Members Only
R&D REPORT
NO. 126

Proficiency testing for sensory ranking tests: statistical guidelines - Part 2

2001

Campden BRI

Campden BRI

Chipping Campden
Gloucestershire
GL55 6LD, UK

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

Members Only
R&D Report No. 126

Proficiency testing for sensory ranking tests: statistical guidelines - Part 2

J McEwan

2001

EXECUTIVE SUMMARY

Proficiency testing in sensory analysis is an important step in demonstrating that data obtained from human instruments are as reliable as one would expect fkom any 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.

This report follows on from a previous document that proposed a procedure to determine the 'expected result' on a ranking test, and subsequently measure panel performance. This document concentrates on testing and validating the proposed procedure through the use of ring trials on apple juice and custard.

Through the use of a validation stage, using both trained and untrained panels, it was possible to demonstrate how to set up the expected result, and to set criteria and limits to measure panel performance.

The results of subsequent ring trials are also reported to demonstrate how the overall performance measured for each panel was achieved.

This research demonstrated that it was possible to establish performance criteria using the concept of validation panels. It was also shown that the use of untrained panels was helpful in setting the performance scheme.

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 were:

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 . Background to Proficiency Testing

Proficiency testing in sensory analysis is an important step towards demonstrating that data obtained from human instruments are as reliable as one would expect from any 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.

A previous report (McEwan, 2000) considered some of the issues concerning proficiency testing for ranking tests undertaken by trained panels. In particular it was discussed that the concept of 'expected result' was much more meaningful for sensory tests than the more common 'true value' measurement used in chemical and other instrumental proficiency testing schemes.

1.2 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 for both research and 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 perforniance of the panel, and therefore the panel will have performed less well.

However, the agreement between members of the panel, as measured by the coefficient of concordance, is of interest, as one measure of a panel's performance. This is a measure of whether the members of the panel are working as a team, and therefore ranking the samples in the same order.

This document is, therefore, mainly concerned with the performance of panels, and not individual assessors within the panel.

1.3 Report Scope

A previous document (McEwan, 2000) proposed a procedure to determine the 'expected result' of a ranking test, and subsequently measure panel performance. This document concentrates on testing and validating the proposed procedure.

Through the use of a validation stage, using both trained and untrained panels, it was possible to demonstrate how to set up the expected result, and to set criteria and limits to measure panel performance. The results of the subsequent ring trial are also reported, and the document demonstrates how the overall performance measured for each panel was achieved.

Details of the statistical methods used in this report can be found in McEwan (2000), where an explanation of the methods can be found together with calculations, where appropriate.

2. STAGES IN ESTABLISHING PANEL PERFORMANCE

2.1 Introduction

This chapter outlines a procedure to evaluate the performance of panels in a proficiency testing scheme of the sensory ranking test. The actual measurement criteria will be illustrated as part of two case studies reported in this document.

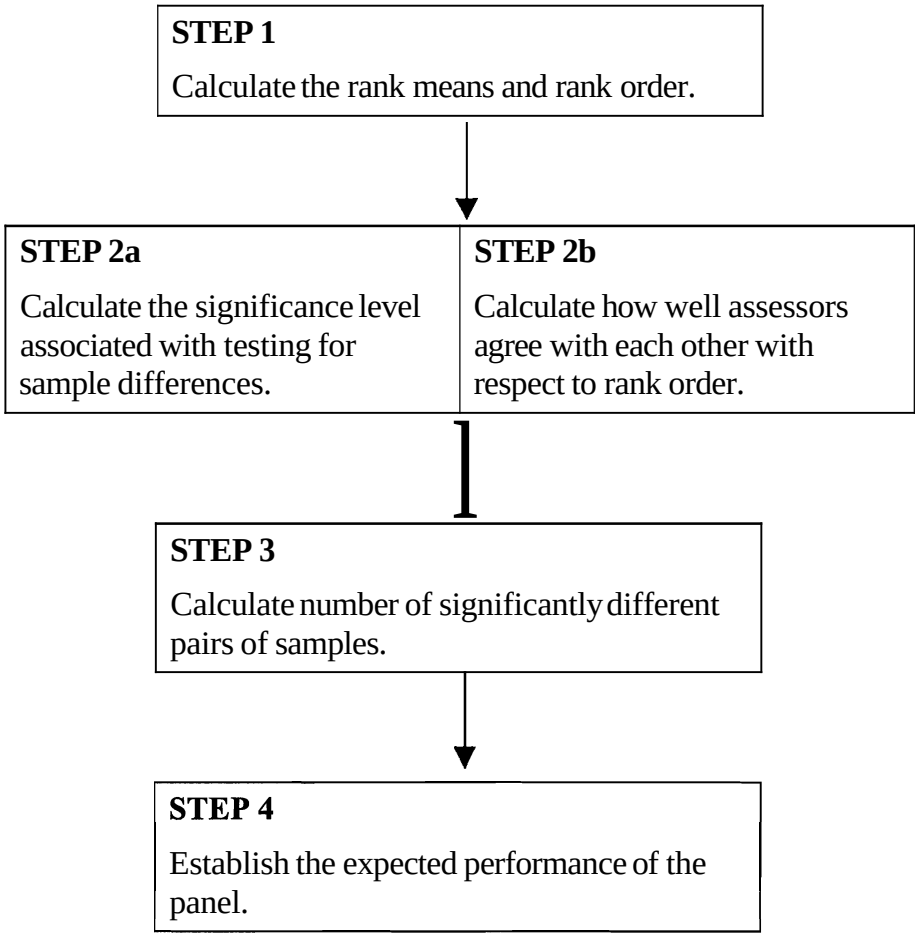
It is important to clarify a change in terminology use from Part 1 (McEwan, 2000). In Part 1, 'screening' was used to describe the stage where levels of spiking in samples (or sample selection) were investigated to ensure that perceptible differences between levels were neither too small nor too large. This was followed by a 'pre-test' stage where selected panels undertook the ranking test and the data were analysed with two aims: firstly to ensure that sample differences were sufficiently challenging; and secondly to determine the 'expected results'.

In this document (Part 2), the term 'pre-testing' is used to refer to the task related to the selection of suitable samples by the laboratory of the Scheme Provider. This stage may involve some panels undertaking a ranking exercise to ensure that sample selection is sound (Lyon, 2001). Once the samples are selected, validation panels will undertake the ranking exercise to obtain data to determine the expected results. This stage is known as the 'validation' stage (Lyon, 2001).

2.2 Establishing the Expected Result

The stepwise procedure for establishing the expected result is shown below through 4 key stages (Figure 2.1). The validation laboratories are those organisations providing trained and/or untrained assessors for the purpose of setting the expected results.

Figure 2.1: Stepwise procedure for establishing the expected result and setting performance criteria.



