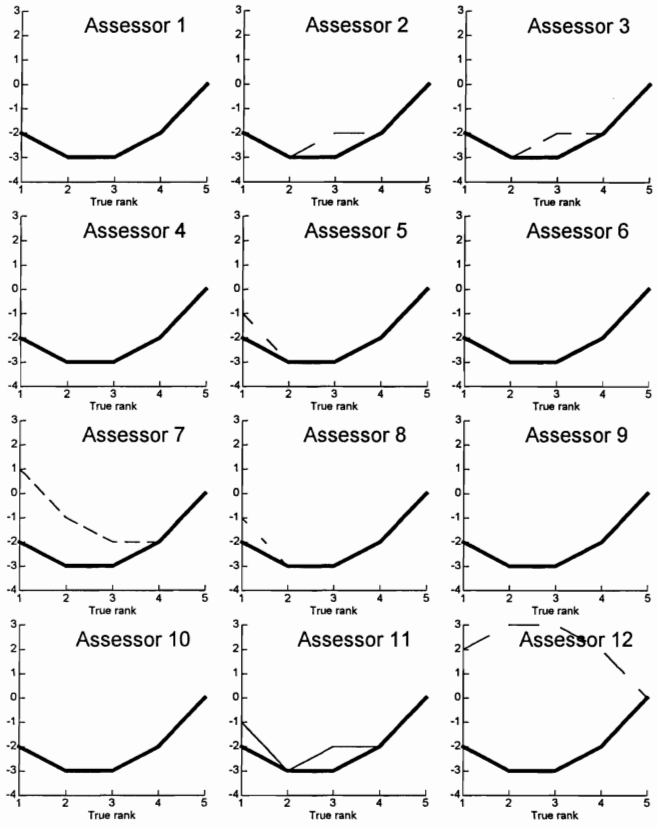


Figure 4



It can be proved that the area between a single assessor's curve and the outer shell is proportional to 1 minus Spearman's rank correlation (Hirst and Næs, 1994). In other words: small area indicates large rank correlation; large area indicates small rank correlation. This is the theoretical basis for stating that in Figure 4, Assessor 7 is in a sense 'worse' than Assessor 5.

APPENDIX 7: THE RANK INTERACTION TEST

Principal

If all assessors in a panel rank the samples in the same order, then there should be no interaction between the assessors and the samples. However, if one or more assessors ranks the samples differently, then a cross-over interaction will occur between the samples and assessors. Such an interaction indicates that the panel is not agreed on how the samples should be ordered with respect to the named attribute. In analysis of variance, it is easy to measure and test for such an interaction, yet this procedure for ranking data is less well known.

Procedure

In experiments where there are replicated rankings for each assessor it is possible to partition the variation between rankings into two parts. One part is due to variation within assessors and another is due to variation between assessors. The latter can be thought of as an interaction between sample effects and assessor effects.

The approach to this problem is repeated use of the Friedman test and was stimulated by reading of De Kroon and Van Der Laan (1981), who tackled the problem of ranking in the context of data collected as continuous variates.

If the Friedman test is applied to each assessor in turn, a Chi-Squared (χ^2) statistic is obtained testing how much each assessor discriminates between samples. Assessors who do fail to rank the samples in a consistent way will have a small χ^2, whilst an assessor who ranks the samples in the same order in each replicate will have a high χ^2. For interest χ^2 values from a Friedman test are bounded by zero and ∞. The individual χ^2 values for each assessor are

summed to give T. Likewise the Friedman process can be applied to the whole data set ignoring the distinction between assessors and replicates – T_1. This provides an overall test of treatment effects. By taking the difference between these two χ^2 figures T_2=T-T_1, a measure of the interaction between assessors and samples is obtained.

Formulae

n = number of samples ranked

m = number of assessors

r = number of replicates

r_{ijk} – rank for i^{th} sample, j^{th} assessor and k^{th} replicate

$$T = T^1 + T^2 + \ldots\ldots + T^m$$

$$= \sum_{j=1}^{m}\left[\frac{12}{r*n*(n+1)}\sum_{i=1}^{n}\left(\sum_{k=1}^{r} r_{ijk}\right)^2 - 3*r*(n+1)\right]$$

$$= \frac{12}{r*n*(n+1)}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\sum_{k=1}^{r} r_{ijk}\right)^2 - 3*m*r*(n+1)$$

$$T_1 = \frac{12}{m*r*n*(n+1)}\sum_{i=1}^{n}\left(\sum_{i=1}^{m}\sum_{k=1}^{r} r_{ijk}\right)^2 - 3*m*r*(n+1)$$

$$T_2 = T - T_1$$

$$= \frac{12}{r*n*(n+1)}\left[\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\sum_{k=1}^{r} r_{ijk}\right) - \frac{1}{m}\sum_{i=1}^{n}\left(\sum_{i=1}^{m}\sum_{k=1}^{r} r_{ijk}\right)^2\right]$$

$T^{1\ldots m}$ and T_1 are asymptotically distributed as χ^2 with n – 1 df

Consequently T is asymptotically distributed as χ^2 with m(n – 1) df

T_2 is therefore asymptotically distributed as χ^2 with (m – 1)(n – 1) df

Example Calculations

Table 1: Rankings of 6 samples as collected by 3 assessors on 3 replicate occasions.

Assessor	Rep	Sample1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	1	2	6	3	4	5	1
	2	3	6	4	5	1	2
	3	1	6	3	5	2	4
2	1	2	6	1	5	4	3
	2	2	5	3	6	4	1
	3	3	5	1	6	2	4
3	1	4	5	1	6	2	3
	2	2	6	3	5	1	4
	3	4	3	1	5	6	2

Table 2: Sum of ranks over replicates.

Assessor	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
1	6	18	10	14	8	7
2	7	16	5	17	10	8
3	10	14	5	16	9	9

Table 3: Sum of ranks over replicates and assessors.

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
23	48	20	47	27	24

Calculations

$$T^1 = \frac{12}{3x6x(6+1)}\left(6^2+18^2+10^2+14^2+8^2+7^2\right) - 3x3x(6+1)$$

$$= 10.24$$

$$T^2 = 11.57$$

$$T^3 = 7.38$$

$$T = T^1 + T^2 + T^3$$

$$= 29.19$$

$$T_1 = \frac{12}{3x3x6*(6+1)}\left(23^2+48^2+20^2+47^2+27^2+24^2\right) - 3x3x(6+1)$$

$$= 25.19$$

$$T_2 = T - T_1$$

$$= 4.00$$

Interpretation

The null hypothesis H_0 is the assessors 1, 2 ....3 cannot order the samples consistently.

Although T_1, T_2 and T_3 are asymptotically distributed as χ^2 with (6-1) = 5 df, this approximation is not sufficiently good for our purpose. Instead values of the test statistic from published Tables must be used:

$T_i = 6.33$ $p = 0.052$

$T_i = 8.33$ $p = 0.012$

$T_i = 10.33$ $p = 0.0017$

Consequently all the assessors show statistical evidence of being able to rank the samples consistently ($p>0.05$). Assessors 1 and 2 are more consistent than assessor 3.

Overall (aggregating over both replicates and assessors) the null hypothesis H_0 is that the samples are not ordered consistently.

The test statistic T_1 which is again asymptotically distributed as χ^2 with (6-1) = 5 df., is clearly statistically significant – $p<0.001$ (see test values above.

The interaction can be tested by the statistic $T_2 = 4.00$. This is asymptotically distributed with (3-1) x (6-1) = 15 df. Clearly there is no statistical evidence of interaction.