Step 1 – Calculate the Rank Means and Rank Order

For each validation panel, tabulate the rank data and work out the panel rank mean for each sample. If all validation panels agree in their rank means, then the average over all validation 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.

The Pearson correlation coefficient is then determined between the 'expected rank means' and

the actual panel rank means and tested 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 close together. In addition, a significant negative correlation (10% level) would indicate that the panel had ranked the samples in the wrong order (or forgot to recode the data). In this event, the participating laboratory to the proficiency test would be failed.

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

To establish how well each validation panel of assessors discriminated between the samples, a Friedman rank test should be undertaken and the level of significance recorded. If all validation panels performed well (i.e. $p \leq 0.01$ (1%)), then the results from an untrained panel may be required to help check whether the ranking 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 validation panels were just very good. If all validation panels perform poorly (i.e. $p > 0.10$ (10%)), then the nature of the samples may have made the ranking test too difficult. Alternatively, the method of preparation and serving may lead to heterogeneity within the samples. The Co-ordinator should be confident that the decisions based on the validation 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'.

Kendall's Coefficient of Concordance (W) can be used to measure the agreement between assessors in a panel, which is related to the overall level of discrimination. Generally, a lack of agreement between assessors would be reflected by a poor result in Steps 2a and **3.** However, W provides a single measure of how well the panel works together to produce a given level of performance. The 'expected concordance level' is then set.

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 (Conover, 1999). 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.

Step 4 – Establishing the expected performance of the panel

Finally, the information gathered in Steps 1-3 should be collated, and rules applied to define the expected level of performance. By allocating a score to each of these categories for Steps 1 to 3, an 'expected overall performance' can be specified.

Comments on the Procedure

In general, if the expected overall performance scores are high, the validation laboratories can discriminate between the samples, can rank the samples in the right order, and can detect differences between the specified samples, and the assessors within the panel agree with each other. This will normally indicate that the Co-ordinator should go ahead with the main trial, unless untrained panels also perform well (see paragraph below). If the validation panels' performance is low, particularly for the trained panels, the selection of may need to be revisited and a repeat validation stage organised with new samples. If the expected overall performance scores 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.

It is also possible that the task is too easy (reflected by high performance scoress), and therefore the ring trial would not be sufficiently demanding. In such circumstances, the Co-ordinator may recommend that the sample differences are made smaller.

Having set the performance criteria, and having made the decision to carry on with the main trial, one performance level should be designated as the 'expected result' for each step in the performance scheme. Participating laboratories will therefore be judged on their performance in each of the critical performance measures, and not solely on the basis of 'overall performance'.

It is important that the expected result is achievable, for each of the performance criteria. For example, if the expected result is set too high (e.g. 'very good' level), 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 whose ability one would normally have confidence in performing sensory ranking tests.

Normally, only one validation stage would be necessary, as prior screening and pre-testing should have sorted out any problenis of sample differences being too large or too small.

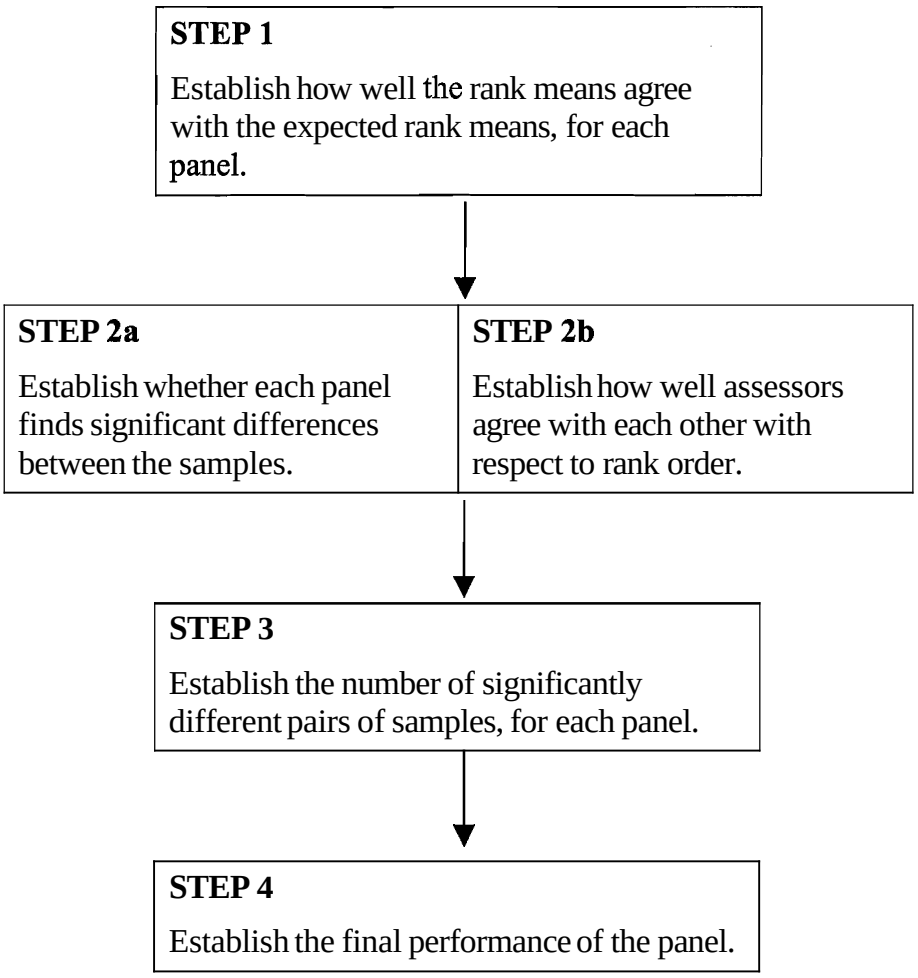
2.3 Determining the Actual Panel Performance

The stepwise procedure for establishing the actual performance of participants for ranking tests in relation to the 'expected result' determined from the validation panel results is according to the following scheme (Figure 2.2).

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 validation stage) and the actual rank means, to establish how well they agree.

Figure 2.2: Stepwise procedure for establishing the perfomance score for each panel



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.

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

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

Perform the multiple comparison test using the multiple comparison value determined at the specified levels of significance in the validation stage, note which samples are different at each level for each panel, and record the performance score.

Step 4 - 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 determined to measure the overall performance of each panel.

3. EXPECTED RESULTS FOR THE TRIALS PERFORMED IN 2000

3.1 Samples and Data

Samples

For the first trial, five samples of apple juice were spiked with different levels of glucose and fructose blends (Table 3.1). Each mixture comprised 50 ml of apple juice, 50 ml of water and 6.5 g of the sugar blend.

Table 3.1: Sugar blends and 3-digit numbers used to code the products for 2 replicate assessments.

	Sugar Blend		Code	
Sample	**Glucose**	**Fructose**	**Rep 1**	**Rep 2**
1	75%	25%	611	853
2	63%	37%	208	460
3	50%	50%	436	798
4	37%	63%	986	199
5	25%	75%	538	887

For the second trial, five samples of custard were spiked with different levels of starch (Table 3.2). In the end two validation stages were required due to large differences between samples being perceived in the first trial.

Table 3.2: Custard samples and 3-digit numbers used to code the products for 2 replicate assessments.

		1st validation stage		2nd validation stage	
Sample	**Level of Thickener**	**Code Rep 1**	**Code Rep** 2	**Code Rep 1**	**Code Rep** 2
1	Level 1 (low)	541	825	513	623
2	Level **2**	638	917	202	915
3	Level 3	575	397	132	217
4	Level 4	398	465	816	494
5	Level 5 (high)	170	280	443	568

Panels

Four validation panels participated in the apple juice trial, of which 2 were trained and 2 untrained. For the main trial, 10 panels, participated of which 8 were trained and 2 untrained.

Four validation panels participated in the custard trial, of which 2 were trained and 2 untrained. For the main trial, 13 panels, participated of which 11 were trained and 2 untrained.

Ranking Procedure

For the apple juice, assessors were asked to rank the samples according to sweetness intensity. Panels ranked samples using their normal procedure: either '1 = most' and '5 = least', or '1 = least' and '5 = most'. Data were converted to the former ranking system.

For the custard, assessors were asked to rank the samples according to perceived thickness. Panels ranked samples using their normal procedure: either '1 = most' and '5 = least', or '1 = least' and '5 = most'. Data were converted to the former ranking system.

3.2 Validation Results – Apple Juice

3.2.1 Panel Rank Means, Friedman Test and Conover Multiple Comparison

For the apple juice trial, 2 trained and 2 untrained panels were selected to investigate the procedure for setting the 'expected results' and the resulting performance criteria.

Table 3.3: Results of the Friedman and multiple comparison tests for the 4 validation panels.

	Panel							
	A		**E**		**F1**		**I1**	
Sample	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**
6111853	5.0	4.8	4.4	4.6	3.9	4.3	4.9	4.6
2081460	4.0	4.1	3.8	3.2	3.4	3.2	3.6	3.4
4361798	2.8	3.1	2.9	3.0	3.6	2.3	2.8	2.5
9861199	1.9	1.7	2.1	2.4	2.3	2.8	2.0	3.0
5381887	1.3	1.3	1.8	1.8	1.8	2.4	1.8	1.5
p-value	0.000	0.000	0.000	0.000	0.003	0.013	0.000	0.002
MC-5%	0.49	0.52	0.98	1.03	1.11	1.16	1.02	1.19
n-assessor	9	9	12	12	12	12	8	8
Significant Differences	10	9	6	6	5	3	6	6

The first step is to calculate the panel mean ranks (Table 3.3), and to then set the ***expected panel rank means.*** The overall sample rank order corresponds well across all panels, and Panel **A** shows the most discrimination between samples. Thus, Panel **A** was used to define the expected sample rank means, as shown in Table 3.4.

Table 3.4: Mean panel data for Panel A over 2 replicates used to calculate expected mean rank.

	A		Expected
Sample	**Rep 1**	**Rep 2**	**Result**
611/853	5.0	4.8	4.9
208/460	4.0	4.1	4.0
436/798	2.8	3.1	3.0
986/199	1.9	1.7	1.8
538/887	1.3	1.3	1.3

Table 3.5 shows the Pearson correlation coefficient between the 'expected rank means' and the rank means for each of the four validation panels. Based on these results, the following performance criteria were set.

Score 0	if $p > 0.10$	or if a negative correlation
Score 1	if $p \leq 0.10$	'expected result'

If a panel provides a significant negative correlation with the expected rank means, then this would result in an immediate decision that the laboratory was not proficient.

Table 3.5: Correlation between the panel mean data and the expected mean rank.

	Panel			
Correlation	**A**	E	**F1**	**I1**
Replicate 1	0.997	0.997	0.924	0.981
(p-value)	(0.000)	(0.000)	(0.012)	(0.002)
Replicate 2	0.998	0.957	0.812	0.997
(p-value)	(0.000)	(0.006)	(0.048)	(0.026)

The second step is to establish the level of significance associated with the sample differences. The results of the Friedman test (Table 3.3) indicate that Panels A and E have $p \leq 0.001$ on

both replicates, whilst Panel F1 has p = 0.003 and p = 0.013, and Panel I1 has p ≤ 0.001 and p = 0.002. Therefore the following performance criteria were set.

Score 0	if p > 0.05	
Score 1	if p ≤ 0.05	
Score 2	if p ≤ 0.01	'expected result'
Score 3	if p ≤ 0.001	

The third step is to identify the number of pairs of samples that are significantly different at the 5% level of significance. There are potentially 10 pairs of samples that can show significant differences. It was decided to look at the 5% level of significance (Step 2a). From Table 3.3 it can be observed that Panel A achieved 10 significant pairs in the 1st replicate, whilst Panel F1 achieved only 3 significant pairs in the 2nd replicate. Therefore, the following performance criteria were set.

Score 0	if 0 or 1 significant differences	
Score 1	if 2 or 3 significant differences	
Score 2	if 4 or 5 significant differences	
Score 3	if 6 or 7 significant differences	
Score 4	if 8 significant differences	'expected result'
Score 5	if 9 significant differences	
Score 6	if 10 significant differences	

3.2.2 Coefficient of Concordance

Table 3.6: Coefficient of concordance (W) for the validation panels.

Panel	Replicate 1	Replicate 2
A	0.91	0.90
E	0.48	0.42
F1	0.33	0.27
I1	0.65	0.53

The coefficient of concordance calculated for the validation panels is shown in Table 3.6 for both replicate assessments. This forms Step 2b of the performance scheme. Based on these results, the following performance criteria were specified.

Score 0	If $W < 0.60$	
Score 1	if $W \geq 0.60$	
Score 2	if $W \geq 0.70$	
Score 3	if $W \geq 0.80$	'expected result'
Score 4	if $W \geq 0.90$	

3.2.3 Setting the Final Expected Result Criteria

Based on the performance criteria scores given for Steps 1-3 above, a total possible score of 14 (1 + 3 + 4 + 6) is achievable. If the 'expected results' from Steps 1-3 are added together a total score of 10 (1 + 2 + 3 + 4) is specified. Given that a panel can score 1 less than the expected result on any step (1, 2a, 2b or 3), then the expected overall score could be set as the interval 9-10.

Score = 10.1-14.0	Better than expected
Score = 9 - 10	'Expected result'
Score < 9.0	Less than expected

3.3 Validation Results – Custard

3.3.1 Panel Rank Means, Friedman Test and Conover Multiple Comparison

For the custard trial, a screening/pre-test was undertaken, followed by a validation stage. Table 3.7 shows the results from the 4 validation panels.

Table 3.7: Results of the Friedman and multiple comparison tests for the 4 validation panels.

	A		D		H1		J1	
Sample	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**
5411825	5.0	5.0	4.4	5.0	4.3	4.8	4.6	4.6
6381917	3.8	3.2	4.6	3.9	3.9	4.1	4.0	4.2
5751397	3.3	3.8	2.8	3.1	2.9	2.8	3.0	2.8
3981465	2.0	1.8	2.0	1.9	2.1	2.2	2.3	2.3
170/280	1.0	1.2	1.2	1.1	1.8	1.1	1.2	1.3
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MC-5%	0.26	0.32	0.52	0.29	0.96	0.44	0.69	0.67
n-assessor	12	12	10	10	12	12	12	12
Significant Differences	10	10	9	10	6	10	9	8

It appears that Panel A switched samples 397 and 917 in the 2nd replicate. This was further confirmed by the fact that all pairs of samples were significantly different.

3.3.2 Coefficient of Concordance

Table 3.8: Coefficient of concordance (W) for the validation panels.

Panel	Replicate 1	Replicate 2
A	0.963	0.944
D	0.880	0.964
H1	0.503	0.893
J1	0.743	0.756

Table 3.8 shows the coefficient of concordance to measure the agreement between assessors

in the panel. It can be seen that Panels A and D show good agreement, whilst the untrained panels, H1 and J1, have less agreement between the assessors.

3.3.3 Setting the Final Expected Result Criteria

From the above results, it was decided that the ranking task was not sufficiently challenging, and the increments in the thickening agent were reduced.

The consequence of this step was that 2 trained and 2 untrained panels from the main trial were allocated as validation panels for the purpose of setting the performance criteria. Section 3.4 works through the results from the validation panels: Panels A, A1, K and K1.

3.4 Validation Results – 2nd Custard Trial

3.4.1 Panel Rank Means, Friedman Test and Conover Multiple Comparison

Table 3.9 shows the results from the 4 panels allocated to be regarded as the second validation stage. There is a possibility that Panel A1 had samples 915 and 623 switched in the 2nd replicate. This raises some questions about the method of preparation used by this panel.

In order to undertake the first step, the mean of the panel mean ranks was calculated for Panels A and K (Table 3.10).

Table 3.9: Results of the Friedman and multiple comparison tests for the 4 validation panels.

	A		A1		K		K1	
Sample	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**
5131623	4.7	4.9	4.2	3.6	4.9	5.0	5.0	4.8
2021915	4.3	3.6	4.7	4.9	3.3	4.0	3.3	3.7
1321217	3.0	2.8	3.1	2.6	3.6	2.5	2.9	3.0
8161494	1.9	2.7	1.8	2.7	2.3	2.4	2.4	2.6
4431568	1.1	1.0	1.2	1.1	1.0	1.1	1.3	1.0
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MC-5%	0.33	0.58	0.43	0.70	0.50	0.37	0.71	0.66
n-assessors	12	12	11	11	12	12	12	12
Significant Differences	10	9	10	9	9	9	8	9

Table 3.10: Mean panel data for Panels A and K over 2 replicates used to calculate expected mean rank.

	A		K		Expected
Sample	**Rep 1**	**Rep 2**	**Rep 1**	**Rep 2**	**Result**
443/568	1.1	1.0	1.0	1.1	1.0
816/494	1.9	2.7	2.3	2.4	2.3
132/217	3.0	2.8	3.6	2.5	3.0
202/915	4.3	3.6	3.3	4.0	3.8
513/623	4.7	4.9	4.9	5.0	4.9

Table 3.11: Correlation between the panel mean data and the expected mean rank.

	Panel			
Correlation	**A**	**A1**	**K**	**K1**
Replicate 1	0.975	0.910	0.965	0.981
(p-value)	(0.005)	(0.032)	(0.008)	(0.003)
Replicate 2	0.986	0.806	0.983	0.995
(p-value)	(0.002)	(0.099)	(0.003)	(0.008)

Table 3.11 shows the Pearson correlation coefficient between the 'expected rank means' and the rank means for each of the four validation panels. Based on these results, the following performance criteria were set.

Score 0	if $p > 0.10$	or if a negative correlation
Score 1	if $p \geq 0.10$	'Expected result'

If a panel provides a significant negative correlation with the expected rank means, then this would result in an immediate decision that the laboratory was not proficient.

The next step is to establish the level of significance associated with the sample differences. The results of the Friedman test (Table 3.9) indicate that all panels show highly significant results ($p < 0.001$). This suggests that the criteria should be a little stricter than for the apple juice. The following criteria were set.

Score 0	if $p > 0.01$	
Score 1	if $p \geq 0.01$	
Score 2	if $p \leq 0.001$	'Expected result'

The third step is to identify what pairs of samples are significantly different at the 5% level of significance. There are potentially 10 pairs of samples that can show significant differences. The 5% level of significance was chosen, as in the Apple Juice scheme. However, there is a good case for choosing the 1% level. Section 3.5 looks at this alternative.

From Table 3.9 it can be observed that all panels found at least 8 significant differences, with 2 assessments finding all 10 pairs significantly different at the 5% level of significance. Thus, the following criteria were set.

Score 0	if ≤ 5 significant differences	
Score 1	if 6 significant differences	
Score 2	if 7 significant differences	
Score 3	if 8 significant differences	
Score 4	if 9 significant differences	'Expected result'
Score 5	if 10 significant differences	

3.4.2 Coefficient of Concordance

Table 3.12: Coefficient of concordance (W) for the validation panels.

Panel	Replicate 1	Replicate 2
A	0.940	0.815
A1	0.909	0.790
K	0.864	0.926
K1	0.724	0.768

The coefficieiit of concordance was calculated for the validation panels (Table 3.12) for both replicate assessments. Based on these results the following performance criteria were set.

Score 0	if W < 0.80	
Score 1	If W ≥ 0.80	
Score 2	if W ≥ 0.85	
Score 3	if W ≥ 0.90	'Expected result'
Score 4	if W ≥ 0.95	

3.4.3 Setting the Final Expected Result Criteria

The results of the 2nd validation stage also suggested that the task of ranking was relatively easy. Whilst for demonstration it would have been better if the task had been harder, the following outlines the criteria set-up to measure the performance of the panels participating in the custard ranking ring trial.

Based on the performance criteria scores given for Steps 1-4 above, a total possible score of 12 (1 + 2 + 4 + 5) is achievable. If the 'expected results' from Steps 1-3 are added together a total score of 10 (1 + 2 + 3 + 4) is specified. Given that a panel can score 1 less than the expected result on any step (1, 2a, 2b or 3), then the expected overall score could be set as the interval 9-10.

Score = 10.1-12.0	Better than expected
Score = 9 - 10	'Expected result'
Score < 9.0	Less than expected

4. 2000 RING TRIAL RESULTS

4.1 Apple Juice Results

4.1.1 Panel Rank Means, Friedman Test and Conover Multiple Comparison

Tables 4.1 and 4.2 show the panel mean ranks, the results of the Friedrnan test (p-value), multiple comparison value at 5% significance and the number of significant pairs at this significance level.

4.1.2 Step 1 – Correlation with Expected Rank Means

Table 4.3 shows the results of the correlation between the panel sample rank means and expected sample rank means. With the exception of Panel G, all panels received a score of 1 on both replicate assessments.

Table 4.1: Panel rank means, Friedman and multiple comparison results: Replicate 1.

	Panel													
Sample	**A**	B	**C**	D	E	**F**	***G***	H	**I**	**J**	**F1**	**I1**	**J1**	**K1**
611	5.0	4.5	4.9	3.8	4.4	4.2	3.1	4.6	4.3	5.0	3.9	4.9	4.8	4.3
208	4.0	3.6	4.1	4.6	3.8	3.4	2.4	3.8	3.5	4.0	3.4	3.6	3.5	3.5
436	2.8	2.4	2.8	3.0	2.9	3.3	3.0	3.2	2.9	2.9	3.6	2.8	2.3	3.6
986	1.9	2.7	2.2	1.3	2.1	2.3	2.8	1.9	2.0	1.8	2.3	2.0	2.2	2.1
538	1.3	1.8	1.0	2.3	1.8	1.8	2.8	1.5	2.3	1.3	1.8	1.8	2.2	1.6
p-value	0.000	0.000	0.000	0.000	0.000	0.003	0.799	0.008	0.000	0.000	0.003	0.000	0.000	0.000
MC-5%	0.49	0.91	0.34	0.99	0.98	1.10	1.25	0.79	1.22	0.39	1.11	1.02	0.92	0.96
n	9	14	10	9	12	12	16	11	10	10	12	8	12	12
Significant Differences	10	5	10	6	6	4	0	7	4	10	5	6	7	7

Table 4.2: Panel rank means, Friedman and multiple comparison results: Replicate 2.

Sample	Panel													
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**F1**	**I1**	**J1**	**K1**
853	4.8	4.6	5.0	4.2	4.6	4.3	2.8	4.5	5.0	5.0	4.3	4.6	4.1	3.9
460	4.1	3.6	3.9	3.3	3.2	3.6	3.1	3.7	3.5	3.7	3.2	3.4	3.3	4.0
798	3.1	3.2	3.0	3.6	3.0	2.5	3.1	3.3	2.9	3.0	2.3	2.5	2.6	3.3
199	1.7	2.3	2.1	2.7	2.4	2.7	2.8	2.3	1.9	2.3	2.8	3.0	3.2	2.3
887	1.3	1.3	1.0	1.1	1.8	1.9	3.3	1.2	1.7	1.0	2.4	1.5	1.9	1.4
p-value	0.000	0.000	0.000	0.000	0.000	0.002	0.844	0.000	0.000	0.000	0.013	0.002	0.000	0.000
MC-5%	0.52	0.73	0.29	1.08	1.03	1.08	1.14	0.81	0.81	0.48	1.16	1.19	1.17	0.97
n	9	14	10	9	12	12	16	11	10	10	12	8	12	12
Significant Differences	9	9	10	5	6	4	0	6	7	10	3	6	4	5

Table 4.3: Pearson correlation between panel and expected sample rank means.

Panel	Replicate 1	Replicate 2
A	0.997	0.998
B	0.922	0.979
C	0.984	0.990
D	0.828	0.873
E	0.998	0.957
F	0.977	0.934
G	0.106	-0.424
H	0.996	0.945
I	0.967	0.979
J	0.999	0.975
F1	0.924	0.812
I1	0.981	0.877
J1	0.910	0.817
K1	0.955	0.943

4.1.3 Step 2a – Level of Significance Associated with Sample Differences

Table 4.4 shows the p-value corresponding to the significance level associated with sample differences, on undertaking the Friedman test. In addition, the performance score achieved is recorded.

Table 4.4: Significance levels (p-values) associated with the Friedman test together with allocated performance score.

	Replicate 1		Replicate 2	
Panel	**p-value**	**Score**	**p-value**	**Score**
A	0.000	3	0.000	3
B	0.000	3	0.000	3
C	0.000	3	0.000	3
D	0.000	3	0.000	3
E	0.000	3	0.000	3
F	0.003	2	0.002	2
G	0.799	0	0.844	0
H	0.008	2	0.000	3
I	0.000	3	0.000	3
J	0.000	3	0.000	3
F1	0.003	2	0.013	1
I1	0.000	3	0.002	2
J1	0.000	3	0.000	3
K1	0.000	3	0.000	3

4.1.4 Step 2b –Agreement Between Assessors – Coefficient of Concordance

Table 4.5 shows the coefficient of concordance to measure the agreement between assessors in a panel. This measure is distinguishing between panels in terms of level of performance.

Table 4.5: Coefficient of concordance (W) for all panels.

	Replicate 1		Replicate 2	
Panel	**W**	**Score**	**W**	**Score**
A	0.91	4	0.90	4
B	0.46	0	0.65	1
C	0.95	4	0.96	4
D	0.62	1	0.56	0
E	0.48	0	0.42	0
F	0.34	0	0.37	0
G	0.04	0	0.02	0
H	0.70	2	0.68	1
I	0.34	0	0.72	2
J	0.93	4	0.90	4
F1	0.33	0	0.27	0
I1	0.65	1	0.53	0
J1	0.54	0	0.26	0
K1	0.50	0	0.49	0

4.1.5 Step 3 – Significantly Different Sample Pairs

Table 4.6 shows the number of significantly different pairs out of a possible total of ten, together with the allocated performance score. It can be seen that panels differ in their ability to discriminate between pairs of samples, in spite of having similar p-values. This illustrates the importance of looking at specific sample differences, rather than just the overall test result.

Table 4.6: Number of significant pairs at the 5% level with performance score.

	Replicate 1		Replicate 2	
Panel	**Pairs**	**Score**	**Pairs**	**Score**
A	10	6	9	5
B	5	2	9	5
C	10	6	10	6
D	6	3	5	2
E	6	3	6	3
F	4	2	4	2
G	0	0	0	0
H	7	3	6	3
I	4	2	7	3
J	10	6	10	6
F1	5	2	3	1
I1	6	3	6	3
J1	7	3	4	2
K1	7	3	5	2

4.2 Custard Results

4.2.1 Panel Rank Means, Friedman Test and Conover Multiple Comparison

Tables 4.7 and 4.8 show the results of the Friedman test, multiple comparison value and the number of significantly different pairs at the 5% level.

Table 4.7: Panel rank means, Friedman and multiple comparison results: Replicate 1.

Sample	Panel												
	A	**A1**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**K1**
443	1.1	1.2	1.0	1.0	2.1	1.3	1.0	1.1	1.0	1.1	1.1	1.0	1.3
816	1.9	1.8	2.0	2.5	1.7	1.7	2.2	1.9	2.1	1.9	1.9	2.3	2.4
132	3.0	3.1	3.0	2.5	2.5	3.0	2.8	3.1	3.1	3.0	3.1	3.6	2.9
202	4.3	4.7	4.0	4.1	3.7	4.2	4.2	3.9	3.8	4.0	3.9	3.3	3.3
513	4.7	4.2	5.0	4.9	5.0	4.8	4.8	5.0	5.0	5.0	5.0	4.9	5.0
MC-5%	0.33	0.43	0.00	0.39	0.79	0.41	0.32	0.32	0.32	0.20	0.29	0.50	0.71
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n	12	11	7	10	10	10	12	9	11	10	10	12	12
Significant Differences	10	10	10	9	8	9	10	10	10	10	10	9	8

Table 4.8: Panel rank means, Friedman and multiple comparison results: Replicate 2.

Sample	Panel												
	A	**A1**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**K1**
568	1.0	1.1	1.0	1.1	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.0
494	2.7	2.7	2.0	1.9	2.0	2.0	2.0	2.2	2.2	2.1	2.0	2.4	2.6
217	2.8	2.6	3.1	3.0	3.2	3.0	3.3	3.1	3.0	2.9	3.1	2.5	3.0
915	3.6	4.9	3.9	4.0	3.8	4.0	3.7	4.4	3.9	4.3	4.5	4.0	3.7
623	4.9	3.6	5.0	5.0	4.9	5.0	5.0	4.2	4.9	4.7	4.4	5.0	4.8
MC-5%	0.58	0.70	0.21	0.20	0.50	0.00	0.29	0.68	0.42	0.37	0.42	0.37	0.66
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
n	12	11	7	10	10	10	12	9	11	10	10	12	12
Significant Differences	9	9	10	10	10	10	10	9	10	10	9	9	9

4.2.2 Step 1 – Correlation with Expected Rank Means

Table 4.9 shows the results of the correlation between the panel sample rank means and expected sample rank means: all were significant at the 10% level.

Table 4.9: Pearson correlation between panel and expected sample rank means.

Panel	Replicate 1	Replicate 2
A	0.975	0.986
A1	0.910	0.806
B	0.995	0.995
C	0.979	0.990
D	0.889	0.992
E	0.966	0.995
F	0.988	0.990
G	0.990	0.949
H	0.997	0.999
I	0.990	0.982
J	0.990	0.954
K	0.965	0.983
K1	0.981	0.995

4.2.3 Step 2a – Level of Significance Associated with Sample Differences

Table 4.10 shows the p-value corresponding to the significance level associated with sample differences, on undertaking the Friedman test, together with the performance score achieved.

4.2.4 Step 2b – Agreement Between Assessors - Coefficient of Concordance

Table 4.11 lists the coefficient of concordance for each panel, together with the allocated performance score.

Table 4.10: Significance levels (p-values) associated with the Friedman test together with allocated performance score.

	Replicate 1		Replicate 2	
Panel	**p-value**	**Score**	**p-value**	**Score**
A	0.000	2	0.000	2
A1	0.000	2	0.000	2
B	0.000	2	0.000	2
C	0.000	2	0.000	2
D	0.000	2	0.000	2
E	0.000	2	0.000	2
F	0.000	2	0.000	2
G	0.000	2	0.000	2
H	0.000	2	0.000	2
I	0.000	2	0.000	2
J	0.000	2	0.000	2
K	0.000	2	0.000	2
K1	0.000	2	0.000	2

Table 4.11: Coefficient of concordance (W) for all panels.

	Replicate 1		Replicate 2	
Panel	**W**	**Score**	**W**	**Score**
A	0.940	3	0.815	1
A1	0.909	3	0.790	0
B	1.000	4	0.976	4
C	0.932	3	0.982	4
D	0.724	0	0.890	2
E	0.926	3	1.000	4
F	0.944	3	0.956	4
G	0.960	4	0.819	1
H	0.950	4	0.914	3
I	0.982	4	0.940	3
J	0.964	4	0.922	3
K	0.864	2	0.926	3
K1	0.724	0	0.768	0

4.2.5 Step 3 – Significantly Different Sample Pairs

Table 4.12 shows the number of sample pairs that were significantly different at the 5% level of significance, together with the allocated performance score.

Table 4.12: Number of significant pairs at the 5% level with performance score.

	Replicate 1		Replicate 2	
Panel	**Pairs**	**Score**	**Pairs**	**Score**
A	10	5	9	4
A1	10	5	9	4
B	10	5	10	5
C	9	4	10	5
D	8	3	10	5
E	9	4	10	5
F	10	5	10	5
G	10	5	9	4
H	10	5	10	5
I	10	5	10	5
J	10	5	9	4
K	9	4	9	4
K1	8	3	9	4

5. PANEL PERFORMANCE

5.1 Apple Juice Panel Performance

Tables 5.1 and 5.2 show the performance scores for each of the 3 steps, for Replicates 1 and 2, respectively. The values shown for Step 4 are simply the sum of Steps 1 to 3.

It can be seen that Panel G scored 0 on all criteria, and in fact on further investigation it was established that this panel failed to use the data entry sheet correctly.

Table 5.1: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: Replicate 1.

Panel	Step 1	Step 2a	Step 2b	Step 3	Step 4
A	1	***3***	4	**6**	14
B	1	***3***	0	2	**6**
C	1	***3***	4	6	14
D	1	3	1	3	8
E	1	**3**	0	3	7
F	1	2	0	2	5
G	0	0	0	0	0
H	1	2	2	3	8
I	1	**3**	0	2	6
J	1	**3**	4	6	14
F1	1	2	0	2	5
I1	1	3	1	3	8
J1	1	3	0	**3**	7
K1	1	3	0	**3**	7

Table 5.2: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: Replicate 2.

Panel	Step 1	Step 2a	Step 2b	Step 3	Step 4
A	1	3	4	*5*	13
B	1	3	1	**5**	10
C	1	**3**	4	6	14
D	1	3	0	2	6
E	1	3	0	3	7
F	1	2	0	2	*5*
G	0	0	0	0	0
H	1	3	1	3	8
I	1	3	2	3	9
J	1	3	4	6	14
F1	1	1	0	1	**3**
I1	1	2	0	3	6
J1	1	3	0	2	6
K1	1	3	0	2	**6**

Table 5.3 shows the average results for each step, where the last row in the table indicates the 'expected result'. With the exception of Panel G, all panels achieved the expected result in Step 1. For Step 2a, only Panels G and F1 were below expected, whilst all 4 untrained panels, Panels B, D, H and I, were below expected. With respect to Step 2b, only Panels A, C and J achieved the expected result.

Table 5.3: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: average over replicates.

Panel	**Step 1**	**Step 2a**	**Step 2b**	**Step** 3	**Step 4**
A	1	3	4	5.5	13.5
B	1	3	0.5	3.5	**8**
C	1	3	4	6	14
D	1	3	0.5	2.5	7
E	1	3	0	3	7
F	1	2	0	2	5
G	0	0	0	0	0
H	1	2.5	1.5	3	8
I	1	3	1	2.5	7.5
J	1	3	4	6	14
F1	1	1.5	0	1.5	4
I1	1	2.5	0.5	3	7.0
J1	1	3	0	2.5	6.5
K1	1	3	0	2.5	6.5
Expected	***1***	***2***	**3**	***4***	9 **–** ***10***

Table 5.4 shows the average overall performance over the 2 replicate assessments, together with the overall performance score. Whilst most panels were below the expected result, six panels (B, D, E, H, I and I1) were only up to 2 points below the lower score of the overall expected result. It may be considered, for example, that Panel F1 needs more training.

Table 5.4: Summary of overall performance.

Panel	Replicate 1	Replicate 2	Average	Grade
A	14	13	13.5	***> expected***
B	**6**	10	8.0	< expected
C	14	14	14.0	***> expected***
D	8	6	7.0	< expected
E	7	7	7.0	< expected
F	5	5	5.0	< expected
G	0	0	0	< expected
H	8	8	8.0	< expected
I	6	9	7.5	< expected
J	14	14	14.0	***> expected***
F1	5	3	4	< expected
I1	8	6	7.0	< expected
J1	7	6	6.5	< expected
K1	7	6	6.5	< expected
Expected			***9 - 1 0***	

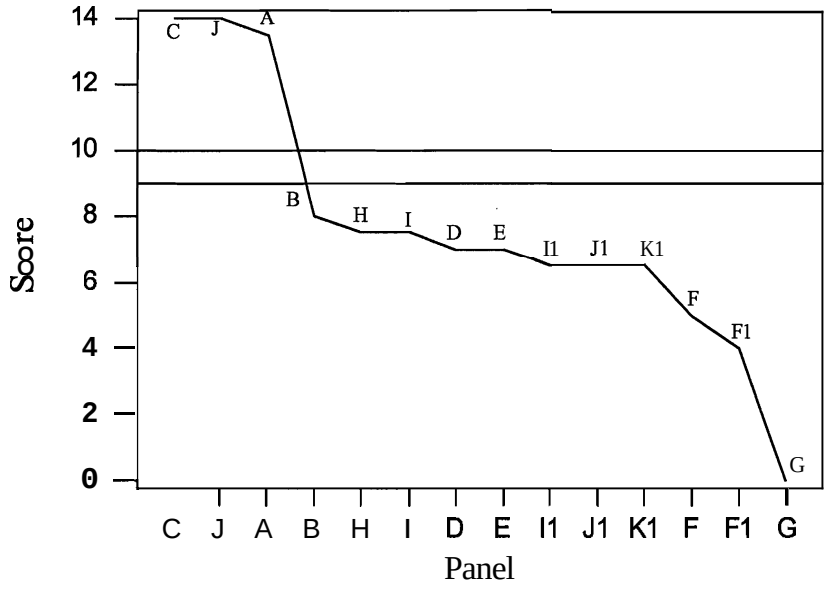
Note on Panel G

As previously noted, Panel G scored 0 on all criteria, which was due to failing to use the data entry sheet correctly. As it was agreed to treat Year 2 data in the spirit of a 'real' ring trial, the data were used in this form to illustrate the seriousness of failing to give attention to all aspects of a ring trial. In fact, as shown in Appendix 1, which provides the results based on the corrected data, this panel achieved a score of 8.5, just below 'expected'.

Graphical Display

Figure 5.1 represents the final overall performance score as a histogram, illustrating the expected result band.

Figure **5.1**: **Apple** juice performance summary.



This graph illustrates that most panels performed around the same level, even though overall they scored slightly below the specified overall expected score. As will be seen later, the graph for custard is quite different.

5.2 Custard Panel Performance

Tables 5.5 and 5.6 show the performance scores for each of the 4 steps, for Replicates 1 and 2, respectively. The values shown for Step 4 are simply the sum of Steps 1 to **3.**

Table 5.5: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: Replicate 1.

Panel	Step 1	Step 2a	Step 2b	Step 3	Step 4
A	1	2	**3**	**5**	11
A1	1	2	**3**	**5**	11
B	1	2	**4**	***5***	12
C	1	2	3	4	10
D	1	2	0	3	6
E	1	2	3	4	10
F	1	2	3	***5***	11
G	1	2	4	***5***	12
H	1	2	4	***5***	12
I	1	2	4	***5***	12
J	1	2	4	***5***	12
K	1	2	2	4	9
K1	1	2	0	3	6

Table 5.6: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: Replicate 2.

Panel	Step 1	Step 2a	Step 2b	Step 3	Step 4
A	1	2	1	**4**	8
A1	1	2	0	**4**	7
B	1	2	**4**	***5***	12
C	1	2	4	***5***	12
D	1	**2**	2	***5***	10
E	1	2	4	***5***	12
F	1	2	4	***5***	12
G	1	2	1	4	8
H	1	2	3	***5***	11
I	1	2	3	***5***	11
J	1	2	3	4	10
K	1	2	3	4	10
K1	1	2	0	4	7

Table 5.7 shows the average results for each step, where the last row in the table indicates the 'expected result'. All panels achieved the expected result in Steps 1 and 2a. Panels G and K were just below expected for Step 2b, as was Panel A, but Panels D, A1 and K1 were more than 1 point below. Only Panel K1 was just below expected for Step 3.

Table 5.7: Summary of performance for each of the 3 Steps, where Step 4 is the sum of the others: average over replicates.

Panel	**Step 1**	**Step 2a**	**Step 2b**	**Step 3**	**Step 4**
A	1	**2**	**2**	4.5	9.5
A1	1	**2**	1.5	4.5	9.0
B	1	**2**	4	5	12.0
C	1	***2***	3.5	4.5	11.0
D	1	***2***	1	4	8
E	1	***2***	3.5	4.5	11.0
F	1	*2*	3.5	5	11.5
G	1	**2**	2.5	4.5	10.0
H	1	**2**	3.5	5	11.5
I	1	***2***	3.5	5	11.5
J	1	***2***	3.5	4.5	11.0
K	1	**2**	2.5	4	9.5
K1	1	***2***	0	3.5	6.5
Expected	***1***	***2***	**3**	***4***	9 – 10

Table 5.8 shows the average overall performance over the 2 replicate assessments, with most panels performing better than the overall expected result. Only Panels D and K1 had a score below the lower limit of the overall expected score, whilst Panels A, A1 and K fell within the overall expected score interval.

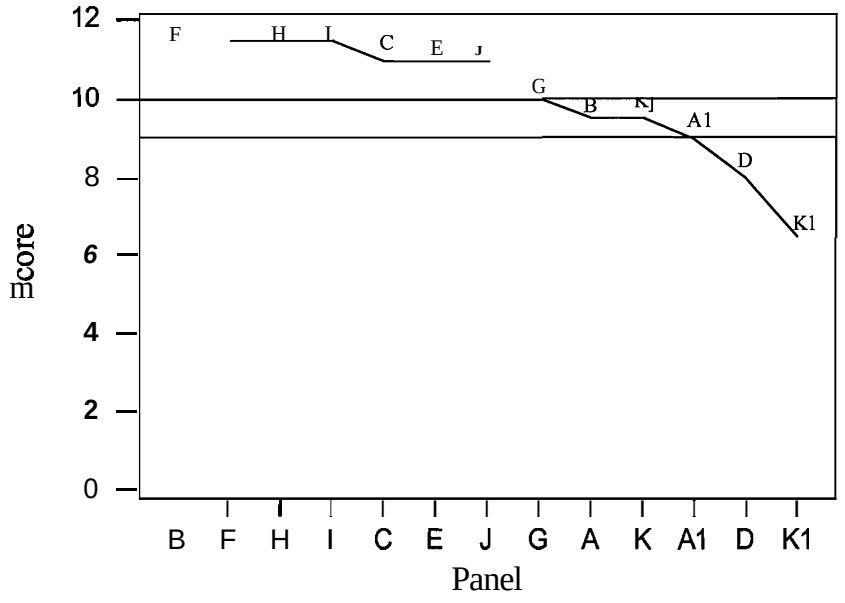
Table 5.8: Summary of overall performance.

Panel	Replicate 1	Replicate 2	Average	Grade
A	11	**8**	9.5	= ***expected***
A1	11	7	9.0	= ***expected***
B	12	12	12.0	> ***expected***
C	10	12	11.0	> ***expected***
D	6	10	8	< expected
E	10	12	11.0	> ***expected***
F	11	12	11.5	> ***expected***
G	12	8	10.0	> ***expected***
H	12	11	11.5	> ***expected***
I	12	11	11.5	> ***expected***
J	12	10	11.0	> ***expected***
K	9	10	9.5	= ***expected***
K1	6	7	6.5	< expected
Expected			***9 - 10***	

Graphical Representation

Figure 5.2 represents the final overall performance score as a histogram, illustrating the expected result band. This graph is quite different from Figure 5.1, and illustrates that the majority of panels performed as expected or better.

Figure 5.2 Custard performance summary.



6. GUIDANCE FOR PERF'ORMANCE MEASURES

Based on the worked examples of the apple juice and custard ring trials, the performance scheme proposed in Chapter 2 is shown to discriminate between the laboratories. However, it is useful to review a number of issues.

6.1 Screening, Pre-testing and Validation

The importance of screening, pre-testing and validation cannot be over-emphasised. As demonstrated in the custard trial, the initial screening and pre-testing resulted in saniples that showed (too) large differences. However, in spite of reducing the increments of thickening agents for the main trial, the ranking task was still 'too easy'. A second screening and pre-test would have had every chance of detecting that the task was still not sufficiently difficult. It is therefore recommended that samples are always screened and pre-tested, thus giving the best chance for the validation phase to be successfully used to set the performance criteria and expected results.

6.2 Setting Performance Criteria

Whilst setting the criteria for the apple juice trial was relatively straightforward, the exercise on custard was not so clear-cut. This indicates the importance of working through several scenarios prior to finalising the performance criteria. In addition, it is also important to consider the experience of the validation laboratories, as panels with expertise in a product category could result in the expected results being set too high.

In this exercise, certain significance levels were chosen, but it should be remembered that

these should be chosen on the basis of the data, and therefore should be reviewed for each new product, or perhaps as a result of experience with previous ring trials.

Moreover, much effort was put into developing the performance criteria through a scoring system. This resulted in the importance of considering performance at each step in the scheme, rather than just considering a final overall score. In addition, the importance of inadvertently over-weighting a step became very apparent in earlier versions of the Performance Scheme, and this has been considered. In fact, it was felt that discrimination between pairs of samples was the most important step.

6.3 Further Research

Clearly, there are still issues that need to be considered in terms of improving the Performance Scheme further. It could be useful to develop a more statistically based weighting procedure for each of the steps. Moreover, the concept of confidence intervals could be an attractive option.

One final issue is the ability to compare results across ring trials. Clearly a laboratory wants to demonstrate improvement over time. However, the Performance Scheme will differ for different products and depending how challenging the task is in terms of perceptible differences between samples. More thought may be required as, at present, results can mainly be compared within a ring trial.

REFERENCES

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

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

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

Lyon, D.H. (2001). Guidelines for Proficiency Testing in Sensory Analysis. In preparation.

McEwan, J.A.. (2000). Proficiency Testing for Sensory Ranking Tests: Statistical Guidelines. Part 1. R&D Report No. 118. CCFRA.

McEwan, J.A.. (2001). Proficiency Testing for Sensory Profile Tests: Statistical Guidelines. Part 2. R&D Report No. 127. CCFRA.

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.

APPENDIX 1: PANEL G CORRECTED RESULTS

Friedman Test Results

Sample	Rep	
	1	2
611/853	4.0	4.6
208/460	3.8	4.3
436/798	2.4	2.8
986/199	2.4	1.9
538/887	1.5	1.5
MC-5%	1.0	0.5
p-value	0.000	0.000
n	16	16
W	0.452	0.778
Significant Differences	6	8

Performance Score

Based on the scheme used for the apple juice, the following table provides the score achieved for each step (Step 4 is the sum of Steps 1 to 3). This panel achieved a score just below the 'expected overall score'.

	Replicate 1	Replicate 2	Mean
Step 1	1	1	1
Step 2a	3	3	3
Step 2b	0	2	1
Step 3	3	4	3.5
Step 4	7	10	8.5